

# **COLT 2011 Proceedings**

**Volume 19:**

**24th Annual Conference on Learning Theory**

*Sham M. Kakade and Ulrike von Luxburg*



## Preface

The 24th International Conference on Learning Theory, COLT 2011, was held in Budapest, Hungary between July 9-11, 2011. It was co-located with the Foundations of Computational Mathematics conference (FOCM 2011, July 4th - 14th). There was a total of 36 accepted papers and a total of 6 open problems. The best paper prize was awarded to Alexander Rakhlin, Karthik Sridharan and Ambuj Tewari for their paper "Online Learning: Beyond Regret" The best student paper prize was awarded to Amit Daniely and Sivan Sabato for "Multiclass Learnability and the ERM Principle" (co-authored with Shai Ben-David and Shai Shalev-Shwartz). The invited speakers were William T. Freeman and David J. Hand.

The organization committee consisted of:

Program Chairs:

Sham Kakade and Ulrike von Luxburg

Local Arrangements Chair:

András György

Publicity Chair:

Jeff Jackson

Open Problems Chair:

Alina Beygelzimer

The program committee consisted of:

Jean-Yves Audibert, Nina Balcan, Peter Bartlett, Shai Ben-David, Gilles Blanchard, Nicolò Cesa-Bianchi, Kamalika Chaudhuri, Koby Crammer, Claudio Gentile, Steve Hanneke, Elad Hazan, Matthias Hein, Daniel Hsu, Satyen Kale, Adam Klivans, Katrina Ligett, Gábor Lugosi, Shie Mannor, Yishay Mansour, Massimiliano Pontil, Sasha Rakhlin, Cynthia Rudin, Shai Shalev-Shwartz, Ohad Shamir, Yoram Singer, Nati Srebro, Ingo Steinwart, Gilles Stoltz, Ambuj Tewari, Manfred Warmuth, Bob Williamson, Jenn Wortman Vaughan and Tong Zhang

The conference was applauded as a success.

Sincerely,

Sham M. Kakade and Ulrike von Luxburg, the program chairs



## Table of Contents

### Preface

---

### Part I Regular papers

---

<i>Abbasi-Yadkori, Szepesvari: Regret Bounds for the Adaptive Control of Linear Quadratic Systems</i>	1
<i>Abernethy, Bartlett, Hazan: Blackwell Approachability and No-Regret Learning are Equivalent</i>	27
<i>Acharya, Das, Jafarpour, Orlitsky, Pan: Competitive Closeness Testing</i>	47
<i>Agarwal, Duchi, Bartlett, Levraud: Oracle inequalities for computationally budgeted model selection</i>	69
<i>Amin, Kearns, Syed: Bandits, Query Learning, and the Haystack Dimension</i>	87
<i>Audibert, Bubeck, Lugosi: Minimax Policies for Combinatorial Prediction Games</i>	107
<i>Bartok, Pal, Szepesvari: Minimax Regret of Finite Partial-Monitoring Games in Stochastic Environments</i>	133
<i>Chaudhuri, Hsu: Sample Complexity Bounds for Differentially Private Learning</i>	155
<i>Comminges, Dalalyan: Tight conditions for consistent variable selection in high dimensional nonparametric regression</i>	187
<i>Daniely, Sabato, Ben-David, Shalev-Shwartz: Multiclass Learnability and the ERM principle</i>	207

<i>Erven, Reid, Williamson: Mixability is Bayes Risk Curvature Relative to Log Loss</i>	<a href="#">233</a>
<i>Feldman: Distribution-Independent Evolvability of Linear Threshold Functions</i>	<a href="#">253</a>
<i>Feldman, Lee, Servedio: Lower Bounds and Hardness Amplification for Learning Shallow Monotone Formulas</i>	<a href="#">273</a>
<i>Foster, Rakhlin, Sridharan, Tewari: Complexity-Based Approach to Calibration with Checking Rules</i>	<a href="#">293</a>
<i>Foygel, Srebro: Concentration-Based Guarantees for Low-Rank Matrix Reconstruction</i>	<a href="#">315</a>
<i>Gao, Zhou: On the Consistency of Multi-Label Learning</i>	<a href="#">341</a>
<i>Garivier, Cappe: The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond</i>	<a href="#">359</a>
<i>Gerchinovitz: Sparsity Regret Bounds for Individual Sequences in Online Linear Regression</i>	<a href="#">377</a>
<i>Grunwald, Smith Jones, Winter, Smith: Safe Learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity</i>	<a href="#">397</a>
<i>Hazan, Kale: Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization</i>	<a href="#">421</a>
<i>Kallweit, Simon, Kallweit, Simon: A Close Look to Margin Complexity and Related Parameters</i>	<a href="#">437</a>
<i>Kotlowski, Grunwald: Maximum Likelihood vs. Sequential Normalized Maximum Likelihood in On-line Density Estimation</i>	<a href="#">457</a>
<i>Li, Zhang: A New Algorithm for Compressed Counting with Applications in Shannon Entropy Estimation in Dynamic Data</i>	<a href="#">477</a>
<i>Maillard, Munos, Stoltz: A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences</i>	<a href="#">497</a>

<i>Mannor, Perchet, Stoltz: Robust approachability and regret minimization in games with partial monitoring</i>	<a href="#">515</a>
<i>Mukherjee, Rudin, Schapire: The Rate of Convergence of Adaboost</i>	<a href="#">537</a>
<i>Rakhlin, Sridharan, Tewari: Online Learning: Beyond Regret</i>	<a href="#">559</a>
<i>Rigollet, Tong: Neyman-Pearson classification under a strict constraint</i>	<a href="#">595</a>
<i>Rudin, Letham, Salleb-Aouissi, Kogan, Madigan: Sequential Event Prediction with Association Rules</i>	<a href="#">615</a>
<i>Salmon, Dalalyan: Optimal aggregation of affine estimators</i>	<a href="#">635</a>
<i>Shamir, Shalev-Shwartz: Collaborative Filtering with the Trace Norm:Learning, Bounding, and Transducing</i>	<a href="#">661</a>
<i>Slivkins: Contextual Bandits with Similarity Information</i>	<a href="#">679</a>
<i>Steinwart: Adaptive Density Level Set Clustering</i>	<a href="#">703</a>
<i>Szita, Szepesvari: Agnostic KWIK learning and efficient approximate reinforcement learning</i>	<a href="#">739</a>
<i>Vainsencher, Mannor, Bruckstein: The Sample Complexity of Dictionary Learning</i>	<a href="#">773</a>
<i>Yang, Hanneke, Carbonell: Identifiability of Priors from Bounded Sample Sizes with Applications to Transfer Learning</i>	<a href="#">789</a>

---

---

## Part II Open problems

---

<i>Abernethy, Mannor: Does an Efficient Calibrated Forecasting Strategy Exist?</i>	<a href="#">809</a>
<i>Grunwald, Kotlowski: Bounds on Individual Risk for Log-loss Predictors</i>	<a href="#">813</a>
<i>Hazan, Kale: A simple multi-armed bandit algorithm with optimal variation-bounded regret</i>	<a href="#">817</a>
<i>Kotlowski, Warmuth: Minimax Algorithm for Learning Rotations</i>	<a href="#">821</a>
<i>Michael: Missing Information Impediments to Learnability</i>	<a href="#">825</a>
<i>Slivkins: Monotone multi-armed bandit allocations</i>	<a href="#">829</a>

# **Part I**

## **Regular papers**



# Regret Bounds for the Adaptive Control of Linear Quadratic Systems

**Yasin Abbasi-Yadkori**

**Csaba Szepesvari**

*Department of Computing Science, University of Alberta*

ABBASIYA@CS.UALBERTA.CA

SZEPESVA@CS.UALBERTA.CA

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We study the average cost Linear Quadratic (LQ) control problem with unknown model parameters, also known as the adaptive control problem in the control community. We design an algorithm and prove that apart from logarithmic factors its regret up to time  $T$  is  $O(\sqrt{T})$ . Unlike previous approaches that use a forced-exploration scheme, we construct a high-probability confidence set around the model parameters and design an algorithm that plays optimistically with respect to this confidence set. The construction of the confidence set is based on the recent results from online least-squares estimation and leads to improved worst-case regret bound for the proposed algorithm. To the best of our knowledge this is the first time that a regret bound is derived for the LQ control problem.

## 1. Introduction

We study the average cost LQ control problem with unknown model parameters, also known as the adaptive control problem in the control community. The problem is to minimize the average cost of a controller that operates in an environment whose dynamics is linear, while the cost is a quadratic function of the state and the control. The optimal solution is a linear feedback controller which can be computed in a closed form from the matrices underlying the dynamics and the cost. In the learning problem, the topic of this paper, the dynamics of the environment is unknown. This problem is challenging since the control actions influence both the cost and the rate at which the dynamics is learned, a topic of adaptive control. The objective in this case is to minimize the *regret* of the controller, i.e. to minimize the difference between the average cost incurred by the learning controller and that of the optimal controller. In this paper, for the first time, we show an adaptive controller and prove that, under some assumptions, its expected regret is bounded by  $\tilde{O}(\sqrt{T})$ . We build on recent works in online linear estimation and adaptive control design, the latter of which we survey next.

When the model parameters are known and the state is fully observed, one can use the principles of dynamic programming to obtain the optimal controller. The version of the problem that deals with the unknown model parameters is called the adaptive control problem. The early attempts to solve this problem relied on the *certainty equivalence principle* (Simon, 1956). The idea was to estimate the unknown parameters from observations and then use the estimated parameters as if they were the true parameters to design a controller. It was soon realized that the certainty equivalence principle does not necessarily provide enough information to reliably estimate the parameters and the estimated parameters can converge to incorrect values with positive probability (Becker et al., 1985). This in turn might lead to suboptimal performance.

To avoid non-identification problem, methods that actively explore the environment to gather information are developed (Lai and Wei, 1982a, 1987; Chen and Guo, 1987; Chen and Zhang, 1990; Fiechter, 1997; Lai and Ying, 2006; Campi and Kumar, 1998; Bittanti and Campi, 2006). However, only asymptotic results are proven for these methods. One exception is the work of Fiechter (1997) who proposes an algorithm for the “discounted” LQ problem and analyzes its performance in a PAC framework.

Most of the aforementioned methods use forced-exploration schemes to provide the sufficient exploratory information. The idea is to take exploratory actions according to a fixed and appropriately designed schedule. However, the forced-exploration schemes lack strong worst-case regret bounds, even in the simplest problems (see e.g. Dani and Hayes (2006), section 6). Unlike the preceding methods, Campi and Kumar (1998) proposes an algorithm that uses the Optimism in the Face of Uncertainty (OFU) principle, which goes back to the work of Lai and Robbins (1985), to deal with the exploration/exploitation dilemma. They call this the Bet On the Best (BOB) principle. The idea is to construct high-probability confidence sets around the model parameters, find the optimal controller for each member of the confidence set, and finally choose the controller whose associated average cost is the smallest. However, Campi and Kumar (1998) only show asymptotic optimality, i.e. the average cost of their algorithm converges to that of the optimal policy in the limit. In this paper, we modify the algorithm and the proof technique of Campi and Kumar (1998) and extend their work to derive a finite time regret bound. Our work also builds upon on the works of Lai and Wei (1982b); Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010) in analyzing the linear estimation with dependent covariates, although we use a more recent, improved confidence bound (see Theorem 1).

Note that the OFU principle has been applied very successfully to a number of challenging learning and control situations. Lai and Robbins (1985), who invented the principle, used it to address learning in bandit problems (i.e., when there is no state) and later this work was picked up and modified by Auer et al. (2002) to make it work in nonparametric bandits. The OFU principle has also been applied to learning in *finite* Markov Decision Processes, both in a regret minimization (e.g., Bartlett and Tewari 2009; Auer et al. 2010) and in a PAC-learning setting (e.g., Kearns and Singh 1998; Brafman and Tennenholtz 2002; Kakade 2003; Strehl et al. 2006; Szita and Szepesvári 2010). In the PAC-MDP framework there has been some work to extend the OFU principle to infinite Markov Decision Problems under various assumptions. For example, Lipschitz assumptions have been used

by Kakade et al. (2003), while Strehl and Littman (2008) explored linear models. However, none of these works consider both continuous state and action spaces. Continuous action spaces in the context of bandits have been explored in a number of works, such as the works of Kleinberg (2005); Auer et al. (2007); Kleinberg et al. (2008) and in a linear setting by Auer (2003); Dani et al. (2008) and Rusmevichientong and Tsitsiklis (2010).

## 2. Notation and conventions

We use  $\|\cdot\|$  and  $\|\cdot\|_F$  to denote the 2-norm and the Frobenius norm, respectively. For a positive semidefinite matrix  $A \in \mathbb{R}^{d \times d}$ , the weighted 2-norm  $\|\cdot\|_A$  is defined by  $\|x\|_A^2 = x^\top Ax$ , where  $x \in \mathbb{R}^d$ . The inner product is denoted by  $\langle \cdot, \cdot \rangle$ . We use  $\lambda_{\min}(A)$  and  $\lambda_{\max}(A)$  to denote the minimum and maximum eigenvalues of the positive semidefinite matrix  $A$ , respectively. We use  $A \succ 0$  to denote that  $A$  is positive definite, while we use  $A \succeq 0$  to denote that it is positive semidefinite. We use  $\mathbb{I}_{\{A\}}$  to denote the indicator function of event  $A$ .

## 3. The Linear Quadratic Problem

We consider the discrete-time infinite-horizon linear quadratic (LQ) control problem:

$$\begin{aligned} x_{t+1} &= A_* x_t + B_* u_t + w_{t+1}, \\ c_t &= x_t^\top Q x_t + u_t^\top R u_t, \end{aligned}$$

where  $t = 0, 1, \dots$ ,  $u_t \in \mathbb{R}^d$  is the control at time  $t$ ,  $x_t \in \mathbb{R}^n$  is the state at time  $t$ ,  $c_t \in \mathbb{R}$  is the cost at time  $t$ ,  $w_{t+1}$  is the ‘‘noise’’,  $A_* \in \mathbb{R}^{n \times n}$  and  $B_* \in \mathbb{R}^{n \times d}$  are unknown matrices while  $Q \in \mathbb{R}^{n \times n}$  and  $R \in \mathbb{R}^{d \times d}$  are known (positive definite) matrices. At time zero, for simplicity,  $x_0 = 0$ . The problem is to design a controller based on past observations to minimize the average expected cost

$$J(u_0, u_1, \dots) = \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T \mathbb{E}[c_t]. \quad (1)$$

Let  $J_*$  be the optimal (lowest) average cost. The *regret* up to time  $T$  of a controller which incurs a cost of  $c_t$  at time  $t$  is defined by

$$R(T) = \sum_{t=0}^T (c_t - J_*),$$

which is the difference between the performance of the controller and the performance of the optimal controller that has full information about the system dynamics. Thus the regret can be interpreted as a measure of the cost of not knowing the system dynamics.

### 3.1. Assumptions

In this section, we state our assumptions on the noise and the system dynamics. In particular, we assume that the noise is sub-Gaussian and the system is controllable and observable<sup>1</sup>. Define

$$\Theta_*^\top = (A_*, \quad B_*) \quad \text{and} \quad z_t = \begin{pmatrix} x_t \\ u_t \end{pmatrix}.$$

Thus, the state transition can be written as

$$x_{t+1} = \Theta_*^\top z_t + w_{t+1}.$$

**Assumption A1** There exists a filtration  $(\mathcal{F}_t)$  such that for the random variables  $(z_0, x_1), \dots, (z_t, x_{t+1})$ , the following hold:

(i)  $z_t, x_t$  are  $\mathcal{F}_t$ -measurable;

(ii) For any  $t \geq 0$ ,

$$\mathbb{E}[x_{t+1} | \mathcal{F}_t] = z_t^\top \Theta_*,$$

i.e.,  $w_{t+1} = x_{t+1} - z_t^\top \theta_*$  is a martingale difference sequence ( $\mathbb{E}[w_{t+1} | \mathcal{F}_t] = 0$ ,  $t = 0, 1, \dots$ );

(iii)  $\mathbb{E}[w_{t+1} w_{t+1}^\top | \mathcal{F}_t] = I_n$ ;

(iv) The random variables  $w_t$  are component-wise sub-Gaussian in the sense that there exists  $L > 0$  such that for any  $\gamma \in \mathbb{R}$ , and index  $j$ ,

$$\mathbb{E}[\exp(\gamma w_{t+1,j}) | \mathcal{F}_t] \leq \exp(\gamma^2 L^2 / 2).$$

The assumption  $\mathbb{E}[w_{t+1} w_{t+1}^\top | \mathcal{F}_t] = I_n$  makes the analysis more readable. However, we shall show it in Section 4 that it is in fact not necessary. Our next assumption on the system uncertainty states that the unknown parameter is a member of a bounded set and is such that the system is controllable and observable. This assumption will let us derive a closed form expression for the optimal control law.

**Assumption A2** The unknown parameter  $\Theta_*$  is a member of set  $\mathcal{S}$  such that

$$\mathcal{S} \subseteq \mathcal{S}_0 \cap \left\{ \Theta \in \mathbb{R}^{n \times (n+d)} \mid \text{trace}(\Theta^\top \Theta) \leq S^2 \right\},$$

where

$$\mathcal{S}_0 = \left\{ \Theta = (A, B) \in \mathbb{R}^{n \times (n+d)} \mid \begin{array}{l} (A, B) \text{ is controllable,} \\ (A, M) \text{ is observable, where } Q = M^\top M \end{array} \right\}.$$

In what follows, we shall always assume that the above assumptions are valid.

---

1. Controllability and observability are defined in Appendix B

### 3.2. Parameter estimation

In order to implement the OFU principle, we need high-probability confidence sets for the unknown parameter matrix. The derivation of the confidence set is based on results from Abbasi-Yadkori et al. (2011) that use techniques from self-normalized processes to estimate the least squares estimation error. Define

$$e(\Theta) = \lambda \text{trace}(\Theta^\top \Theta) + \sum_{s=0}^{t-1} \text{trace}((x_{s+1} - \Theta^\top z_s)(x_{s+1} - \Theta^\top z_s)^\top).$$

Let  $\hat{\Theta}_t$  be the  $\ell^2$ -regularized least-squares estimate of  $\Theta_*$  with regularization parameter  $\lambda > 0$ :

$$\hat{\Theta}_t = \underset{\Theta}{\operatorname{argmin}} e(\Theta) = (Z^\top Z + \lambda I)^{-1} Z^\top X, \quad (2)$$

where  $Z$  and  $X$  are the matrices whose rows are  $z_0^\top, \dots, z_{t-1}^\top$  and  $x_1^\top, \dots, x_t^\top$ , respectively.

**Theorem 1** *Let  $(z_0, x_1), \dots, (z_t, x_{t+1})$ ,  $z_i \in \mathbb{R}^{n+d}$ ,  $x_i \in \mathbb{R}^n$  satisfy the linear model Assumption A1 with some  $L > 0$ ,  $\Theta_* \in \mathbb{R}^{(n+d) \times n}$ ,  $\text{trace}(\Theta_*^\top \Theta_*) \leq S^2$  and let  $\mathcal{F} = (\mathcal{F}_t)$  be the associated filtration. Consider the  $\ell^2$ -regularized least-squares parameter estimate  $\hat{\Theta}_t$  with regularization coefficient  $\lambda > 0$  (cf. (2)). Let*

$$V_t = \lambda I + \sum_{i=0}^{t-1} z_i z_i^\top$$

*be the regularized design matrix underlying the covariates. Define*

$$\beta_t(\delta) = \left( nL \sqrt{2 \log \left( \frac{\det(V_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + \lambda^{1/2} S \right)^2. \quad (3)$$

*Then, for any  $0 < \delta < 1$ , with probability at least  $1 - \delta$ ,*

$$\text{trace}((\hat{\Theta}_t - \Theta_*)^\top V_t (\hat{\Theta}_t - \Theta_*)) \leq \beta_t(\delta).$$

*In particular,  $\mathbb{P}(\Theta_* \in \mathcal{C}_t(\delta), t = 1, 2, \dots) \geq 1 - \delta$ , where*

$$\mathcal{C}_t(\delta) = \left\{ \Theta \in \mathbb{R}^{n \times (n+d)} : \text{trace} \left\{ (\Theta - \hat{\Theta}_t)^\top V_t (\Theta - \hat{\Theta}_t) \right\} \leq \beta_t(\delta) \right\}.$$

### 3.3. The design of the controller

Let  $(A, B) = \Theta \in \mathcal{S}_0$ , where  $\mathcal{S}_0$  is defined in Assumption A2. Then there is a unique solution  $P(\Theta)$  in the class of positive semidefinite symmetric matrices to the *Riccati equation*

$$P(\Theta) = Q + A^\top P(\Theta) A - A^\top P(\Theta) B (B^\top P(\Theta) B + R)^{-1} B^\top P(\Theta) A.$$

Under the same assumptions, the matrix  $A + BK(\Theta)$  is stable, i.e. its norm-2 is less than one, where

$$K(\Theta) = -(B^\top P(\Theta)B + R)^{-1}B^\top P(\Theta)A$$

is the *gain matrix* (Bertsekas, 2001). Further, by the boundedness of  $\mathcal{S}$ , we also obtain the boundedness of  $P(\Theta)$  (Anderson and Moore, 1971). The corresponding constant will be denoted by  $D$ :

$$D = \sup \{\|P(\Theta)\| \mid \Theta \in \mathcal{S}\}. \quad (4)$$

The *optimal control law* for a LQ system with parameters  $\Theta$  is

$$u_t = K(\Theta)x_t, \quad (5)$$

i.e., this controller achieves the optimal average cost which satisfies  $J(\Theta) = \text{trace}(P(\Theta))$  (Bertsekas, 2001). In particular, the average cost of control law (5) with  $\Theta = \Theta_*$  is the optimal average cost  $J_* = J(\Theta_*) = \text{trace}(P(\Theta_*))$ .

We assume that the bound on the norm of the unknown parameter,  $S$ , and the sub-Gaussianity constant,  $L$ , are known:

**Assumption A3** Constants  $L$  and  $S$  in Assumptions A1 and A2 are known.

The algorithm that we propose implements the OFU principle as follows: At time  $t$ , the algorithm chooses a parameter  $\tilde{\Theta}_t$  from  $\mathcal{C}_t(\delta) \cap \mathcal{S}$  such that

$$J(\tilde{\Theta}_t) \leq \inf_{\Theta \in \mathcal{C}_t(\delta) \cap \mathcal{S}} J(\Theta) + \frac{1}{\sqrt{t}}$$

and then uses the optimal feedback controller (5) underlying the chosen parameter. In order to prevent too frequent changes to the controller (which might harm performance), the algorithm changes controllers only after the current parameter estimates are significantly refined. The details of the algorithm are given in Algorithm 1.

## 4. Analysis

In this section we give our main result together with its proof. Before stating the main theorem, we make one more assumption in addition to the assumptions we made before.

**Assumption A4** The set  $\mathcal{S}$  is such that  $\rho := \sup_{(A,B) \in \mathcal{S}} \|A + BK(A, B)\| < 1$ . Further, there exists a positive number  $C$  such that  $C = \sup_{\Theta \in \mathcal{S}} \|K(\Theta)\| < \infty$ .

Our main result is the following theorem:

```

Inputs:  $T, S > 0, \delta > 0, Q, L, \lambda > 0$ .
Set  $V_0 = \lambda I$  and  $\hat{\Theta}_0 = 0$ .
 $(\tilde{A}_0, \tilde{B}_0) = \tilde{\Theta}_0 = \operatorname{argmin}_{\Theta \in \mathcal{C}_0(\delta) \cap S} J(\Theta)$ .
for  $t := 0, 1, 2, \dots$  do
    if  $\det(V_t) > 2 \det(V_0)$  then
        Calculate  $\hat{\Theta}_t$  by (2).
        Find  $\tilde{\Theta}_t$  such that  $J(\tilde{\Theta}_t) \leq \inf_{\Theta \in \mathcal{C}_t(\delta) \cap S} J(\Theta) + \frac{1}{\sqrt{t}}$ .
        Let  $V_0 = V_t$ .
    else
         $\tilde{\Theta}_t = \tilde{\Theta}_{t-1}$ .
    end if
    Calculate  $u_t$  based on the current parameters,  $u_t = K(\tilde{\Theta}_t)x_t$ .
    Execute control, observe new state  $x_{t+1}$ .
    Save  $(z_t, x_{t+1})$  into the dataset, where  $z_t^\top = (x_t^\top, u_t^\top)$ .
     $V_{t+1} := V_t + z_t z_t^\top$ .
end for

```

Table 1: The proposed adaptive algorithm for the LQ problem

**Theorem 2** For any  $0 < \delta < 1$ , for any time  $T$ , with probability at least  $1 - \delta$ , the regret of Algorithm 1 is bounded as follows:

$$R(T) = \tilde{O}\left(\sqrt{T \log(1/\delta)}\right),$$

where the constant hidden is a problem dependent constant.<sup>2</sup>

**Remark 3** The assumption  $\mathbb{E}[w_{t+1} w_{t+1}^\top | \mathcal{F}_t] = I_n$  makes the analysis more readable. Alternatively, we could assume that  $\mathbb{E}[w_{t+1} w_{t+1}^\top | \mathcal{F}_t] = G_*$  and  $G_*$  be unknown. Then the optimal average cost becomes  $J(\Theta_*, G_*) = \text{trace}(P(\Theta_*)G_*)$ . The only change in Algorithm 1 is in the computation of  $\tilde{\Theta}_t$ , which will have the following form:

$$(\tilde{\Theta}_t, \tilde{G}) = \operatorname{argmin}_{(\Theta, G) \in \mathcal{C}_t} J(\Theta),$$

where  $\mathcal{C}_t$  is now a confidence set over  $\Theta_*$  and  $G_*$ . The rest of the analysis remains identical, provided that an appropriate confidence set is constructed.

The least squares estimation error from Theorem 1 scales with the size of the state and action vectors. Thus, in order to prove Theorem 2, we first prove a high-probability bound on the norm of the state vector. Given the boundedness of the state, we decompose the regret and analyze each term using appropriate concentration inequalities.

---

2. Here,  $\tilde{O}$  hides logarithmic factors.

#### 4.1. Bounding $\|x_t\|$

We choose an error probability,  $\delta > 0$ . Given this, we define two “good events” in the probability space  $\Omega$ . In particular, we define the event that the confidence sets hold for  $s = 0, \dots, t$ ,

$$E_t = \{\omega \in \Omega : \forall s \leq t, \Theta_* \in \mathcal{C}_s(\delta/4)\},$$

and the event that the state vector stays “small”:

$$F_t = \{\omega \in \Omega : \forall s \leq t, \|x_s\| \leq \alpha_t\}$$

where

$$\begin{aligned} \alpha_t &= \frac{1}{1-\rho} \left( \frac{\eta}{\rho} \right)^{n+d} \left[ G Z_T^{\frac{n+d}{n+d+1}} \beta_t(\delta/4)^{\frac{1}{2(n+d+1)}} + 2L \sqrt{n \log \frac{4nt(t+1)}{\delta}} \right], \\ \eta &= 1 \vee \sup_{\Theta \in \mathcal{S}} \|A_* + B_* K(\Theta)\|, \\ Z_T &= \max_{0 \leq t \leq T} \|z_t\|, \\ G &= 2 \left( \frac{2S(n+d)^{n+d+1/2}}{U^{1/2}} \right)^{1/(n+d+1)}, \\ U &= \frac{U_0}{H}, \\ U_0 &= \frac{1}{16^{n+d-2} (1 \vee S^{2(n+d-2)})}, \end{aligned}$$

and  $H$  is any number satisfying<sup>3</sup>

$$H > \left( 16 \vee \frac{4S^2 M^2}{(n+d)U_0} \right),$$

where

$$M = \sup_{Y \geq 0} \frac{\left( nL \sqrt{(n+d) \log \left( \frac{1+TY/\lambda}{\delta} \right)} + \lambda^{1/2} S \right)}{Y}.$$

In what follows, we let  $E = E_T$  and  $F = F_T$ . First, we show that  $E \cap F$  holds with high probability and on  $E \cap F$ , the state vector does not explode.

**Lemma 4**  $\mathbb{P}(E \cap F) \geq 1 - \delta/2$ .

The proof is in Appendix D. It first shows that  $\|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|$  is well controlled except for a small number of occasions. Given this and proper decomposition of the state update equation, we can prove that the state vector  $x_t$  stays smaller than  $\alpha_t$ . Notice that  $\alpha_t$  itself depends  $\beta_t$  and  $Z_T$ , which in turn depend on  $x_t$ . Thus, we need one more step to have a bound on  $\|x_t\|$ .

---

3. We use  $\wedge$  and  $\vee$  to denote the minimum and the maximum, respectively.

**Lemma 5** *For appropriate problem dependent constants  $C_1 > 0, C_2 > 0$  (which are independent of  $t, \delta, T$ ), for any  $t \geq 0$ , it holds that  $\mathbb{I}_{\{F_t\}} \max_{1 \leq s \leq t} \|x_s\| \leq X_t$ , where*

$$X_t = Y_t^{n+d+1}$$

and

$$Y_t \stackrel{\text{def}}{=} (e \vee \lambda(n+d)(e-1) \vee 4(C_1 \log(1/\delta) + C_2 \log(t/\delta)) \log^2(4(C_1 \log(1/\delta) + C_2 \log(t/\delta)))) .$$

**Proof** Fix  $t$ . On  $F_t$ ,  $\hat{X}_t := \max_{0 \leq s \leq t} \|x_s\| \leq \alpha_t$ . With appropriate constants, this implies that

$$x \leq D_1 \sqrt{\beta_t(\delta)} \log(t) x^{\frac{n+d}{n+d+1}} + D_2 \sqrt{\log \frac{t}{\delta}},$$

or

$$x \leq \left( D_1 \sqrt{\beta_t(\delta)} \log(t) + D_2 \sqrt{\log \frac{t}{\delta}} \right)^{n+d+1}, \quad (6)$$

holds for  $x = \hat{X}_t$ . Let  $X_t$  be the largest value of  $x \geq 0$  that satisfies (6). Thus,

$$X_t \leq \left( D_1 \sqrt{\beta_t(\delta)} \log(t) + D_2 \sqrt{\log \frac{t}{\delta}} \right)^{n+d+1}, \quad (7)$$

Clearly,  $\hat{X}_t \leq X_t$ . Because  $\beta_t(\delta)$  is a function of  $\log \det(V_t)$ , (7) has the form of

$$X_t \leq f(\log(X_t))^{n+d+1}. \quad (8)$$

Let  $a_t = X_t^{1/(n+d+1)}$ . Then, (8) is equivalent to

$$a_t \leq f(\log a_t^{n+d+1}) = f((n+d+1) \log a_t).$$

Let  $c = \max(1, \max_{1 \leq s \leq t} \|a_s\|)$ . Assume that  $t \geq \lambda(n+d)$ . By the construction of  $F_t$ , Lemma 10, tedious, but elementary calculations, it can then be shown that

$$c \leq A \log^2(c) + B_t, \quad (9)$$

where  $A = G_1 \log(1/\delta)$  and  $B_t = G_2 \log(t/\delta)$ . From this, further elementary calculations show that the maximum value that  $c$  can take on subject to the constraint (9) is bounded from above by  $Y_t$ . ■

## 4.2. Regret Decomposition

From the Bellman optimality equations for the LQ problem, we get that (Bertsekas, 1987)[V. 2, p. 228–229]

$$\begin{aligned} J(\tilde{\Theta}_t) + x_t^\top P(\tilde{\Theta}_t) x_t &= \min_u \left\{ x_t^\top Q x_t + u^\top R u + \mathbb{E} \left[ \tilde{x}_{t+1}^{uT} P(\tilde{\Theta}_t) \tilde{x}_{t+1}^u \middle| \mathcal{F}_t \right] \right\} \\ &= x_t^\top Q x_t + u_t^\top R u_t + \mathbb{E} \left[ \tilde{x}_{t+1}^{u_t T} P(\tilde{\Theta}_t) \tilde{x}_{t+1}^{u_t} \middle| \mathcal{F}_t \right], \end{aligned}$$

where  $\tilde{x}_{t+1}^u = \tilde{A}_t x_t + \tilde{B}_t u + w_{t+1}$  and  $(\tilde{A}_t, \tilde{B}_t) = \tilde{\Theta}_t$ . Hence,

$$\begin{aligned}
 J(\tilde{\Theta}_t) + x_t^\top P(\tilde{\Theta}_t)x_t &= x_t^\top Qx_t + u_t^\top Ru_t \\
 &\quad + \mathbb{E} \left[ (\tilde{A}_t x_t + \tilde{B}_t u_t + w_{t+1})^\top P(\tilde{\Theta}_t)(\tilde{A}_t x_t + \tilde{B}_t u_t + w_{t+1}) \middle| \mathcal{F}_t \right] \\
 &= x_t^\top Qx_t + u_t^\top Ru_t + \mathbb{E} \left[ (\tilde{A}_t x_t + \tilde{B}_t u_t)^\top P(\tilde{\Theta}_t)(\tilde{A}_t x_t + \tilde{B}_t u_t) \middle| \mathcal{F}_t \right] \\
 &\quad + \mathbb{E} \left[ w_{t+1}^\top P(\tilde{\Theta}_t)w_{t+1} \middle| \mathcal{F}_t \right] \\
 &= x_t^\top Qx_t + u_t^\top Ru_t + \mathbb{E} \left[ (\tilde{A}_t x_t + \tilde{B}_t u_t)^\top P(\tilde{\Theta}_t)(\tilde{A}_t x_t + \tilde{B}_t u_t) \middle| \mathcal{F}_t \right] \\
 &\quad + \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_t)x_{t+1} \middle| \mathcal{F}_t \right] \\
 &\quad - \mathbb{E} \left[ (A_* x_t + B_* u_t)^\top P(\tilde{\Theta}_t)(A_* x_t + B_* u_t) \middle| \mathcal{F}_t \right] \\
 &= x_t^\top Qx_t + u_t^\top Ru_t + \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_t)x_{t+1} \middle| \mathcal{F}_t \right] \\
 &\quad + (\tilde{A}_t x_t + \tilde{B}_t u_t)^\top P(\tilde{\Theta}_t)(\tilde{A}_t x_t + \tilde{B}_t u_t) \\
 &\quad - (A_* x_t + B_* u_t)^\top P(\tilde{\Theta}_t)(A_* x_t + B_* u_t),
 \end{aligned}$$

where in the one before last equality we have used  $x_{t+1} = A_* x_t + B_* u_t + w_{t+1}$  and the martingale property of the noise. Hence,

$$\sum_{t=0}^T J(\tilde{\Theta}_t) + R_1 = \sum_{t=0}^T (x_t^\top Qx_t + u_t^\top Ru_t) + R_2 + R_3,$$

where

$$R_1 = \sum_{t=0}^T \left\{ x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E} \left[ x_{t+1}^\top P(\tilde{\Theta}_{t+1})x_{t+1} \middle| \mathcal{F}_t \right] \right\}, \quad (10)$$

$$R_2 = \sum_{t=0}^T \mathbb{E} \left[ x_{t+1}^\top (P(\tilde{\Theta}_t) - P(\tilde{\Theta}_{t+1}))x_{t+1} \middle| \mathcal{F}_t \right], \quad (11)$$

and

$$R_3 = \sum_{t=0}^T \left( (\tilde{A}_t x_t + \tilde{B}_t u_t)^\top P(\tilde{\Theta}_t)(\tilde{A}_t x_t + \tilde{B}_t u_t) - (A_* x_t + B_* u_t)^\top P(\tilde{\Theta}_t)(A_* x_t + B_* u_t) \right). \quad (12)$$

Thus, on  $E \cap F$ ,

$$\begin{aligned}
 \sum_{t=0}^T (x_t^\top Qx_t + u_t^\top Ru_t) &= \sum_{t=0}^T J(\tilde{\Theta}_t) + R_1 - R_2 - R_3 \\
 &\leq TJ(\Theta_*) + R_1 - R_2 - R_3 + 2\sqrt{T},
 \end{aligned}$$

where the last inequality follows from the choice of  $\tilde{\Theta}_t$  and the fact that on  $E$ ,  $\Theta_* \in C_t(\delta)$ . Thus, on  $E \cap F$ ,

$$R(T) \leq R_1 - R_2 - R_3 + 2\sqrt{T}. \quad (13)$$

In the following subsections, we bound  $R_1, R_2$ , and  $R_3$ .

### 4.3. Bounding $\mathbb{I}_{\{E \cap F\}} R_1$

We start by showing that with high probability all noise terms are small.

**Lemma 6** *With probability  $1 - \delta/8$ , for any  $k \leq T$ ,  $\|w_k\| \leq Ln\sqrt{2n \log(8nT/\delta)}$ .*

**Proof** From sub-Gaussianity Assumption A1, we have that for any index  $1 \leq i \leq n$  and any time  $k$ ,

$$|w_{k,i}| \leq L\sqrt{2 \log(8/\delta)}.$$

Thus, with probability  $1 - \delta/8$ , for any  $k \leq T$ ,  $\|w_k\| \leq Ln\sqrt{2n \log(8nT/\delta)}$ .  $\blacksquare$

**Lemma 7** *Let  $R_1$  be as defined by (10). With probability at least  $1 - \delta/2$ ,*

$$\mathbb{I}_{\{E \cap F\}} R_1 \leq 2DW^2 \sqrt{2T \log \frac{8}{\delta}} + n\sqrt{B'_\delta},$$

where  $W = Ln\sqrt{2n \log(8nT/\delta)}$  and

$$B'_\delta = (v + TD^2S^2X^2(1 + C^2)) \log \left( \frac{4nv^{-1/2}}{\delta} (v + TD^2S^2X^2(1 + C^2))^{1/2} \right).$$

**Proof** Let  $f_{t-1} = A_*x_{t-1} + B_*u_{t-1}$  and  $P_t = P(\tilde{\Theta}_t)$ . Write

$$\begin{aligned} R_1 &= x_0^\top P(\tilde{\Theta}_0)x_0 - x_{T+1}^\top P(\tilde{\Theta}_{T+1})x_{T+1} \\ &\quad + \sum_{t=1}^T \left( x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E} \left[ x_t^\top P(\tilde{\Theta}_t)x_t | \mathcal{F}_{t-1} \right] \right). \end{aligned}$$

Because  $P$  is positive semi-definite and  $x_0 = 0$ , the first term is bounded by zero. The second term can be decomposed as follows

$$\begin{aligned} \sum_{t=1}^T \left( x_t^\top P_t x_t - \mathbb{E} \left[ x_t^\top P_t x_t | \mathcal{F}_{t-1} \right] \right) &= \sum_{t=1}^T f_{t-1}^\top P_t w_t \\ &\quad + \sum_{t=1}^T \left( w_t^\top P_t w_t - \mathbb{E} \left[ w_t^\top P_t w_t | \mathcal{F}_{t-1} \right] \right). \end{aligned}$$

We bound each term separately. Let  $v_t^\top = f_{t-1}^\top P_t$  and

$$G_1 = \mathbb{I}_{\{E \cap F\}} \sum_{t=1}^T v_t^\top w_t = \mathbb{I}_{\{E \cap F\}} \sum_{t=1}^T \sum_{i=1}^n v_{k,i} w_{k,i} = \sum_{i=1}^n \mathbb{I}_{\{E \cap F\}} \sum_{t=1}^T v_{k,i} w_{k,i}.$$

Let  $M_{T,i} = \sum_{t=1}^T v_{k,i} w_{k,i}$ . By Theorem 16, on some event  $G_{\delta,i}$  that holds with probability at least  $1 - \delta/(4n)$ , for any  $T \geq 0$ ,

$$M_{T,i}^2 \leq 2R^2 \left( v + \sum_{t=1}^T v_{t,i}^2 \right) \log \left( \frac{4nv^{-1/2}}{\delta} \left( v + \sum_{t=1}^T v_{t,i}^2 \right)^{1/2} \right) = B_{\delta,i}.$$

On  $E \cap F$ ,  $\|v_t\| \leq DSX\sqrt{1+C^2}$  and thus,  $v_{t,i} \leq DSX\sqrt{1+C^2}$ . Thus, on  $G_{\delta,i}$ ,  $\mathbb{I}_{\{E \cap F\}} M_{t,i}^2 \leq B'_{\delta,i}$ . Thus, we have  $G_1 \leq \sum_{i=1}^n \sqrt{B'_{\delta,i}}$  on  $\cap_{i=1}^n G_{\delta,i}$ , that holds w.p.  $1 - \delta/4$ .

Define  $X_t = w_t^\top P_t w_t - \mathbb{E}[w_t^\top P_t w_t | \mathcal{F}_{t-1}]$  and its truncated version  $\tilde{X}_t = X_t \mathbb{I}_{\{X_t \leq 2DW^2\}}$ . Define  $G_2 = \sum_{t=1}^T X_t$  and  $\tilde{G}_2 = \sum_{t=1}^T \tilde{X}_t$ . By Lemma 14,

$$\mathbb{P}\left(G_2 > 2DW^2 \sqrt{2T \log \frac{8}{\delta}}\right) \leq \mathbb{P}\left(\max_{1 \leq t \leq T} X_t \geq 2DW^2\right) + \mathbb{P}\left(\tilde{G}_2 > 2DW^2 \sqrt{2T \log \frac{8}{\delta}}\right).$$

By Lemma 6 and Azuma's inequality, each term on the right hand side is bounded by  $\delta/8$ . Thus, w.p.  $1 - \delta/4$ ,

$$G_2 \leq 2DW^2 \sqrt{2T \log \frac{8}{\delta}}$$

Summing up the bounds on  $G_1$  and  $G_2$  gives the result that holds w.p. at least  $1 - \delta/2$ ,

$$\mathbb{I}_{\{E \cap F\}} R_1 \leq 2DW^2 \sqrt{2T \log \frac{8}{\delta}} + n \sqrt{B'_\delta}.$$

■

#### 4.4. Bounding $\mathbb{I}_{\{E \cap F\}} |R_2|$

We can bound  $\mathbb{I}_{\{E \cap F\}} |R_2|$  by simply showing that Algorithm 1 rarely changes the policy, and hence most terms in (11) are zero.

**Lemma 8** *On the event  $E \cap F$ , Algorithm 1 changes the policy at most*

$$(n+d) \log_2 (1 + TX_T^2(1+C^2)/\lambda)$$

*times up to time  $T$ .*

**Proof** If we have changed the policy  $K$  times up to time  $T$ , then we should have that  $\det(V_T) \geq \lambda^{n+d} 2^K$ . On the other hand, we have

$$\lambda_{\max}(V_T) \leq \lambda + \sum_{t=0}^{T-1} \|z_t\|^2 \leq \lambda + TX_T^2(1+C^2),$$

where  $C$  is the bound on the norm of  $K(\cdot)$  as defined in Assumption A4. Thus, it holds that

$$\lambda^{n+d} 2^K \leq (\lambda + TX_T^2(1+C^2))^{n+d}.$$

Solving for  $K$ , we get

$$K \leq (n+d) \log_2 \left( 1 + \frac{TX_T^2(1+C^2)}{\lambda} \right).$$

■

**Lemma 9** *Let  $R_2$  be as defined by Equation (11). Then we have*

$$\mathbb{I}_{\{E \cap F\}} |R_2| \leq 2DX_T^2(n+d) \log_2 (1 + TX_T^2(1+C^2)/\lambda).$$

**Proof** On event  $E \cap F$ , we have at most  $K = (n+d) \log_2 (1 + TX_T^2(1+C^2)/\lambda)$  policy changes up to time  $T$ . So at most  $K$  terms in the summation (11) are non-zero. Each term in the summation is bounded by  $2DX_T^2$ . Thus,

$$\mathbb{I}_{\{E \cap F\}} |R_2| \leq 2DX_T^2(n+d) \log_2 (1 + TX_T^2(1+C^2)/\lambda).$$

■

#### 4.5. Bounding $\mathbb{I}_{\{E \cap F\}} |R_3|$

The summation  $\sum_{t=0}^T \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2$  will appear in the analysis while bounding  $|R_3|$ . So we first bound this summation, whose analysis requires the following two results.

**Lemma 10** *The following holds for any  $t \geq 1$ :*

$$\sum_{k=0}^{t-1} \left( \|z_k\|_{V_k^{-1}}^2 \wedge 1 \right) \leq 2 \log \frac{\det(V_t)}{\det(\lambda I)}.$$

Further, when the covariates satisfy  $\|z_t\| \leq c_m$ ,  $t \geq 0$  with some  $c_m > 0$  w.p.1 then

$$\log \frac{\det(V_t)}{\det(\lambda I)} \leq (n+d) \log \left( \frac{\lambda(n+d) + tc_m^2}{\lambda(n+d)} \right).$$

The proof of the lemma can be found in Abbasi-Yadkori et al. (2011).

**Lemma 11** *Let  $A \in \mathbb{R}^{m \times m}$  and  $B \in \mathbb{R}^{m \times m}$  be positive semi-definite matrices such that  $A \succeq B$ . Then, we have*

$$\sup_{X \neq 0} \frac{\|X^\top AX\|}{\|X^\top BX\|} \leq \frac{\det(A)}{\det(B)}.$$

The proof of this lemma is in Appendix C.

**Lemma 12** *On  $E \cap F$ , it holds that*

$$\sum_{t=0}^T \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 \leq \frac{16}{\lambda} (1 + C^2) X_T^2 \beta_T(\delta/4) \log \frac{\det(V_T)}{\det(\lambda I)}.$$

**Proof** Consider timestep  $t$ . Let  $s_t = (\Theta_* - \tilde{\Theta}_t)^\top z_t$ . Let  $\tau \leq t$  be the last timestep when the policy is changed. So  $s_t = (\Theta_* - \tilde{\Theta}_\tau)^\top z_t$ . We have

$$\|s_t\| \leq \|(\Theta_* - \hat{\Theta}_\tau)^\top z_t\| + \|(\hat{\Theta}_\tau - \tilde{\Theta}_\tau)^\top z_t\|. \quad (14)$$

For all  $\Theta \in \mathcal{C}_\tau$ ,

$$\begin{aligned} \|(\Theta - \hat{\Theta}_\tau)^\top z_t\| &\leq \|V_t^{1/2}(\Theta - \hat{\Theta}_\tau)\| \|z_t\|_{V_t^{-1}} && \text{(Cauchy-Schwartz inequality)} \\ &\leq \|V_\tau^{1/2}(\Theta - \hat{\Theta}_\tau)\| \sqrt{\frac{\det(V_t)}{\det(V_\tau)}} \|z_t\|_{V_t^{-1}} && \text{(Lemma 11)} \\ &\leq \sqrt{2} \|V_\tau^{1/2}(\Theta - \hat{\Theta}_\tau)\| \|z_t\|_{V_t^{-1}} && \text{(Choice of } \tau\text{)} \\ &\leq \sqrt{2\beta_\tau(\delta/4)} \|z_t\|_{V_t^{-1}}, && (\lambda_{\max}(M) \leq \text{trace}(M) \text{ for } M \succeq 0) \end{aligned}$$

Applying the last inequality to  $\Theta_*$  and  $\tilde{\Theta}_\tau$ , together with (14) gives

$$\|s_t\|^2 \leq 8\beta_\tau(\delta/4) \|z_t\|_{V_t^{-1}}^2.$$

Now, by Assumption A4 and the fact that  $\tilde{\Theta}_t \in \mathcal{S}$  we have that

$$\|z_t\|_{V_t^{-1}}^2 \leq \frac{\|z_t\|^2}{\lambda} \leq \frac{(1 + C^2) X_T^2}{\lambda}.$$

It follows then that

$$\begin{aligned} \sum_{t=0}^T \|s_t\|^2 &\leq \frac{8}{\lambda} (1 + C^2) X_T^2 \beta_T(\delta/4) \sum_{t=0}^T (\|z_t\|_{V_t^{-1}}^2 \wedge 1) \\ &\leq \frac{16}{\lambda} (1 + C^2) X_T^2 \beta_T(\delta/4) \log \frac{\det(V_T)}{\det(\lambda I)}. && \text{(Lemma 10).} \end{aligned}$$

■

Now, we are ready to bound  $R_3$ .

**Lemma 13** *Let  $R_3$  be as defined by Equation (12). Then we have*

$$\mathbb{I}_{\{E \cap F\}} |R_3| \leq \frac{8}{\sqrt{\lambda}} (1 + C^2) X_T^2 SD \left( \beta_T(\delta/4) \log \frac{\det(V_T)}{\det(\lambda I)} \right)^{1/2} \sqrt{T}.$$

**Proof** We have that

$$\begin{aligned}
 \mathbb{I}_{\{E \cap F\}} |R_3| &\leq \mathbb{I}_{\{E \cap F\}} \sum_{t=0}^T \left| \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\|^2 - \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\|^2 \right| \quad (\text{Tri. ineq.}) \\
 &\leq \mathbb{I}_{\{E \cap F\}} \left( \sum_{t=0}^T \left( \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\| - \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \quad (\text{C.-S. ineq.}) \\
 &\quad \times \left( \sum_{t=0}^T \left( \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\
 &\leq \mathbb{I}_{\{E \cap F\}} \left( \sum_{t=0}^T \left\| P(\tilde{\Theta}_t)^{1/2} (\tilde{\Theta}_t - \Theta_*)^\top z_t \right\|^2 \right)^{1/2} \quad (\text{Tri. ineq.}) \\
 &\quad \times \left( \sum_{t=0}^T \left( \left\| P(\tilde{\Theta}_t)^{1/2} \tilde{\Theta}_t^\top z_t \right\| + \left\| P(\tilde{\Theta}_t)^{1/2} \Theta_*^\top z_t \right\| \right)^2 \right)^{1/2} \\
 &\leq \frac{8}{\sqrt{\lambda}} (1 + C^2) X_T^2 S D \left( \beta_T(\delta/4) \log \frac{\det(V_T)}{\det(\lambda I)} \right)^{1/2} \sqrt{T}. \quad ((4), \text{L. 12})
 \end{aligned}$$

■

Now we are ready to prove Theorem 2.

#### 4.6. Putting Everything Together

**Proof** [Proof of Theorem 2] By (13) and Lemmas 7, 9, 13 we have that with probability at least  $1 - \delta/2$ ,

$$\begin{aligned}
 \mathbb{I}_{\{E \cap F\}} (R_1 - R_2 - R_3) &\leq 2DX_T^2(n+d) \log_2 \left( 1 + TX_T^2(1+C^2)/\lambda \right) \\
 &\quad + 2DW^2 \sqrt{2T \log \frac{8}{\delta}} + n\sqrt{B'_\delta} \\
 &\quad + \frac{8}{\sqrt{\lambda}} (1 + C^2) X_T^2 S D \left( \beta_T(\delta/4) \log \frac{\det(V_T)}{\det(\lambda I)} \right)^{1/2} \sqrt{T}.
 \end{aligned}$$

Thus, on  $E \cap F$ , with probability at least  $1 - \delta/2$ ,

$$\begin{aligned}
 R(T) &\leq 2DX_T^2(n+d) \log_2 \left( 1 + TX_T^2(1+C^2)/\lambda \right) \\
 &\quad + 2DW^2 \sqrt{2T \log \frac{8}{\delta}} + n\sqrt{B'_\delta} \\
 &\quad + \frac{8}{\sqrt{\lambda}} (1 + C^2) X_T^2 S D \left( \beta_T(\delta/4) \log \frac{\det(V_T)}{\det(\lambda I)} \right)^{1/2} \sqrt{T}.
 \end{aligned}$$

Further, on  $E \cap F$ , by Lemmas 5 and 10,

$$\log \det V_T \leq (n+d) \log \left( \frac{\lambda(n+d) + T(1+C^2)X_T^2}{\lambda(n+d)} \right) + \log \det \lambda I.$$

Plugging in this gives the final bound, which, by Lemma 4, holds with probability  $1 - \delta$ . ■

## References

- Y. Abbasi-Yadkori, D. Pal, and Cs. Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. Arxiv preprint <http://arxiv.org/abs/1102.2670>, 2011.
- B. D. O. Anderson and J. B. Moore. *Linear Optimal Control*. Prentice-Hall, 1971.
- P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2003. ISSN 1533-7928.
- P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.
- P. Auer, R. Ortner, and Cs. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT-07)*, pages 454–468, 2007.
- P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563—1600, 2010.
- P. L. Bartlett and A. Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence*, 2009.
- A. Becker, P. R. Kumar, and C. Z. Wei. Adaptive control with the stochastic approximation algorithm: Geometry and convergence. *IEEE Trans. on Automatic Control*, 30(4):330–338, 1985.
- D. Bertsekas. *Dynamic Programming*. Prentice-Hall, 1987.
- D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2nd edition, 2001.
- S. Bittanti and M. C. Campi. Adaptive control of linear time invariant systems: the “bet on the best” principle. *Communications in Information and Systems*, 6(4):299–320, 2006.
- R. I. Brafman and M. Tennenholtz. R-MAX - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.
- M. C. Campi and P. R. Kumar. Adaptive linear quadratic Gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6):1890–1907, 1998.
- H. Chen and L. Guo. Optimal adaptive control and consistent parameter estimates for armax model with quadratic cost. *SIAM Journal on Control and Optimization*, 25(4):845–867, 1987.

- H. Chen and J. Zhang. Identification and adaptive control for systems with unknown orders, delay, and coefficients. *Automatic Control, IEEE Transactions on*, 35(8):866–877, August 1990.
- V. Dani and T. P. Hayes. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 937–943, 2006.
- V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. *COLT-2008*, pages 355–366, 2008.
- C. Fiechter. Pac adaptive control of linear systems. In *in Proceedings of the 10th Annual Conference on Computational Learning Theory, ACM*, pages 72–80. Press, 1997.
- S. M. Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
- S. M. Kakade, M. J. Kearns, and J. Langford. Exploration in metric state spaces. In T. Fawcett and N. Mishra, editors, *ICML 2003*, pages 306–312. AAAI Press, 2003.
- M. Kearns and S. P. Singh. Near-optimal performance for reinforcement learning in polynomial time. In J. W. Shavlik, editor, *ICML 1998*, pages 260–268. Morgan Kauffmann, 1998.
- R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704, 2005.
- R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 681–690, 2008.
- T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.
- T. L. Lai and C. Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):pp. 154–166, 1982a.
- T. L. Lai and C. Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982b.
- T. L. Lai and C. Z. Wei. Asymptotically efficient self-tuning regulators. *SIAM Journal on Control and Optimization*, 25:466–481, March 1987.
- T. L. Lai and Z. Ying. Efficient recursive estimation and adaptive control in stochastic regression and armax models. *Statistica Sinica*, 16:741–772, 2006.
- P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

- H. A. Simon. dynamic programming under uncertainty with a quadratic criterion function. *Econometrica*, 24(1):74–81, 1956.
- A. L. Strehl and M. L. Littman. Online linear regression and its application to model-based reinforcement learning. In *NIPS*, pages 1417–1424, 2008.
- A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman. PAC model-free reinforcement learning. In *ICML*, pages 881–888, 2006.
- I. Szita and Cs. Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML 2010*, pages 1031–1038, 2010.

## Appendix A. Tools from Probability Theorem

**Lemma 14** Let  $X_1, \dots, X_t$  be random variables. Let  $a \in \mathbb{R}$ . Let  $S_t = \sum_{s=1}^t X_s$  and  $\tilde{S}_t = \sum_{s=1}^t X_s \mathbb{I}_{\{X_s \leq a\}}$ . Then it holds that

$$\mathbb{P}(S_t > x) \leq \mathbb{P}\left(\max_{1 \leq s \leq t} X_s \geq a\right) + \mathbb{P}\left(\tilde{S}_t > x\right).$$

### Proof

$$\begin{aligned} \mathbb{P}(S_t \geq x) &\leq \mathbb{P}\left(\max_{1 \leq s \leq t} X_s \geq a\right) + \mathbb{P}\left(S_t \geq x, \max_{1 \leq s \leq t} X_s \leq a\right) \\ &\leq \mathbb{P}\left(\max_{1 \leq s \leq t} X_s \geq a\right) + \mathbb{P}\left(\tilde{S}_t \geq x\right). \end{aligned}$$

■

**Theorem 15 (Azuma's inequality)** Assume that  $(X_s; s \geq 0)$  is a supermartingale and  $|X_s - X_{s-1}| \leq c_s$  almost surely. Then for all  $t > 0$  and all  $\epsilon > 0$ ,

$$P(|X_t - X_0| \geq \epsilon) \leq 2 \exp\left(\frac{-\epsilon^2}{2 \sum_{s=1}^t c_s^2}\right).$$

**Theorem 16 (Self-normalized bound for vector-valued martingales)** Let  $(\mathcal{F}_k; k \geq 0)$  be a filtration,  $(m_k; k \geq 0)$  be an  $\mathbb{R}^d$ -valued stochastic process adapted to  $(\mathcal{F}_k)$ ,  $(\eta_k; k \geq 1)$  be a real-valued martingale difference process adapted to  $(\mathcal{F}_k)$ . Assume that  $\eta_k$  is conditionally sub-Gaussian with constant  $R$ . Consider the martingale

$$S_t = \sum_{k=1}^t \eta_k m_{k-1}$$

and the matrix-valued processes

$$V_t = \sum_{k=1}^t m_{k-1} m_{k-1}^\top, \quad \bar{V}_t = V + V_t, \quad t \geq 0,$$

Then for any  $0 < \delta < 1$ , with probability  $1 - \delta$ ,

$$\forall t \geq 0, \quad \|S_t\|_{\bar{V}_t^{-1}}^2 \leq 2R^2 \log \left( \frac{\det(\bar{V}_t)^{1/2} \det(V)^{-1/2}}{\delta} \right).$$

## Appendix B. Controllability and Observability

**Definition 1 (Bertsekas (2001))** A pair  $(A, B)$ , where  $A$  is an  $n \times n$  matrix and  $B$  is an  $n \times d$  matrix, is said to be controllable if the  $n \times nd$  matrix

$$[B \ AB \ \dots \ A^{n-1}B]$$

has full rank. A pair  $(A, C)$ , where  $A$  is an  $n \times n$  matrix and  $C$  is an  $d \times n$  matrix, is said to be observable if the pair  $(A^\top, C^\top)$  is controllable.

## Appendix C. Proof of Lemma 11

**Proof** [Proof of Lemma 11]

We consider first a simple case. Let  $A = B + mm^\top$ ,  $B$  positive definite. Let  $X \neq 0$  be an arbitrary matrix. Using the Cauchy-Schwartz inequality and the fact that for any matrix  $M$ ,  $\|M^\top M\| = \|M\|^2$ , we get

$$\|X^\top mm^\top X\| = \|m^\top X\|^2 = \|m^\top B^{-1/2} B^{1/2} X\|^2 \leq \|m^\top B^{-1/2}\|^2 \|B^{1/2} X\|^2.$$

Thus,

$$\begin{aligned} \|X^\top (B + mm^\top) X\| &\leq \|X^\top BX\| + \|m^\top B^{-1/2}\|^2 \|B^{1/2} X\|^2 \\ &= \left(1 + \|m^\top B^{-1/2}\|^2\right) \|B^{1/2} X\|^2 \end{aligned}$$

and so

$$\frac{\|X^\top AX\|}{\|X^\top BX\|} \leq 1 + \|m^\top B^{-1/2}\|^2.$$

We also have that

$$\det(A) = \det(B + mm^\top) = \det(B) \det(I + B^{-1/2} m (B^{-1/2} m)^\top) = \det(B)(1 + \|m\|_{B^{-1}}^2),$$

thus finishing the proof of this case.

More generally, if  $A = B + m_1 m_1^\top + \cdots + m_{t-1} m_{t-1}^\top$  then define  $V_s = B + m_1 m_1^\top + \cdots + m_{s-1} m_{s-1}^\top$  and use

$$\frac{\|X^\top AX\|}{\|X^\top BX\|} = \frac{\|X^\top V_t X\|}{\|X^\top V_{t-1} X\|} \frac{\|X^\top V_{t-1} X\|}{\|X^\top V_{t-2} X\|} \cdots \frac{\|X^\top V_2 X\|}{\|X^\top BX\|}.$$

By the above argument, since all the terms are positive, we get

$$\frac{\|X^\top AX\|}{\|X^\top BX\|} \leq \frac{\det(V_t)}{\det(V_{t-1})} \frac{\det(V_{t-1})}{\det(V_{t-2})} \cdots \frac{\det(V_2)}{\det(B)} = \frac{\det(V_t)}{\det(B)} = \frac{\det(A)}{\det(B)},$$

the desired inequality.

Finally, by SVD, if  $C \succ 0$ ,  $C$  can be written as the sum of at most  $m$  rank-one matrices, finishing the proof for the general case.  $\blacksquare$

## Appendix D. Bounding $\|x_t\|$

We show that  $E \cap F$  holds with high probability.

**Proof** [Proof of Lemma 4] Let  $M_t = \Theta_* - \tilde{\Theta}_t$ . On event  $E$ , for any  $t \leq T$  we have that

$$\text{trace} \left( M_t^\top \left( \sum_{s=0}^{t-1} \lambda I + z_s z_s^\top \right) M_t \right) \leq \beta_t(\delta/4).$$

Since  $\lambda > 0$  we get that,

$$\text{trace} \left( \sum_{s=0}^{t-1} M_t^\top z_s z_s^\top M_t \right) \leq \beta_t(\delta/4).$$

Thus,

$$\sum_{s=0}^{t-1} \text{trace}(M_t^\top z_s z_s^\top M_t) \leq \beta_t(\delta/4).$$

Since  $\lambda_{\max}(M) \leq \text{trace}(M)$  for  $M \succeq 0$ , we get that

$$\sum_{s=0}^{t-1} \|M_t^\top z_s\|^2 \leq \beta_t(\delta/4).$$

Thus, for all  $t \geq 1$ ,

$$\max_{0 \leq s \leq t-1} \|M_t^\top z_s\| \leq \beta_t(\delta/4)^{1/2} \leq \beta_T(\delta/4)^{1/2}. \quad (15)$$

Choose

$$H > \left( 16 \vee \frac{4S^2 M^2}{(n+d)U_0} \right),$$

where

$$U_0 = \frac{1}{16^{n+d-2}(1 \vee S^{2(n+d-2)})},$$

and

$$M = \sup_{Y \geq 0} \frac{\left( nL \sqrt{(n+d) \log \left( \frac{1+TY/\lambda}{\delta} \right)} + \lambda^{1/2} S \right)}{Y}.$$

Fix a real number  $0 \leq \epsilon \leq 1$ , and consider the time horizon  $T$ . Let  $\pi(v, \mathcal{B})$  and  $\pi(M, \mathcal{B})$  be the projections of vector  $v$  and matrix  $M$  onto subspace  $\mathcal{B} \subset \mathbb{R}^{(n+d)}$ , where the projection of matrix  $M$  is done column-wise. Let  $\mathcal{B} \oplus \{v\}$  be the span of  $\mathcal{B}$  and  $v$ . Let  $\mathcal{B}^\perp$  be the subspace orthogonal to  $\mathcal{B}$  such that  $\mathbb{R}^{(n+d)} = \mathcal{B} \oplus \mathcal{B}^\perp$ .

Define a sequence of subspaces  $\mathcal{B}_t$  as follows: Set  $\mathcal{B}_{T+1} = \emptyset$ . For  $t = T, \dots, 1$ , initialize  $\mathcal{B}_t = \mathcal{B}_{t+1}$ . Then while  $\|\pi(M_t, \mathcal{B}_t^\perp)\|_F > (n+d)\epsilon$ , choose a column of  $M_t$ ,  $v$ , such that  $\|\pi(v, \mathcal{B}_t^\perp)\|_F > \epsilon$  and update  $\mathcal{B}_t = \mathcal{B}_t \oplus \{v\}$ . After finishing with timestep  $t$ , we will have

$$\|\pi(M_t, \mathcal{B}_t^\perp)\| \leq \|\pi(M_t, \mathcal{B}_t^\perp)\|_F \leq (n+d)\epsilon. \quad (16)$$

Let  $\mathcal{T}_T$  be the set of timesteps at which subspace  $\mathcal{B}_t$  expands. The cardinality of this set,  $m$ , is at most  $n+d$ . Denote these timesteps by  $t_1 > t_2 > \dots > t_m$ . Let  $i(t) = \max\{1 \leq i \leq m : t_i \geq t\}$ .

**Lemma 17** *For any vector  $x \in \mathbb{R}^{n+d}$*

$$U\epsilon^{2(n+d)} \|\pi(x, \mathcal{B}_t)\|^2 \leq \sum_{i=1}^{i(t)} \|M_{t_i}^\top x\|^2,$$

where  $U = U_0/H$ .

**Proof** Let  $N = \{v_1, \dots, v_m\}$  be the set of vectors that are added to  $\mathcal{B}_t$  during the expansion timesteps. By construction,  $N$  is a subset of the set of all columns of  $M_{t_1}, M_{t_2}, \dots, M_{t_{i(t)}}$ . Thus, we have that

$$\sum_{i=1}^{i(t)} \|M_{t_i}^\top x\|^2 \geq x^\top (v_1 v_1^\top + \dots + v_m v_m^\top) x.$$

Thus, in order to prove the statement of the lemma, it is enough to show that

$$\forall x, \forall j \in \{1, \dots, m\}, \sum_{k=1}^j \langle v_k, x \rangle^2 \geq \frac{\epsilon^4}{H} \frac{\epsilon^{2(j-2)}}{16^{j-2}(1 \vee S^{2(j-2)})} \|\pi(x, B_j)\|^2, \quad (17)$$

where  $B_j = \text{span}(v_1, \dots, v_j)$  for any  $1 \leq j \leq m$ . We have  $B_m = \mathcal{B}_t$ . We can write  $v_k = w_k + u_k$ , where  $w_k \in B_{k-1}$ ,  $u_k \perp B_{k-1}$ ,  $\|u_k\| \geq \epsilon$ , and  $\|v_k\| \leq 2S$ .

The inequality (17) is proven by induction. First, we prove the induction base for  $j = 1$ . Without loss of generality, assume that  $x = Cv_1$ . From condition  $H > 16$ , we get that

$$16^{-1}H(1 \vee S^{-1}) \geq 1.$$

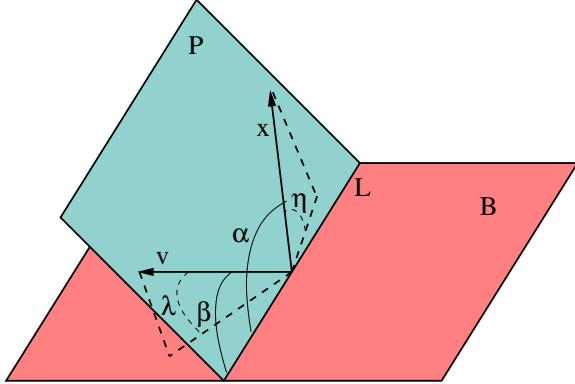


Figure 1: The geometry used in the inductive step.  $v = v_{l+1}$  and  $B = B_l$ .

Thus,

$$\epsilon^2 \geq \frac{\epsilon^2}{16^{-1}H(1 \vee S^{-1})}.$$

Thus,

$$C^2 \|v_1\|^4 \geq \frac{\epsilon^2 C^2 \|v_1\|^2}{16^{-1}H(1 \vee S^{-1})},$$

where we have used the fact that  $\|v_1\| \geq \epsilon$ . Thus,

$$\langle v_1, x \rangle^2 \geq \frac{\epsilon^4}{H} \frac{\epsilon^{-2}}{16^{-1}(1 \vee S^{-2})} \|\pi(x, B_1)\|^2,$$

which establishes the base of induction.

Next, we prove that if the inequality (17) holds for  $j = l$ , then it also holds for  $j = l + 1$ . Figure 1 contains all relevant quantities that are used in the following argument.

Assume that the inequality (17) holds for  $j = l$ . Without loss of generality, assume that  $x$  is in  $B_{l+1}$ , and thus  $\|x\| = \|\pi(x, B_{l+1})\|$ . Let  $P \subset B_{l+1}$  be the 2-dimensional subspace that passes through  $x$  and  $v_{l+1}$ . The 2-dimensional subspace  $P$  and the  $l$ -dimensional subspace  $B_l$  can, respectively, be identified by  $l - 1$  and one equations in  $B_{l+1}$ . Because  $P$  is not a subset of  $B_l$ , the intersection of  $P$  and  $B_l$  is a line in  $B_{l+1}$ . Let's call this line  $L$ . The line  $L$  creates two half-planes on  $P$ . Without loss of generality, assume that  $x$  and  $v_{l+1}$  are on the same half-plane (notice that we can always replace  $x$  by  $-x$  in (17)).

Let  $0 \leq \beta \leq \pi/2$  be the angle between  $v_{l+1}$  and  $L$ . Let  $0 < \lambda < \pi/2$  be the orthogonal angle between  $v_{l+1}$  and  $B_l$ . We know that  $\beta > \lambda$ ,  $v_{l+1}$  is orthogonal to  $B_l$ , and  $\|v_{l+1}\| \geq \epsilon$ . Thus,  $\beta \geq \arcsin(\epsilon / \|v_{l+1}\|)$ . Let  $0 \leq \alpha \leq \pi$  be the angle between  $x$  and  $L$  ( $\alpha < \pi$ , because  $x$  and  $v_{l+1}$  are on the same half-plane). The direction of  $\alpha$  is chosen so that it is consistent with the direction of  $\beta$ . Finally, let  $0 \leq \eta \leq \pi/2$  be the orthogonal angle between  $x$  and  $B_l$ .

By the induction assumption

$$\begin{aligned} \sum_{k=1}^{l+1} \langle v_k, x \rangle^2 &= \langle v_{l+1}, x \rangle^2 + \sum_{k=1}^l \langle v_k, x \rangle^2 \\ &\geq \langle v_{l+1}, x \rangle^2 + \frac{\epsilon^4}{H} \frac{\epsilon^{2(l-2)}}{16^{l-2}(1 \vee S^{2(l-2)})} \|\pi(x, B_l)\|^2. \end{aligned}$$

If  $\alpha < \pi/2 + \beta/2$  or  $\alpha > \pi/2 + 3\beta/2$ , then

$$|\langle v_{l+1}, x \rangle| = \|\|v_{l+1}\| \|x\| \cos \angle(v_{l+1}, x)\| \geq \left| \|\|v_{l+1}\| \|x\| \sin \left( \frac{\beta}{2} \right) \right| \geq \frac{\epsilon \|x\|}{4}.$$

Thus,

$$\langle v_{l+1}, x \rangle^2 \geq \frac{\epsilon^2 \|x\|^2}{16} \geq \frac{\epsilon^4}{H} \frac{\epsilon^{2(l-1)}}{16^{l-1}(1 \vee S^{2(l-1)})} \|\pi(x, B_{l+1})\|^2,$$

where we use  $0 \leq \epsilon \leq 1$  and  $x \in B_{l+1}$  in the last inequality.

If  $\pi/2 + \beta/2 < \alpha < \pi/2 + 3\beta/2$ , then  $\eta < \pi/2 - \beta/2$ . Thus,

$$\|\pi(x, B_l)\| = \|x\| |\cos(\eta)| \geq \|x\| \left| \sin \left( \frac{\beta}{2} \right) \right| \geq \frac{\epsilon \|x\|}{4S}.$$

Thus,

$$\|\pi(x, B_l)\|^2 \geq \frac{\epsilon^2 \|x\|^2}{16S^2},$$

and

$$\frac{\epsilon^4}{H} \frac{\epsilon^{2(l-2)}}{16^{l-2}(1 \vee S^{2(l-2)})} \|\pi(x, B_l)\|^2 \geq \frac{\epsilon^4}{H} \frac{\epsilon^{2(l-1)}}{16^{l-1}(1 \vee S^{2(l-1)})} \|x\|^2,$$

which finishes the proof. ■

Next we show that  $\|M_t^\top z_t\|$  is well controlled except when  $t \in \mathcal{T}_T$ .

**Lemma 18** *We have that for any  $0 \leq t \leq T$ ,*

$$\max_{s \leq t, s \notin \mathcal{T}_t} \|M_s^\top z_s\| \leq G Z_t^{\frac{n+d}{n+d+1}} \beta_t (\delta/4)^{\frac{1}{2(n+d+1)}},$$

where

$$G = 2 \left( \frac{2S(n+d)^{n+d+1/2}}{U^{1/2}} \right)^{1/(n+d+1)},$$

and

$$Z_t = \max_{s \leq t} \|z_s\|.$$

**Proof** From Lemma 17 we get that

$$\sqrt{U}\epsilon^{n+d} \|\pi(z_s, \mathcal{B}_s)\| \leq \sqrt{i(s)} \max_{1 \leq i \leq i(s)} \|M_{t_i}^\top z_s\|,$$

which implies that

$$\|\pi(z_s, \mathcal{B}_s)\| \leq \sqrt{\frac{n+d}{U}} \frac{1}{\epsilon^{n+d}} \max_{1 \leq i \leq i(s)} \|M_{t_i}^\top z_s\|. \quad (18)$$

Now we can write

$$\begin{aligned} \|M_s^\top z_s\| &= \|(\pi(M_s, \mathcal{B}_s^\perp) + \pi(M_s, \mathcal{B}_s))^\top (\pi(z_s, \mathcal{B}_s^\perp) + \pi(z_s, \mathcal{B}_s))\| \\ &= \|\pi(M_s, \mathcal{B}_s^\perp)^\top \pi(z_s, \mathcal{B}_s^\perp) + \pi(M_s, \mathcal{B}_s)^\top \pi(z_s, \mathcal{B}_s)\| \\ &\leq \|\pi(M_s, \mathcal{B}_s^\perp)^\top \pi(z_s, \mathcal{B}_s^\perp)\| + \|\pi(M_s, \mathcal{B}_s)^\top \pi(z_s, \mathcal{B}_s)\| \\ &\leq (n+d)\epsilon \|z_s\| + 2S\sqrt{\frac{n+d}{U}} \frac{1}{\epsilon^{n+d}} \max_{1 \leq i \leq i(s)} \|M_{t_i}^\top z_s\|. \quad \text{by ((18) and (16))} \end{aligned} \quad (19)$$

Thus,

$$\max_{s \leq t, s \notin \mathcal{T}_t} \|M_s^\top z_s\| \leq (n+d)\epsilon Z_t + 2S\sqrt{\frac{n+d}{U}} \frac{1}{\epsilon^{n+d}} \max_{s \notin \mathcal{T}_t, s \leq t} \max_{1 \leq i \leq i(s)} \|M_{t_i}^\top z_s\|.$$

From  $1 \leq i \leq i(s), s \notin \mathcal{T}_t$ , we conclude that  $s < t_i$ . Thus,

$$\max_{s \leq t, s \notin \mathcal{T}_t} \|M_s^\top z_s\| \leq (n+d)\epsilon Z_t + 2S\sqrt{\frac{n+d}{U}} \frac{1}{\epsilon^{n+d}} \max_{0 \leq s < t} \|M_t^\top z_s\|.$$

By (15) we get that

$$\max_{s \leq t, s \notin \mathcal{T}_t} \|M_s^\top z_s\| \leq (n+d)\epsilon Z_t + 2S\sqrt{\frac{n+d}{U}} \frac{1}{\epsilon^{n+d}} \beta_t (\delta/4)^{1/2}.$$

Now if we choose

$$\epsilon = \left( \frac{2S\beta_t (\delta/4)^{1/2}}{Z_t(n+d)^{1/2} U^{1/2} H} \right)^{1/(n+d+1)}$$

we get that

$$\begin{aligned} \max_{s \leq t, s \notin \mathcal{T}_t} \|M_s^\top z_s\| &\leq 2 \left( \frac{2S\beta_t (\delta/4)^{1/2} Z_t^{n+d} (n+d)^{n+d+1/2}}{U^{1/2}} \right)^{1/(n+d+1)} \\ &= G Z_t^{\frac{n+d}{n+d+1}} \beta_t (\delta/4)^{\frac{1}{2(n+d+1)}}. \end{aligned}$$

Finally, we show that this choice of  $\epsilon$  satisfies  $\epsilon < 1$ . From the chose of  $H$ , we have that

$$H > \frac{4S^2 M^2}{(n+d)U_0}.$$

Thus,

$$\left( \frac{4S^2M^2}{(n+d)U_0H} \right)^{\frac{1}{2(n+d+1)}} < 1.$$

Thus,

$$\epsilon = \left( \frac{2S\beta_t(\delta/4)}{Z_t(n+d)^{1/2}U_0^{1/2}H^{1/2}} \right)^{\frac{1}{n+d+1}} < 1.$$

■

We can write the state update as

$$x_{t+1} = \Gamma_t x_t + r_{t+1},$$

where

$$\Gamma_{t+1} = \begin{cases} \tilde{A}_t + \tilde{B}_t K(\tilde{\Theta}_t) & t \notin \mathcal{T}_T \\ A_* + B_* K(\tilde{\Theta}_t) & t \in \mathcal{T}_T \end{cases}$$

and

$$r_{t+1} = \begin{cases} M_t^\top z_t + w_{t+1} & t \notin \mathcal{T}_T \\ w_{t+1} & t \in \mathcal{T}_T \end{cases}$$

Hence we can write

$$\begin{aligned} x_t &= \Gamma_{t-1} x_{t-1} + r_t = \Gamma_{t-1} (\Gamma_{t-2} x_{t-2} + r_{t-1}) + r_t = \Gamma_{t-1} \Gamma_{t-2} x_{t-2} + r_t + \Gamma_{t-1} r_{t-1} \\ &= \Gamma_{t-1} \Gamma_{t-2} \Gamma_{t-3} x_{t-3} + r_t + \Gamma_{t-1} r_{t-1} + \Gamma_{t-1} \Gamma_{t-2} r_{t-2} = \dots = \Gamma_{t-1} \dots \Gamma_{t-t} x_{t-t} \\ &\quad + r_t + \Gamma_{t-1} r_{t-1} + \Gamma_{t-1} \Gamma_{t-2} r_{t-2} + \dots + \Gamma_{t-1} \Gamma_{t-2} \dots \Gamma_{t-(t-1)} r_{t-(t-1)} \\ &= \sum_{k=1}^t \left( \prod_{s=k}^{t-1} \Gamma_s \right) r_k. \end{aligned}$$

From Section 4, we have that

$$\eta \geq \max_{t \leq T} \|A_* + B_* K(\tilde{\Theta}_t)\|, \rho \geq \max_{t \leq T} \|\tilde{A}_t + \tilde{B}_t K(\tilde{\Theta}_t)\|.$$

So we have that

$$\prod_{s=k}^{t-1} \|\Gamma_s\| \leq \eta^{n+d} \rho^{t-k-(n+d)+1}.$$

Hence, we have that

$$\begin{aligned} \|x_t\| &\leq \left( \frac{\eta}{\rho} \right)^{n+d} \sum_{k=1}^t \rho^{t-k+1} \|r_{k+1}\| \\ &\leq \frac{1}{1-\rho} \left( \frac{\eta}{\rho} \right)^{n+d} \max_{0 \leq k \leq t-1} \|r_{k+1}\|. \end{aligned}$$

Now,  $\|r_{k+1}\| \leq \|M_k^\top z_k\| + \|w_{k+1}\|$  when  $k \notin \mathcal{T}_T$ , and  $\|r_{k+1}\| = \|w_{k+1}\|$ , otherwise. Hence,

$$\max_{k < t} \|r_{k+1}\| \leq \max_{k < t, k \notin \mathcal{T}_T} \|M_k^\top z_k\| + \max_{k < t} \|w_{k+1}\|.$$

The first term can be bounded by Lemma 18. The second term can be bounded as follows: notice that from the sub-Gaussianity Assumption A1, we have that for any index  $1 \leq i \leq n$  and any time  $k \leq t$ , with probability  $1 - \delta/(t(t+1))$

$$|w_{k,i}| \leq L \sqrt{2 \log \frac{t(t+1)}{\delta}}.$$

As a result, with a union bound argument, on some event  $H$  with  $\mathbb{P}(H) \geq 1 - \delta/4$ ,  $\|w_t\| \leq 2L\sqrt{n \log \frac{4nt(t+1)}{\delta}}$ . Thus, on  $H \cap E$ ,

$$\|x_t\| \leq \frac{1}{1-\rho} \left(\frac{\eta}{\rho}\right)^{n+d} \left[ GZ_T^{\frac{n+d}{n+d+1}} \beta_t(\delta/4)^{\frac{1}{2(n+d+1)}} + 2L\sqrt{n \log \frac{4nt(t+1)}{\delta}} \right] = \alpha_t.$$

By the definition of  $F$ ,  $H \cap E \subset F \cap E$ . Since, by the union bound,  $\mathbb{P}(H \cap E) \geq 1 - \delta/2$ ,  $\mathbb{P}(E \cap F) \geq 1 - \delta/2$  also holds, finishing the proof.  $\blacksquare$

# Blackwell Approachability and No-Regret Learning are Equivalent

**Jacob Abernethy**

*University of California, Berkeley  
Division of Computer Science*

JAKE@CS.BERKELEY.EDU

**Peter L. Bartlett**

*Univ. of California, Berkeley  
Division of Computer Science  
Department of Statistics*

BARTLETT@CS.BERKELEY.EDU

**Elad Hazan**

*Technion - Israel Institute of Technology  
Industrial Engineering and Management*

EHAZAN@IE.TECHNION.AC.IL

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We consider the celebrated Blackwell Approachability Theorem for two-player games with vector payoffs. Blackwell himself previously showed that the theorem implies the existence of a “no-regret” algorithm for a simple online learning problem. We show that this relationship is in fact much stronger, that Blackwell’s result is equivalent to, in a very strong sense, the problem of regret minimization for Online Linear Optimization. We show that any algorithm for one such problem can be efficiently converted into an algorithm for the other. We provide one novel application of this reduction: the first *efficient* algorithm for calibrated forecasting.

## 1. Introduction

Von Neumann’s minimax theorem (1928) establishes a central result in the theory of two-player zero-sum games, essentially by providing a prescription to both players. This prescription is in the form of a pair of optimal strategies, either of which attains the optimal worst-case value of the game even without knowledge of the opponent’s strategy. However, the theorem fundamentally requires that both players have utility that can be expressed as a *scalar*. In 1956, in response to von Neumann’s result, David Blackwell posed an intriguing question: what guarantee can we hope to achieve when playing a two-player game with a *vector-valued payoff*?

When our payoffs are non-scalar quantities, it does not make sense to ask “can we earn at least  $x$ ?”. A sensible generalization is, “can we guarantee that our vector payoff lies in some convex set  $S$ ?”. In this case the story is more difficult, and Blackwell observed that an oblivious strategy does not suffice—in short, we do not achieve “minimax duality” for vector-payoff games as we can when the payoff is a scalar. Blackwell was able to prove that this negative result applies only for *one-shot games*. In his celebrated Approachability Theorem (Blackwell, 1956), one can achieve a duality statement *in the limit* when the game

is played repeatedly, and the player may learn from his opponent’s prior actions. Blackwell constructed an algorithm (that is, an adaptive strategy) that guarantees the average payoff vector “approaches”  $S$ .

Blackwell’s Approachability Theorem has the flavor of learning in repeated games, a topic which has received much interest. In particular, there are a wealth of recent results on so-called *no-regret learning algorithms* for making repeated decisions given an arbitrary (and potentially adversarial) sequence of cost functions. The first no-regret algorithm for a “discrete action” setting was given in a seminal paper by James Hannan in 1956 (Hannan, 1957). That same year, David Blackwell pointed out (Blackwell, 1954) that his Approachability result leads, as a special case, to an algorithm with essentially the same low-regret guarantee proven by Hannan.

Over the years several other problems have been reduced to Blackwell approachability, including asymptotic calibration (Foster and Vohra, 1998), online learning with global cost functions (Even-Dar et al., 2009) and more (Mannor and Shimkin, 2008). Indeed, it has been presumed that approachability, while establishing the existence of a no-regret algorithm, is strictly more powerful than regret-minimization; hence its utility in such a wide range of problems. In the present paper we prove, to the contrary, that Blackwell’s Approachability Theorem is equivalent, in a very strong sense, to no-regret learning for the setting of *Online Linear Optimization*. This shows that the connection discovered by Blackwell, between regret and approachability, is much stronger than originally supposed.

More specifically, we show how any no-regret algorithm can be converted into an algorithm for Approachability and vice versa. This algorithmic equivalence is achieved via the use of *conic duality*: an approachability problem over a convex cone  $K$  can be reduced to an online linear optimization instance where we must “learn” within the *polar cone*  $K^0$ . The reverse direction is similar. This equivalence provides a range of benefits and one such is “asymptotic calibrated forecasting”. The calibration problem was reduced to Blackwell’s Approachability Theorem by Foster (1999), and a handful of other calibration techniques have been proposed, yet none have provided any efficiency guarantees on the strategy. Using a similar reduction from calibration to approachability, and by carefully constructing the reduction from approachability to online linear optimization, we achieve the first efficient calibration algorithm.

**Related work** There is by now vast literature on all three main topics of this paper: approachability, online learning and calibration, see (Cesa-Bianchi and Lugosi, 2006) for an excellent exposition.

Calibration is a fundamental notion in prediction theory and has found numerous applications in economics and learning. Dawid (1982) was the first to define calibration, with numerous algorithms later given by Foster and Vohra (1998), Fudenberg and Levine (1999), Hart and Mas-Colell (2000) and more (see e.g. (Sandroni et al., 2003; Perchet, 2009)). Foster has given a calibration algorithm based on approachability (Foster, 1999). There are numerous definitions (mostly asymptotic) of calibration in the literature. In this paper we give precise finite-time rates of calibration. Furthermore, we give the first *efficient* algorithm for calibration: attaining  $\varepsilon$ -calibration (formally defined later) required a running time of  $\text{poly}(\frac{1}{\varepsilon})$  for all previous algorithms, whereas our algorithm runs in time proportional to  $\log \frac{1}{\varepsilon}$ .

## 2. Game Theory Preliminaries

### 2.1. Two-Player Games

Formally, a two-player normal-form game is defined by a pair of action sets  $[n]$  and  $[m]$ , for natural numbers  $n, m$ , and a pair of utility functions  $u_1, u_2 : [n] \times [m] \rightarrow \mathbb{R}$ . When player 1 chooses action  $i$  and player 2 chooses action  $j$ , player 1 and player 2 receive utilities  $u_1(i, j)$  and  $u_2(i, j)$  respectively. An important class of two-player games are known as *zero-sum*, in that  $u_1 \equiv -u_2$ . For zero-sum games we drop the subscripts on  $u_1, u_2$  and simply write  $u(i, j)$  for player 1's utility, and  $-u(i, j)$  for player 2's utility. For the remainder of this section, we shall be concerned entirely with zero-sum games, hence we will refer to player 1 as the Player and player 2 as the Adversary.

It is natural to assume that the players in a game may include randomness in their choice of action; simple games such as Rock-Paper-Scissors require randomness to achieve optimality. When the players choose their actions randomly according to the distributions  $p \in \Delta_n$  and  $q \in \Delta_m$ , respectively, the *expected utility* for the Player is  $\sum_{i,j} p(i)q(j)u(i, j)$ . Von Neumann's minimax theorem, widely considered the first key result in game theory, tells us that both the Player and the Adversary have an “optimal” randomized strategy that can be played without knowledge of the strategy of their respective opponent.

**Theorem 1 (Von Neumann's Minimax Theorem (Neumann et al., 1947))** *For any integers  $n, m > 0$  and any utility function  $u : [n] \times [m] \rightarrow \mathbb{R}$ ,*

$$\max_{p \in \Delta_n} \min_{q \in \Delta_m} \sum_{i,j} p(i)q(j)u(i, j) = \min_{q \in \Delta_m} \max_{p \in \Delta_n} \sum_{i,j} p(i)q(j)u(i, j)$$

The statement of the minimax theorem is often referred to as *duality* as it swaps the min and max. This result can be used to establish strong duality for linear programming. It was proven by Maurice Sion in the 1950's that von Neumann's notion of duality can be extended further, for a much larger class of input spaces and a more general class of functions.

**Theorem 2 (Sion (1958)<sup>1</sup>)** *Given convex compact sets  $\mathcal{X} \subset \mathbb{R}^n, \mathcal{Y} \subset \mathbb{R}^m$ , and a function  $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  convex and concave in its first and second arguments respectively, we have*

$$\inf_{\mathbf{x} \in \mathcal{X}} \sup_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) = \sup_{\mathbf{y} \in \mathcal{Y}} \inf_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}).$$

In the present work we shall not need anything quite so general, although we use this theorem to generalize slightly the class of two-player zero-sum games. Rather than define the actions of our players as being drawn randomly from discrete sets  $[n]$  and  $[m]$ , let the players' decision space be characterized by given compact convex sets  $\mathcal{X} \subset \mathbb{R}^n$  and  $\mathcal{Y} \subset \mathbb{R}^m$  respectively. In addition, we shall assume that the utility is characterized by a *biaffine* function  $u : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ ; that is,  $u(\alpha\mathbf{x} + (1 - \alpha)\mathbf{x}', \mathbf{y}) = \alpha u(\mathbf{x}, \mathbf{y}) + (1 - \alpha)u(\mathbf{x}', \mathbf{y})$  and  $u(\mathbf{x}, \alpha\mathbf{y} + (1 - \alpha)\mathbf{y}') = \alpha u(\mathbf{x}, \mathbf{y}) + (1 - \alpha)u(\mathbf{x}, \mathbf{y}')$  for every  $0 \leq \alpha \leq 1$ ,  $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$  and  $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$ . Following Sion's theorem, we arrive at the following.

**Corollary 3** *For compact convex sets  $\mathcal{X} \subset \mathbb{R}^n$  and  $\mathcal{Y} \subset \mathbb{R}^m$  and any biaffine function  $u : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ , we have*

$$\max_{\mathbf{x} \in \mathcal{X}} \min_{\mathbf{y} \in \mathcal{Y}} u(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{y} \in \mathcal{Y}} \max_{\mathbf{x} \in \mathcal{X}} u(\mathbf{x}, \mathbf{y})$$

This alternative description of a zero-sum game has two advantages. First, we now assume that both players are deterministic. That is, we have converted the notion of a randomized strategy on a discrete action space to a deterministic strategy  $\mathbf{x}$  inside of a convex set  $\mathcal{X}$ . Rather than evaluate the expected utility of a randomized action, this expectation is now incorporated via the linearity of  $u(\cdot, \cdot)$ . Note, crucially, that the assumptions that  $u$  is biaffine and  $\mathcal{X}$  and  $\mathcal{Y}$  are convex imply that neither player gains from randomness, as  $\mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y}} u(\mathbf{x}, \mathbf{y}) = u(\mathbb{E}_{\mathbf{x}} \mathbf{x}, \mathbb{E}_{\mathbf{y}} \mathbf{y})$ .

A second advantage of this framework is that it allows us to work with action spaces that might seem prohibitively large. For example, we can imagine a game in which each player must select a route in a graph  $G$  between two endpoints, and the utility is the amount of overlap of their paths. The set of paths in a graph is exponential, and even counting the number of such paths is  $\#P$ -hard. However, we may instead set  $\mathcal{X}$  and  $\mathcal{Y}$  to be the *flow polytope* of  $G$ . The flow polytope can be described by a polynomially-sized number of constraints, and hence is much easier to work with.

## 2.2. Vector-Valued Games

Let us now turn our attention to Blackwell’s question: what can be guaranteed when the utility function of the zero-sum game is *vector-valued*? Following the definition in the previous section, we can define a vector-valued game in terms of some biaffine utility function  $\mathbf{u} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$  from a product of two convex compact decision spaces  $\mathcal{X} \subset \mathbb{R}^n$  and  $\mathcal{Y} \subset \mathbb{R}^m$  to  $d$ -dimensional space. The biaffine property is defined in the natural way.

Note that we may not apply our usual notions of utility maximization when dealing with vector-valued games—what does it mean to “maximize” a vector? Furthermore, the concept of “zero-sum” is not immediately clear. Blackwell proposed the following framework: suppose that the Player, who selects  $\mathbf{x} \in \mathcal{X}$ , would like his vector payoff  $\mathbf{u}(\mathbf{x}, \mathbf{y})$  to land inside of a particular closed convex set  $S \subset \mathbb{R}^d$ , where  $S$  is fixed and known to both players. We shall say that the Player wants to *satisfy*  $S$ . The Adversary, who selects  $\mathbf{y} \in \mathcal{Y}$ , would like to prevent the Player from satisfying  $S$ .

Let us return our attention to the simple case of scalar-valued games discussed in Section 2.1. The duality statement achieved in the Minimax Theorem, typically stated in terms of swapping the order of min and max, can instead be formulated in terms of swapping quantifiers  $\forall$  and  $\exists$ .

**Proposition 1** *For any convex compact sets  $\mathcal{X} \subset \mathbb{R}^n$  and  $\mathcal{Y} \subset \mathbb{R}^m$ , and any biaffine utility function  $u : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ , we have the following implication for any  $c \in \mathbb{R}$ :*

$$\forall \mathbf{y} \in \mathcal{Y} \exists \mathbf{x} \in \mathcal{X} : u(\mathbf{x}, \mathbf{y}) \in [c, \infty) \implies \exists \mathbf{x} \in \mathcal{X} \forall \mathbf{y} \in \mathcal{Y} : u(\mathbf{x}, \mathbf{y}) \in [c, \infty).$$

This proposition is simply another way to state duality, in the following form:

$$\min_{\mathbf{y} \in \mathcal{Y}} \max_{\mathbf{x} \in \mathcal{X}} u(\mathbf{x}, \mathbf{y}) \geq c \implies \max_{\mathbf{x} \in \mathcal{X}} \min_{\mathbf{y} \in \mathcal{Y}} u(\mathbf{x}, \mathbf{y}) \geq c.$$

Put another way, if the Player can earn  $c$  by choosing his strategy *with knowledge of* the Adversary’s strategy, then he can earn  $c$  obviously as well.

Here we have simply taken the Minimax Theorem and stated it in terms of satisfying a set, namely the set  $S = [c, \infty)$  for some value  $c$ . This interpretation begs the question: can

we achieve a similar “duality” statement for vector-valued games? In other words, given a biaffine utility function  $\mathbf{u} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$  and any convex set  $S \subset \mathbb{R}^d$ , does the statement

$$\forall \mathbf{y} \in \mathcal{Y} \exists \mathbf{x} \in \mathcal{X} : \mathbf{u}(\mathbf{x}, \mathbf{y}) \in S \implies \exists \mathbf{x} \in \mathcal{X} \forall \mathbf{y} \in \mathcal{Y} : \mathbf{u}(\mathbf{x}, \mathbf{y}) \in S$$

hold in general? The answer, unfortunately, is *no!* Consider the following easy example:  $\mathcal{X} = \mathcal{Y} := [0, 1]$ , the payoff is simply  $\mathbf{u}(x, y) := (x, y)$  for  $x, y \in [0, 1]$ , and the set in question is  $S := \{(z, z) \mid z \in [0, 1]\}$ . Certainly the premise is true, since for every  $y$  there exists an  $x$ , namely  $x = y$ , such that  $\mathbf{u}(x, y) \in S$ . On the other hand, there is no such single  $x$  for which  $\mathbf{u}(x, y) \in S$  for any  $y$ .

### 2.3. Blackwell Approachability

While we might hope that minimax duality, framed in terms of set satisfiability, would extend from scalar-valued games to vector-valued games, the previous example appears to be a nail in the coffin. But in fact the story is not quite so bad: the proposed example is difficult because it is a *one-shot* game. What Blackwell observed, and led to the Approachability Theorem, is that if the game is played *repeatedly* then one can achieve duality “in the limit.” To make this precise we introduce some definitions.

**Definition 4** A Blackwell instance is a tuple  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$ , with  $\mathcal{X} \subset \mathbb{R}^n$  and  $\mathcal{Y} \subset \mathbb{R}^m$  compact and convex,  $\mathbf{u} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$  biaffine, and  $S \subset \mathbb{R}^d$  convex and closed. For any instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$ , we say that

- $S$  is satisfiable if  $\exists \mathbf{x} \in \mathcal{X} \forall \mathbf{y} \in \mathcal{Y} : \mathbf{u}(\mathbf{x}, \mathbf{y}) \in S$ .
- $S$  is response-satisfiable if  $\forall \mathbf{y} \in \mathcal{Y} \exists \mathbf{x} \in \mathcal{X} : \mathbf{u}(\mathbf{x}, \mathbf{y}) \in S$ .
- $S$  is halfspace-satisfiable if, for any halfspace  $H \supseteq S$ ,  $H$  is satisfiable.

To recap, when our utility function  $\mathbf{u}$  is scalar-valued, i.e. for zero-sum games where  $d = 1$ , then minimax duality holds and, according to Proposition 1, this be rephrased as “If  $S := [c, \infty)$  is response-satisfiable then  $S$  is satisfiable.” On the other hand, for vector-valued games it is not the case in general that “ $S$  is response-satisfiable  $\implies S$  is satisfiable” for arbitrary sets  $S$ . What Blackwell showed is that response-satisfiability does lead to a weaker condition, termed *approachability*. Before we define this precisely, let us use the notation  $\text{dist}(\mathbf{z}, U)$  to denote the distance between a point  $\mathbf{z}$  and some convex set  $U$ , that is  $\inf_{\mathbf{x} \in U} \|\mathbf{z} - \mathbf{x}\|$ .

**Definition 5** Given a Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$ , we say that  $S$  is approachable if there exists some algorithm  $\mathcal{A}$  which selects points in  $\mathcal{X}$  such that, for any sequence  $\mathbf{y}_1, \mathbf{y}_2, \dots \in \mathcal{Y}$ , we have

$$\text{dist}\left(\frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{x}_t, \mathbf{y}_t), S\right) \rightarrow 0 \quad \text{as } T \rightarrow \infty,$$

where  $\mathbf{x}_t \leftarrow \mathcal{A}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{t-1})$ .

Under this new notion, we now allow the Player to implement an *adaptive strategy* for a repeated version of the game, and we require that the average utility vector becomes arbitrarily close to  $S$ . Intuitively, we may think of approachability as “satisfiability in the limit”.

**Theorem 6 (Blackwell's Approachability Theorem (Blackwell, 1956))** *For any Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$ ,  $S$  is approachable if and only if it is response-satisfiable.*

The beauty of this theorem is that, while we may not be able to satisfy  $S$  in a one-shot version of the game, we can satisfy the set “on average” if we may play the game indefinitely.

This version of the theorem, which appears in Even-Dar et al. (2009), is not the one usually attributed to Blackwell. The original theorem uses the concept of halfspace satisfiability. It is not difficult to establish the equivalence of the two statements via the following lemma, whose proof uses a nice application of minimax duality.

**Lemma 7** *For any Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$ ,  $S$  is response-satisfiable if and only if it is halfspace-satisfiable.*

**Proof** ( $\implies$ ) Assume that  $S$  is response-satisfiable. Hence, for any  $\mathbf{y}$  there is an  $\mathbf{x}_y$  such that  $\mathbf{u}(\mathbf{x}_y, \mathbf{y}) \in S$ . Now take any halfspace  $H \supset S$  parameterized by  $\boldsymbol{\theta}, c$ , that is  $H = \{\mathbf{z} : \langle \boldsymbol{\theta}, \mathbf{z} \rangle \leq c\}$ . Then let us define a scalar-valued game with utility  $u(\mathbf{x}, \mathbf{y}) = \langle \boldsymbol{\theta}, \mathbf{u}(\mathbf{x}, \mathbf{y}) \rangle$ . Notice that  $H \supset S$  implies that  $\langle \boldsymbol{\theta}, \mathbf{z} \rangle \leq c$  for all  $\mathbf{z} \in S$ . Since  $S$  is response-satisfiable, for every  $\mathbf{y}$  there is an  $\mathbf{x}_y$  such that  $\mathbf{u}(\mathbf{x}_y, \mathbf{y}) \in S \implies u(\mathbf{x}_y, \mathbf{y}) \leq c$ . We then immediately see that

$$\max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{x} \in \mathcal{X}} u(\mathbf{x}, \mathbf{y}) \leq \max_{\mathbf{y} \in \mathcal{Y}} u(\mathbf{x}_y, \mathbf{y}) \leq c.$$

It follows from Corollary 3 that  $\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} u(\mathbf{x}, \mathbf{y}) \leq c$ . Let  $\mathbf{x}^* \in \mathcal{X}$  be any minimizer of the latter expression and notice that, for any  $\mathbf{y} \in \mathcal{Y}$ , we have that  $u(\mathbf{x}^*, \mathbf{y}) \leq c$ . It follows immediately that  $H$  is satisfiable.

( $\impliedby$ ) Assume that  $S$  is not response-satisfiable. Hence, there must exist some  $\mathbf{y}_0 \in \mathcal{Y}$  such that  $\mathbf{u}(\mathbf{x}, \mathbf{y}_0) \notin S$  for every  $\mathbf{x} \in \mathcal{X}$ . Consider the set  $U := \{\mathbf{u}(\mathbf{x}, \mathbf{y}_0) \text{ for all } \mathbf{x} \in \mathcal{X}\}$  and notice that  $U$  is convex since  $\mathcal{X}$  is convex and  $\mathbf{u}(\cdot, \mathbf{y}_0)$  is affine. Furthermore, because  $S$  is convex and  $S \cap U = \emptyset$  by assumption, there must exist some halfspace  $H$  separating the two sets, that is  $S \subseteq H$  and  $H \cap U = \emptyset$ . By construction, we see that for any  $\mathbf{x}$ ,  $\mathbf{u}(\mathbf{x}, \mathbf{y}_0) \notin H$  and hence  $H$  is not satisfiable. It follows immediately that  $S$  is not halfspace-satisfiable. ■

Although it is not posed in this language, Blackwell's original theorem uses the concept of a *halfspace oracle*. Given a Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$ , define a halfspace oracle to be a function  $\mathcal{O}$  that takes as input any halfspace  $H \supset S$  and returns a point  $\mathcal{O}(H) = \mathbf{x}_H \in \mathcal{X}$ , and we shall refer to a halfspace oracle as *valid* if it satisfies that for each halfspace  $H \supset S$ ,  $\mathbf{u}(\mathbf{x}_H, \mathbf{y}) \in H$  for any  $\mathbf{y} \in \mathcal{Y}$ .

**Theorem 8** *For any Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$ , the set  $S$  is approachable if and only if there exists a valid halfspace oracle.*

Notice that the existence of a valid halfspace oracle is equivalent to the halfspace-satisfiability condition. Hence, via Lemma 7, this theorem is equivalent to Theorem 6.

To achieve approachability, following Definition 5 one must construct an algorithm  $\mathcal{A}$  that maps the observed subsequence  $\mathbf{y}_1, \dots, \mathbf{y}_{t-1} \in \mathcal{Y}$  to a point  $\mathbf{x}_t \in \mathcal{X}$ . By the previous theorem, in order for the set  $S$  to be approachable, there must be a valid halfspace oracle  $\mathcal{O}$ , and hence  $\mathcal{A}$  may make calls to  $\mathcal{O}$ . Blackwell actually provides such an algorithm, quite

elegant for its simplicity, which can be found in his original work (Blackwell, 1956) as well as in the book of Cesa-Bianchi and Lugosi (2006).

We note that, when an approachability algorithm  $\mathcal{A}$  is adapted to a Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$ , and makes calls to a halfspace oracle  $\mathcal{O}$ , we may write  $\mathcal{A}_{\mathcal{X}, \mathcal{Y}, \mathbf{u}, S}^{\mathcal{O}}$  to make the dependence clear.

### 3. Online Linear Optimization

Online Convex Optimization (OCO) has become a popular topic within Machine Learning since it was introduced by Zinkevich (2003), and there has been much followup work (Shalev-Shwartz and Singer, 2007; Rakhlin et al., 2010; Hazan, 2010; Abernethy et al., 2009). It provides a generic problem template and was shown to generalize several existing problems in the realm of online learning and repeated decision making. Among these are online pattern classification, the “experts” or “hedge” setting, and sequential portfolio optimization (Freund and Schapire, 1995; Hazan et al., 2007).

In the OCO setting, we imagine an online game between Player and Nature. Assume the Player is given a convex decision set  $\mathcal{K} \subset \mathbb{R}^d$  and must make a sequence of decisions  $\mathbf{x}_1, \mathbf{x}_2, \dots \in \mathcal{K}$ . After committing to  $\mathbf{x}_t$ , Nature reveals a convex loss function  $\ell_t$ , and Player pays  $\ell_t(\mathbf{x}_t)$ . The performance of the Player is typically measured by *regret* which we shall define below. In the present work we shall be concerned with the more specific problem of Online Linear Optimization (OLO) where the loss functions are assumed to be linear,  $\ell_t(\mathbf{x}) = \langle \mathbf{f}_t, \mathbf{x} \rangle$  for some  $\mathbf{f}_t \in \mathbb{R}^d$ .

We define the Player’s adaptive strategy  $\mathcal{L}$ , which we refer to as an *OLO algorithm*, as a function which takes as input a subsequence of loss vectors  $\mathbf{f}_1, \dots, \mathbf{f}_{t-1}$  and returns a point  $\mathbf{x}_t \leftarrow \mathcal{L}(\mathbf{f}_1, \dots, \mathbf{f}_{t-1})$ , where  $\mathbf{x}_t \in \mathcal{K}$ .

**Definition 9** Given an OLO algorithm  $\mathcal{L}$  and a sequence of loss vectors  $\mathbf{f}_1, \mathbf{f}_2, \dots \in \mathbb{R}^d$ , let  $\text{Regret}(\mathcal{L}; \mathbf{f}_{1:T}) := \sum_{t=1}^T \langle \mathbf{f}_t, \mathbf{x}_t \rangle - \min_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^T \langle \mathbf{f}_t, \mathbf{x} \rangle$ . When the sequence of loss vectors is clear, we may simply write  $\text{Regret}_T(\mathcal{L})$ .

An important question is whether an OLO algorithm has a regret rate which scales *sublinearly* in  $T$ . A sublinear regret is key, for then our average performance, in the long run, is essentially no worse than the best in hindsight. We use the term *no-regret* algorithm when it possesses this property.

**Theorem 10** For any bounded decision set  $\mathcal{K} \subset \mathbb{R}^d$  there exists an algorithm  $\mathcal{L}_{\mathcal{K}}$  such that  $\text{Regret}_T(\mathcal{L}_{\mathcal{K}}) = o(T)$  for any sequence of loss vectors  $\{\mathbf{f}_t\}$  with bounded norm.

Later in the paper we provide one such algorithm, known as Online Gradient Descent, proposed by Zinkevich (2003).

Before proceeding, let us demonstrate the value of no-regret algorithms by proving an aforementioned result. We shall sketch a proof of the minimax statement of Corollary 3. Assume we are given convex and compact decision space  $\mathcal{X} \subset \mathbb{R}^n$  and  $\mathcal{Y} \subset \mathbb{R}^m$ , and without loss of generality assume we have a utility function  $u : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  of the form  $u(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top M \mathbf{y}$  for some  $M \in \mathbb{R}^{n \times m}$ . Weak duality, i.e.  $\min_{\mathbf{y} \in \mathcal{Y}} \max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top M \mathbf{y} \geq \max_{\mathbf{x} \in \mathcal{X}} \min_{\mathbf{y} \in \mathcal{Y}} \mathbf{x}^\top M \mathbf{y}$  is trivial, and so we turn our attention to the reverse inequality. We shall imagine our game is played repeatedly, where on round  $t$  the first player chooses

$\mathbf{x}_t$  and the second chooses  $\mathbf{y}_t$ , but where both players select their strategies according to a no-regret algorithm. For every  $t$  we shall set  $\mathbf{x}_t \leftarrow \mathcal{L}_{\mathcal{X}}(\mathbf{f}_1, \dots, \mathbf{f}_{t-1})$  and  $\mathbf{y}_t \leftarrow \mathcal{L}_{\mathcal{Y}}(\mathbf{g}_1, \dots, \mathbf{g}_{t-1})$ , where we define the vectors  $\mathbf{f}_t := -M\mathbf{y}_t$  and  $\mathbf{g}_t^\top := \mathbf{x}_t^\top M$ . By applying the definition of regret twice, we have

$$\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t^\top M \mathbf{y}_t = \min_{\mathbf{y} \in \mathcal{Y}} \left( \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \right)^\top M \mathbf{y} + \frac{\text{Regret}_T(\mathcal{L}_{\mathcal{Y}})}{T} \leq \max_{\mathbf{x} \in \mathcal{X}} \min_{\mathbf{y} \in \mathcal{Y}} \mathbf{x}^\top M \mathbf{y} + \frac{o(T)}{T}, \quad (1)$$

$$\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t^\top M \mathbf{y}_t = \max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top M \left( \frac{1}{T} \sum_{t=1}^T \mathbf{y}_t \right) - \frac{\text{Regret}(\mathcal{L}_{\mathcal{X}})}{T} \geq \min_{\mathbf{y} \in \mathcal{Y}} \max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top M \mathbf{y} - \frac{o(T)}{T}. \quad (2)$$

Combining these two statements gives  $\min_{\mathbf{y} \in \mathcal{Y}} \max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top M \mathbf{y} \leq \max_{\mathbf{x} \in \mathcal{X}} \min_{\mathbf{y} \in \mathcal{Y}} \mathbf{x}^\top M \mathbf{y} + \frac{o(T)}{T}$ . Of course, we can let  $T \rightarrow \infty$  which immediately gives the desired inequality.

The previous example foreshadows a key result of this paper, which is that any no-regret learning algorithm can be converted into an approachability strategy. If we interpret Blackwell Approachability as a generalized form of Minimax Duality for vector-valued games then it may come as no surprise that regret-minimizing algorithms would provide a tool in establishing both game-theoretic results. However, in a certain sense regret-minimization is too heavy a hammer for proving Minimax Duality. For one, the above proof requires that we imagine a repeated version of the game, whereas scalar-valued game duality holds even for one-shot. Indeed, more standard proofs of von Neumann's result do not rely on repeated play. Blackwell Approachability, on the other hand, fundamentally involves repeated play, and in fact we shall show that regret-minimization is the perfectly-sized hammer, as it is *algorithmically equivalent* to approachability.

## 4. Equivalence of Approachability and Regret Minimization

### 4.1. Convex Cones and Conic Duality

We shall define some basic notions and then state some simple lemmas. Henceforth we use the notation  $B_2(r)$  to refer to the  $\ell_2$ -norm ball of radius  $r$ . The notation  $\mathbf{x}' \oplus \mathbf{x}$  is the vector concatenation of  $\mathbf{x}$  and  $\mathbf{x}'$ .

**Definition 11** A set  $X \subset \mathbb{R}^d$  is a cone if it is closed under multiplication by nonnegative scalars, and  $X$  is a convex cone if it is also closed under element addition. Given any set  $K \subset \mathbb{R}^d$ , define the conic hull  $\mathbf{cone}(K) := \{\alpha \mathbf{x} : \alpha \in \mathbb{R}_+, \mathbf{x} \in K\}$  which is also a cone in  $\mathbb{R}^d$ . Also, given any convex cone  $C \subset \mathbb{R}^d$ , we can define the polar cone of  $C$  as

$$C^0 := \{\boldsymbol{\theta} \in \mathbb{R}^d : \langle \boldsymbol{\theta}, \mathbf{x} \rangle \leq 0 \text{ for all } \mathbf{x} \in C\}.$$

It is easily checked that if  $K$  is convex then  $\mathbf{cone}(K)$  is also convex. The following Lemma is folklore.

**Lemma 12** If  $C$  is a convex cone then (1)  $(C^0)^0 = C$  and (2) supporting hyperplanes in  $C^0$  correspond to points  $\mathbf{x} \in C$ , and vice versa. That is, given any supporting hyperplane  $H$  of  $C^0$ ,  $H$  can be written exactly as  $\{\boldsymbol{\theta} \in \mathbb{R}^d : \langle \boldsymbol{\theta}, \mathbf{x} \rangle = 0\}$  for some vector  $\mathbf{x} \in C$  that is unique up to scaling.

The distance to a cone can conveniently be measured via a “dual formulation,” as we now show.

**Lemma 13** *For every convex cone  $C$  in  $\mathbb{R}^d$*

$$\text{dist}(\mathbf{x}, C) = \max_{\boldsymbol{\theta} \in C^0 \cap B_2(1)} \langle \boldsymbol{\theta}, \mathbf{x} \rangle \quad (3)$$

**Proof** We need two simple observations. Define  $\pi_C(\mathbf{x})$  as the projection of  $\mathbf{x}$  onto  $C$ . Then clearly, for any  $\mathbf{x}$ ,

$$\text{dist}(\mathbf{x}, C) = \|\mathbf{x} - \pi_C(\mathbf{x})\| \quad (4)$$

$$\langle \mathbf{x} - \pi_C(\mathbf{x}), \mathbf{y} \rangle \leq 0 \quad \forall \mathbf{y} \in C \text{ and hence } \mathbf{x} - \pi_C(\mathbf{x}) \in C^0 \quad (5)$$

$$\langle \mathbf{x} - \pi_C(\mathbf{x}), \pi_C(\mathbf{x}) \rangle = 0 \quad (6)$$

Given any  $\boldsymbol{\theta} \in C^0$  with  $\|\boldsymbol{\theta}\| \leq 1$ , since  $\pi_C(\mathbf{x}) \in C$  we have that

$$\langle \boldsymbol{\theta}, \mathbf{x} \rangle \leq \langle \boldsymbol{\theta}, \mathbf{x} - \pi_C(\mathbf{x}) \rangle \leq \|\boldsymbol{\theta}\| \|\mathbf{x} - \pi_C(\mathbf{x})\| \leq \|\mathbf{x} - \pi_C(\mathbf{x})\|,$$

which immediately implies that  $\max_{\boldsymbol{\theta} \in C^0, \|\boldsymbol{\theta}\| \leq 1} \langle \boldsymbol{\theta}, \mathbf{x} \rangle \leq \text{dist}(\mathbf{x}, C)$ . Furthermore, by selecting  $\boldsymbol{\theta} = \frac{\mathbf{x} - \pi_C(\mathbf{x})}{\|\mathbf{x} - \pi_C(\mathbf{x})\|}$  which has norm one and, by (4), is in  $C^0$ , we see that

$$\max_{\boldsymbol{\theta} \in C^0, \|\boldsymbol{\theta}\| \leq 1} \langle \boldsymbol{\theta}, \mathbf{x} \rangle \geq \left\langle \frac{\mathbf{x} - \pi_C(\mathbf{x})}{\|\mathbf{x} - \pi_C(\mathbf{x})\|}, \mathbf{x} \right\rangle = \left\langle \frac{\mathbf{x} - \pi_C(\mathbf{x})}{\|\mathbf{x} - \pi_C(\mathbf{x})\|}, \mathbf{x} - \pi_C(\mathbf{x}) \right\rangle = \|\mathbf{x} - \pi_C(\mathbf{x})\|,$$

which implies that  $\max_{\boldsymbol{\theta} \in C^0, \|\boldsymbol{\theta}\| \leq 1} \langle \boldsymbol{\theta}, \mathbf{x} \rangle \geq \text{dist}(\mathbf{x}, C)$  and hence we are done.  $\blacksquare$

Our results require looking at convex cones rather than convex sets, hence we must consider the process of converting a set into a cone. In order to not lose information about the underlying set  $\mathcal{K} \subset \mathbb{R}^d$ , we shall embed the set into a higher dimension, and instead look at  $\text{cone}(\{\kappa\} \times \mathcal{K}) \subset \mathbb{R}^{d+1}$ , where  $\kappa := \max_{\mathbf{x} \in \mathcal{K}} \|\mathbf{x}\|$  is the diameter of  $\mathcal{K}$ . We prove that this process of “lifting” and conifying does not perturb distances by more than a constant.

**Lemma 14** *Consider a compact convex set  $\mathcal{K} \subseteq \mathcal{H}$  in  $\mathbb{R}^d$  and  $\mathbf{x} \notin \mathcal{K}$ . Let  $\tilde{\mathbf{x}} := \kappa \oplus \mathbf{x}$  and  $\tilde{\mathcal{K}} := \{\kappa\} \times \mathcal{K}$ . Then we have*

$$\text{dist}(\tilde{\mathbf{x}}, \text{cone}(\tilde{\mathcal{K}})) \leq \text{dist}(\mathbf{x}, \mathcal{K}) \leq 2\text{dist}(\tilde{\mathbf{x}}, \text{cone}(\tilde{\mathcal{K}})) \quad (7)$$

**Proof** Since  $\text{dist}(\tilde{\mathbf{x}}, \tilde{\mathcal{K}}) = \text{dist}(\mathbf{x}, \mathcal{K})$  and  $\tilde{\mathcal{K}} \subset \text{cone}(\tilde{\mathcal{K}})$ , the first inequality follows immediately.

For notational convenience let  $\mathbf{w} = \pi_{\text{cone}(\tilde{\mathcal{K}})}(\mathbf{y})$  be the projection of  $\mathbf{y}$  onto  $\text{cone}(\tilde{\mathcal{K}})$  and  $\mathbf{v} = \pi_{\tilde{\mathcal{K}}}(\mathbf{y})$  be the projection onto  $\tilde{\mathcal{K}}$ . Consider the plane determined by the three points  $\tilde{\mathbf{x}}, \mathbf{w}, \mathbf{v}$ . Notice that the triangle  $\Delta(\tilde{\mathbf{x}}, \mathbf{w}, \mathbf{v})$  is similar to the triangle  $\Delta(\mathbf{0}, \kappa \oplus \mathbf{0}, \mathbf{v})$ , and hence by triangle similarity

$$\frac{\|\mathbf{v}\|}{\|\kappa \oplus \mathbf{0}\|} = \frac{\|\tilde{\mathbf{x}} - \mathbf{v}\|}{\|\tilde{\mathbf{x}} - \mathbf{w}\|} = \frac{\text{dist}(\tilde{\mathbf{x}}, \tilde{\mathcal{K}})}{\text{dist}(\tilde{\mathbf{x}}, \text{cone}(\tilde{\mathcal{K}}))}$$

For a visual aid, we provide a picture of this triangle similarity in Figure 1. Since  $\mathbf{v} \in \tilde{\mathcal{K}}$  we have  $\|\mathbf{v}\| \leq \|\tilde{\mathcal{K}}\| \leq 2\kappa$ . In addition  $\|\kappa \oplus \mathbf{0}\| = \kappa$  and the result follows.  $\blacksquare$

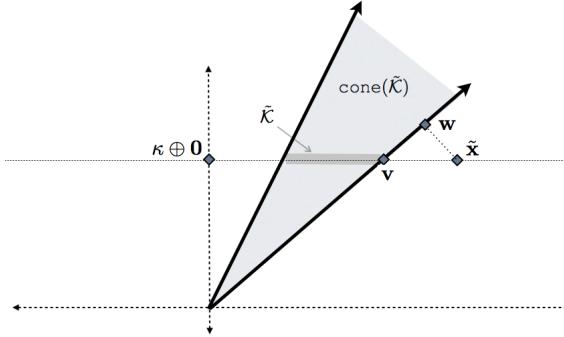


Figure 1: A geometric interpretation of the proof of Lemma 14.

#### 4.2. Duality Theorems

In the previous sections we have presented two sequential decision problems, summarized in Figure 2. We now show that these two decision problems are *algorithmically equivalent*: any strategy (algorithm) that achieves approachability can be converted into an algorithm that achieves low-regret, and vice versa.

Blackwell Approachability Problem	Online Linear Optimization Problem
<p>Given a Blackwell instance <math>(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)</math> and a valid halfspace oracle <math>\mathcal{O} : H \mapsto \mathbf{x}_H \in \mathcal{X}</math>, construct an algorithm <math>\mathcal{A}</math> so that, for any sequence <math>\mathbf{y}_1, \mathbf{y}_2, \dots \in \mathcal{Y}</math>,</p> $\text{dist}\left(\frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{x}_t, \mathbf{y}_t), S\right) \rightarrow 0$ <p>where <math>\mathbf{x}_t \leftarrow \mathcal{A}(\mathbf{y}_1, \dots, \mathbf{y}_{t-1})</math>.</p>	<p>Given a compact convex set <math>\mathcal{K} \subset \mathbb{R}^d</math>, construct a learning algorithm <math>\mathcal{L}</math> so that, for any sequence of loss vectors <math>\mathbf{f}_1, \mathbf{f}_2, \dots \in \mathbb{R}^d</math> we have vanishing regret, that is</p> $\sum_{t=1}^T \langle \mathbf{f}_t, \mathbf{x}_t \rangle - \min_{\mathbf{x} \in K} \sum_{t=1}^T \langle \mathbf{f}_t, \mathbf{x} \rangle = o(T),$ <p>where <math>\mathbf{x}_t \leftarrow \mathcal{L}(\mathbf{f}_1, \dots, \mathbf{f}_{t-1})</math>.</p>

Figure 2: A summary of Blackwell Approachability and Online Linear Optimization

We present this equivalence as a pair of reductions. In Algorithm 1 we show how a learner, presented with a OLO problem characterized by a decision set  $\mathcal{K}$  and an arriving sequence of loss vectors  $\mathbf{f}_1, \mathbf{f}_2, \dots$ , can minimize regret with only oracle access to some approachability algorithm  $\mathcal{A}$ . In Algorithm 2 we show how a player, presented with a Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$  and a valid halfspace oracle  $\mathcal{O}$ , can achieve approachability when only given oracle access to a no-regret OLO algorithm  $\mathcal{L}$ . For the remainder of the paper, for a given Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$  and approachability algorithm

$\mathcal{A}$ ,  $D(\mathcal{A}; \mathbf{y}_1, \dots, \mathbf{y}_T)$  shall refer to the rate of approachability  $\text{dist}\left(\frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{x}_t, \mathbf{y}_t), S\right)$ . We shall write  $D_T(\mathcal{A})$  when the input sequence is clear. For the convex set  $\mathcal{K}$ , we shall let  $\kappa := \max_{\mathbf{x} \in \mathcal{K}} \|\mathbf{x}\|$ , the “norm” of the set  $\mathcal{K}$ .

---

**Algorithm 1** Conversion of Approachability Alg.  $\mathcal{A}$  to Online Linear Optimization Alg.  $\mathcal{L}$

---

- 1: Input: compact convex decision set  $\mathcal{K} \subset \mathbb{R}^d$
  - 2: Input: sequence of cost functions  $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_T \in B_2(1)$
  - 3: Input: approachability oracle  $\mathcal{A}$
  - 4: Set: Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$ , where  $\mathcal{X} := \mathcal{K}$ ,  $\mathcal{Y} := B_2(1)$ ,  $\mathbf{u}(\mathbf{x}, \mathbf{f}) = \frac{\langle \mathbf{f}, \mathbf{x} \rangle}{\kappa} \oplus -\mathbf{f}$ , and  $S := \text{cone}(\{\kappa\} \times \mathcal{K})^0$
  - 5: Construct: valid halfspace oracle  $\mathcal{O}$  // Existence established in Lemma 15
  - 6: **for**  $t = 1, \dots, T$  **do**
  - 7:   Let:  $\mathcal{L}(\mathbf{f}_1, \dots, \mathbf{f}_{t-1}) := \mathcal{A}_{\mathcal{X}, \mathcal{Y}, \mathbf{u}, S}^{\mathcal{O}}(\mathbf{f}_1, \dots, \mathbf{f}_{t-1})$
  - 8:   Receive: cost function  $\mathbf{f}_t$
  - 9: **end for**
- 

In Algorithm 1 we require the construction of a valid halfspace oracle. In the lemma below we give one such oracle and prove that it is valid, but we note that this construction may not be the most efficient in general; any particular scenario may give rise to a simpler and faster construction.

**Lemma 15** *There exists a valid halfspace oracle for the Blackwell instance in Algorithm 1.*

**Proof** Assume we have some halfspace  $H$  which contains  $S = \text{cone}(\{\kappa\} \times \mathcal{K})^0$ . We can assume without loss of generality that  $H$  is tangent to  $S$  and, since  $S$  is a cone,  $H$  meets the origin; that is,  $H = \{\boldsymbol{\theta} : \langle \boldsymbol{\theta}, \mathbf{z}_H \rangle \leq 0\}$  for some  $\mathbf{z}_H \in \mathbb{R}^d$ . Furthermore,  $H \supset \text{cone}(\{\kappa\} \times \mathcal{K})^0$  implies that  $\mathbf{z}_H \in (\text{cone}(\{\kappa\} \times \mathcal{K})^0)^0 = \text{cone}(\{\kappa\} \times \mathcal{K})$ . Equivalently,  $\mathbf{z}_H = \alpha(\kappa \oplus \mathbf{x}_H)$  for some  $\mathbf{x}_H \in \mathcal{K}$  and some  $\alpha > 0$ . With this in mind, we construct our oracle by setting  $\mathbf{x}_H \leftarrow \mathcal{O}(H)$ .

It remains to prove that this halfspace oracle is valid. We compute  $\langle \mathbf{u}(\mathbf{x}_H, \mathbf{f}), \mathbf{z}_H \rangle$ :

$$\langle \mathbf{u}(\mathbf{x}_H, \mathbf{f}), \mathbf{z}_H \rangle = \langle \kappa^{-1} \langle \mathbf{f}, \mathbf{x}_H \rangle \oplus -\mathbf{f}, \alpha \kappa \oplus \alpha \mathbf{x}_H \rangle = \alpha \langle \mathbf{f}, \mathbf{x}_H \rangle + \langle -\mathbf{f}, \alpha \mathbf{x}_H \rangle = 0.$$

By definition,  $\langle \mathbf{u}(\mathbf{x}_H, \mathbf{f}), \mathbf{z}_H \rangle \leq 0$  implies that  $\mathbf{u}(\mathbf{x}_H, \mathbf{f}) \in H$  for any  $\mathbf{f}$  and we are done. ■

**Theorem 16** *The reduction defined in Algorithm 1, for any input algorithm  $\mathcal{A}$ , produces an OLO algorithm  $\mathcal{L}$  such that  $\frac{\text{Regret}(\mathcal{L})}{T} \leq 2\kappa D_T(\mathcal{A})$ .*

**Proof** Applying Lemmas 13 and 12 to the definition of  $D_T(\mathcal{A})$  gives

$$D_T(\mathcal{A}) \equiv \text{dist}\left(\frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{x}_t, \mathbf{f}_t), S\right) = \max_{\mathbf{w} \in \text{cone}(\kappa \oplus \mathcal{K}) \cap B_2^d(1)} \left\langle \frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{x}_t, \mathbf{f}_t), \mathbf{w} \right\rangle \quad (8)$$

Notice that, in this optimization, we can assume w.l.o.g. that  $\|\mathbf{w}\| = 1$ , or  $\mathbf{w} = \mathbf{0}$ . In the former case we can write  $\mathbf{w} = \frac{\kappa \oplus \mathbf{x}}{\|\kappa \oplus \mathbf{x}\|}$  for some  $\mathbf{x} \in \mathcal{K}$ , and we drop the latter case to obtain the inequality

$$\begin{aligned} D_T(\mathcal{A}) &\geq \max_{\mathbf{x} \in \mathcal{K}} \left\langle \frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{x}_t, \mathbf{f}_t), \frac{\kappa \oplus \mathbf{x}}{\|\kappa \oplus \mathbf{x}\|} \right\rangle \\ &= \frac{1}{T} \max_{\mathbf{x} \in \mathcal{K}} \frac{\left( \sum_{t=1}^T \langle \mathbf{f}_t, \mathbf{x}_t \rangle - \sum_{t=1}^T \langle \mathbf{f}_t, \mathbf{x} \rangle \right)}{\|\kappa \oplus \mathbf{x}\|} \\ &\geq \frac{\frac{1}{T} \left( \sum_{t=1}^T \langle \mathbf{f}_t, \mathbf{x}_t \rangle - \sum_{t=1}^T \langle \mathbf{f}_t, \mathbf{x}^* \rangle \right)}{\|\kappa \oplus \mathbf{x}^*\|} \geq \frac{\frac{1}{T} \text{Regret}_T(\mathcal{A})}{2\kappa}, \end{aligned}$$

where we set  $\mathbf{x}^* := \arg \min_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^T \langle \mathbf{f}_t, \mathbf{x} \rangle$ . ■

We turn our attention to the second reduction.

---

**Algorithm 2** Conversion of Online Linear Optimization Alg.  $\mathcal{L}$  to Approachability Alg.  $\mathcal{A}$

---

- 1: Input: Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$ , with  $S$  a cone; and a valid halfspace oracle  $\mathcal{O}$
- 2: Input: Online Linear Optimization oracle  $\mathcal{L}$
- 3: Set:  $\mathcal{K} = S^0 \cap B_2(1)$
- 4: **for**  $t = 1, \dots, T$  **do**
- 5:   Query  $\mathcal{L}$ :  $\boldsymbol{\theta}_t \leftarrow \mathcal{L}_{\mathcal{K}}(\mathbf{f}_1, \dots, \mathbf{f}_{t-1})$ , where  $\mathbf{f}_s \leftarrow -\mathbf{u}(\mathbf{x}_s, \mathbf{y}_s)$
- 6:   Query  $\mathcal{O}$ :  $\mathbf{x}_t \leftarrow \mathcal{O}(H_{\boldsymbol{\theta}_t})$  where  $H_{\boldsymbol{\theta}_t} := \{\mathbf{z} : \langle \boldsymbol{\theta}_t, \mathbf{z} \rangle \leq 0\}$
- 7:   Let:  $\mathcal{A}(\mathbf{y}_1, \dots, \mathbf{y}_{t-1}) := \mathbf{x}_t$
- 8:   Receive:  $\mathbf{y}_t \in \mathcal{Y}$
- 9: **end for**

---

We now prove a similar rate for reverse direction. Here we assume that  $S$  is a cone, but we relax this restriction next.

**Theorem 17** *The reduction in Algorithm 2, when  $S$  is a cone, leads to a rate of approachability of algorithm  $\mathcal{A}$  of  $D_T(\mathcal{A}; \mathbf{y}_{1:T}) \leq \frac{\text{Regret}(\mathcal{L}_{\mathcal{K}}, \mathbf{f}_{1:T})}{T}$ .*

**Proof** We state precisely the halfspace oracle guarantee from line 6. We know that  $\mathbf{u}(\mathbf{x}_t, \mathbf{y}) \in H_{\boldsymbol{\theta}_t}$  or equivalently  $\langle \boldsymbol{\theta}_t, \mathbf{u}(\mathbf{x}_t, \mathbf{y}) \rangle \leq 0$  for any  $\mathbf{y} \in \mathcal{Y}$ . In particular, since  $\mathbf{u}(\mathbf{x}_t, \mathbf{y}_t) = -\mathbf{f}_t$ , we have  $\langle \boldsymbol{\theta}_t, \mathbf{f}_t \rangle \geq 0$ . We bound  $D_T(\mathcal{A})$  by applying Lemma 13 to obtain:

$$\begin{aligned} D_T(\mathcal{A}) &= \text{dist} \left( \frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{x}_t, \mathbf{y}_t), S \right) = \max_{\boldsymbol{\theta} \in \mathcal{K}} \left\langle \frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{x}_t, \mathbf{y}_t), \boldsymbol{\theta} \right\rangle = \frac{1}{T} \max_{\boldsymbol{\theta} \in \mathcal{K}} \left( - \sum_{t=1}^T \langle \mathbf{f}_t, \boldsymbol{\theta} \rangle \right) \\ &\leq \frac{1}{T} \left( \sum_{t=1}^T \langle \mathbf{f}_t, \boldsymbol{\theta}_t \rangle - \min_{\boldsymbol{\theta} \in \mathcal{K}} \sum_{t=1}^T \langle \mathbf{f}_t, \boldsymbol{\theta} \rangle \right) = \frac{1}{T} \text{Regret}_T(\mathcal{A}) \end{aligned} \tag{9}$$

where the inequality follows by the halfspace oracle guarantee. ■

For a Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$ , even when  $S$  is not a cone we can still use Algorithm 2 by *lifting*  $S$ : apply Algorithm 2 to the instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}'(\cdot, \cdot), S')$ , where  $S' := \text{cone}(\{\kappa\} \times S)$  and  $\mathbf{u}'(\mathbf{x}, \mathbf{y}) := \kappa \oplus \mathbf{u}(\mathbf{x}, \mathbf{y})$ .

**Corollary 18** *Given a Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$  with compact  $S$ , and let its lifted instance be  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}'(\cdot, \cdot), S')$  as described above. Then*

$$\text{dist}\left(\frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{x}_t, \mathbf{y}_t), S\right) \leq 2 \cdot \text{dist}\left(\frac{1}{T} \sum_{t=1}^T \mathbf{u}'(\mathbf{x}_t, \mathbf{y}_t), S'\right) \leq \frac{2}{T} \text{Regret}_T(\mathcal{A})$$

**Proof** Apply Lemma 14 to Theorem 17. ■

We include the compactness assumption only because Lemma 14 requires it yet it is not necessary; the size of  $S$  does not enter into the bound. For any Blackwell instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S)$  with non-compact  $S$ , we may always consider a functionally equivalent instance  $(\mathcal{X}, \mathcal{Y}, \mathbf{u}(\cdot, \cdot), S_0)$ , where  $S_0 \subset S$  is compact. Letting  $U := \{\mathbf{u}(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}\}$ , which is compact, we may simply let  $S_0$  be the convex hull of all projections of points in  $U$  onto  $S$ . Hence  $\text{dist}(\mathbf{z}, S) = \text{dist}(\mathbf{z}, S_0)$  for all  $\mathbf{z} \in U$ .

## 5. Efficient Calibration via Approachability and OLO

Imagine a sequence of binary outcomes, say ‘rain’ or ‘shine’ on a given day, and imagine a forecaster, say the weatherman, that wants to predict the probability of this outcome on each day. A natural question to ask is, on the days when the weatherman actually predicts “30% chance of rain”, does it actually rain (roughly) 30% of the time? This exactly the problem of *calibrated forecasting* which we now discuss.

There have been a range of definitions of calibration given throughout the literature, some equivalent and some not, but from a computational viewpoint there are significant differences. We thus give a clean definition of calibration, first introduced by Foster (1999), which is convenient to asses computationally.

We let  $y_1, y_2, \dots \in \{0, 1\}$  be a sequence of outcomes, and  $p_1, p_2, \dots \in [0, 1]$  a sequence of probability predictions by a forecaster. We define for every  $T$  and every probability interval  $[p - \varepsilon/2, p + \varepsilon/2]$  for  $p \in [0, 1]$  and  $\varepsilon > 0$ , the quantities

$$n_T(p, \varepsilon) := \sum_{t=1}^T \mathbb{I}[p_t \in [p - \varepsilon/2, p + \varepsilon/2]], \quad \rho_T(p, \varepsilon) := \frac{\sum_{t=1}^T y_t \mathbb{I}[p_t \in [p - \varepsilon/2, p + \varepsilon/2]]}{n_T(p, \varepsilon)}.$$

The quantity  $\rho_T(p, \varepsilon)$  should be interpreted as the empirical frequency of  $y_t = 1$ , up to round  $T$ , on only those rounds where the forecaster’s prediction was within  $\varepsilon/2$  of  $p$ . The goal of calibration, of course, is to have this empirical frequency  $\rho_T(p, \varepsilon)$  be close to the estimated frequency  $p$  in the limit. The standard definition of a calibrated forecaster is one that satisfies

$$\text{for all } p \in [0, 1], \varepsilon > 0 : \quad \limsup_{T \rightarrow \infty} |\rho_T(p, \varepsilon) - p| \leq O(\varepsilon) \quad \text{unless} \quad n_T(p, \varepsilon) = o(T). \quad (10)$$

Requiring that  $n_T(p, \varepsilon)$  does not grow too slowly is an important condition, as we can not expect the forecaster to be calibrated in regions on which he predicts only a small number of times. On the other hand, this case-sensitive condition is somewhat awkward, and we instead use the following equivalent notion.

**Definition 19** *Let the  $(\ell_1, \varepsilon)$ -calibration rate for forecaster  $\mathcal{A}$  be*

$$C_T^\varepsilon(\mathcal{A}) = \max \left\{ 0, \sum_{i=0}^{\lfloor \varepsilon^{-1} \rfloor} \frac{n_T(i\varepsilon, \varepsilon)}{T} |i\varepsilon - \rho_T(i\varepsilon, \varepsilon)| - \frac{\varepsilon}{2} \right\}.$$

We say that a forecaster is  $(\ell_1, \varepsilon)$ -calibrated if  $C_T^\varepsilon(\mathcal{A}) = o(1)$ .

The definition of asymptotic calibration considers the “total error” over an  $\varepsilon$ -grid, and it adjusts the normalization for each term to  $\frac{1}{T}$ . The benefit here is that we can ignore intervals in this grid for which  $n_T(p, \varepsilon) = o(T)$ . In addition, we subtract the constant  $\varepsilon/2$  which is an artifact of the discretization by  $\varepsilon$ ; this is the smallest constant which allows for  $\limsup_{T \rightarrow \infty} C_T^\varepsilon(\mathcal{A}) \leq 0$ . A standard reduction in the literature (see e.g. (Cesa-Bianchi and Lugosi, 2006)) shows that a fully-calibrated algorithm (i.e. one satisfying (10)) can be constructed from any  $(\ell_1, \varepsilon)$ -calibrated algorithm. Henceforth we only consider the  $(\ell_1, \varepsilon)$  condition.

As our goal is to minimize the calibration score  $C_T^\varepsilon$ , we can interpret this value instead as a distance to the  $\ell_1$ -norm ball. Define the *calibration vector*  $\mathbf{c}_T \in \mathbb{R}^{\lfloor \varepsilon^{-1} \rfloor}$  at time  $T$  as:  $\mathbf{c}_T(i) = \frac{n_T(i\varepsilon, \varepsilon)}{T} (i\varepsilon - \rho_T(i\varepsilon, \varepsilon))$ .

**Claim 1** Whenever  $\mathbf{c}_T \notin B_1(\varepsilon/2)$ , we have

$$C_T^\varepsilon = \text{dist}_1(\mathbf{c}_T, B_1(\varepsilon/2)).$$

**Proof** Notice that for any  $\mathbf{x}$ :  $\text{dist}_1(\mathbf{x}, B_1(\varepsilon/2)) := \min_{\mathbf{y}: \|\mathbf{y}\|_1 \leq \varepsilon/2} \|\mathbf{x} - \mathbf{y}\|_1 = \max\{0, -\varepsilon/2 + \|\mathbf{x}\|_1\}$ . The second equality follows by noting that an optimally chosen  $\mathbf{y}$  will lie in the same quadrant as  $\mathbf{x}$ . When we set  $\mathbf{x} = \mathbf{c}_T$ , it is clear that  $\|\mathbf{c}_T\|_1 > \varepsilon/2$  given our assumption that  $\mathbf{c}_T \notin B_1(\varepsilon/2)$ .  $\blacksquare$

The utility of this claim shall be to convert the problem of  $(\ell_1, \varepsilon)$ -calibration to a problem of approachability; that is, can we approach the set  $B_1(\varepsilon/2)$  for a particular vector-valued game? In the following section we describe this construction in detail.

### 5.1. Existence of Calibrated Forecaster via Blackwell Approachability

A surprising fact is that it is possible to achieve calibration even when the outcome sequence  $\{y_t\}$  is chosen by an adversary, although this requires a randomized strategy of the forecaster. Algorithms for calibrated forecasting under adversarial conditions have been given in Foster and Vohra (1998), Fudenberg and Levine (1999), and Hart and Mas-Colell (2000).

Interestingly, the calibration problem was reduced to Blackwell’s Approachability Theorem in a short paper by Foster (1999). Foster’s reduction uses Blackwell’s original theorem, proving that a given set is halfspace-satisfiable, in particular by providing a construction for

each such halfspace. Here we provide a reduction to Blackwell Approachability using the response-satisfiability condition – that is by using Theorem 6 – which is both significantly easier and more intuitive than Foster’s construction<sup>2</sup>. We also show, using the reduction to Online Linear Optimization from the previous section, how to achieve the most efficient known algorithm for calibration by taking advantage of the Online Gradient Descent algorithm of Zinkevich (2003), using the results of Section 4.

We now describe the construction that allows us to reduce calibration to approachability. For any  $\varepsilon > 0$  we will show how to construct an  $(\ell_1, \varepsilon)$ -calibrated forecaster. Notice that from here, it is straightforward to produce a well-calibrated forecaster (Foster and Vohra, 1998). For simplicity, assume  $\varepsilon = 1/m$  for some positive integer  $m$ . On each round  $t$ , a forecaster will now randomly predict a probability  $p_t \in \{0/m, 1/m, 2/m, \dots, (m-1)/m, 1\}$ , according to the distribution  $\mathbf{w}_t$ , that is  $\Pr(p_t = i/m) = \mathbf{w}_t(i)$ . We now define a vector-valued game. Let the player choose  $\mathbf{w}_t \in \mathcal{X} := \Delta_{m+1}$ , and the adversary choose  $y_t \in \mathcal{Y} := [0, 1]$ , and the payoff vector will be

$$\mathbf{u}(\mathbf{w}_t, y_t) := \left\langle \mathbf{w}_t(0) \left( y_t - \frac{0}{m} \right), \mathbf{w}_t(1) \left( y_t - \frac{1}{m} \right), \dots, \mathbf{w}_t(m) \left( y_t - 1 \right) \right\rangle \quad (11)$$

**Lemma 20** *Consider the vector-valued game described above and let  $S := B_1(\varepsilon/2)$ . If we have a strategy for choosing  $\mathbf{w}_t$  that guarantees approachability of  $S$ , that is  $\frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{w}_t, y_t) \rightarrow S$ , then a randomized forecaster that selects  $p_t$  according to  $\mathbf{w}_t$  is  $(\ell_1, \varepsilon)$ -calibrated with high probability.*

The proof of this lemma is straightforward, and is similar to the construction in Foster (1999). The key fact is that  $\frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{w}_t, y_t) = \mathbb{E}[\mathbf{c}_T]$ , where the expectation is taken over the algorithms draws of every  $p_t$  according to the distribution  $\mathbf{w}_t$ . Since each  $p_t$  is drawn independently, by standard concentration arguments we can see that if  $\frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{w}_t, y_t)$  is close to the  $\ell_1$  ball of radius  $\varepsilon/2$ , then the  $(\ell_1, \varepsilon)$ -calibration vector is close to the  $\varepsilon/2$  ball with high probability.

We can now apply Theorem 6 to prove the existence of a calibrated forecaster.

**Theorem 21** *For the vector-valued game defined in (11), the set  $B_1(\varepsilon/2)$  is response-satisfiable and, hence, approachable.*

**Proof** To show response-satisfiability, we need only show that, for every strategy  $y \in [0, 1]$  played by the adversary, there is a strategy  $\mathbf{w} \in \Delta_m$  for which  $\mathbf{u}(\mathbf{w}, y) \in S$ . This can be achieved by simply setting  $i$  so as to minimize  $|i\varepsilon - y|$ , which can always be made smaller than  $\varepsilon/2$ . We then choose our distribution  $\mathbf{w} \in \Delta_{m+1}$  to be a point mass on  $i$ , that is we set  $w(i) = 1$  and  $w(j) = 0$  for all  $j \neq i$ . Then  $\mathbf{u}(\mathbf{w}, y)$  is identically 0 everywhere except the  $i$ th coordinate, which has the value  $y - i/m$ . By construction,  $y - i/m \in [-1/m, 1/m]$ , and we are done. ■

---

2. A similar existence proof was discovered concurrently by Mannor and Stoltz (2009)

## 5.2. Efficient Algorithm for Calibration via Online Linear Optimization

We now show how the results in the previous Section lead to the first efficient algorithm for calibrated forecasting. The previous theorem provides a natural existence proof for Calibration, but it does not immediately provide us with a simple and efficient algorithm. We proceed according to the reduction outlined in the previous section to prove:

**Theorem 22** *There exists a  $(\ell_1, \varepsilon)$ -calibration algorithm that runs in time  $O(\log \frac{1}{\varepsilon})$  per iteration and satisfies  $C_T^\varepsilon = O\left(\frac{1}{\sqrt{\varepsilon T}}\right)$*

The reduction developed in Theorem 17 has some flexibility, and we shall modify it for the purposes of this problem. The objects we shall need, as well as the required conditions, are as follows:

1. A convex set  $\mathcal{K}$
2. An efficient algorithm  $\mathcal{A}$  which, for any sequence  $\mathbf{f}_1, \mathbf{f}_2, \dots$ , can select a sequence of points  $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots \in \mathcal{K}$  with the guarantee that  $\sum_{t=1}^T \langle \mathbf{f}_t, \boldsymbol{\theta}_t \rangle - \min_{\boldsymbol{\theta} \in \mathcal{K}} \sum_{t=1}^T \langle \mathbf{f}_t, \boldsymbol{\theta} \rangle = o(T)$ . For the reduction, we shall set  $\mathbf{f}_t \leftarrow -\mathbf{u}(\mathbf{w}_t, y_t)$ .
3. An efficient oracle that can select a particular  $\mathbf{w}_t \in \mathcal{X}$  for each  $\boldsymbol{\theta}_t \in \mathcal{K}$  with the guarantee that

$$\text{dist}\left(\frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{w}_t, y_t), S\right) \leq \frac{1}{T} \left( \sum_{t=1}^T \langle -\mathbf{u}(\mathbf{w}_t, y_t), \boldsymbol{\theta}_t \rangle - \min_{\boldsymbol{\theta} \in \mathcal{K}} \sum_{t=1}^T \langle -\mathbf{u}(\mathbf{w}_t, y_t), \boldsymbol{\theta} \rangle \right) \quad (12)$$

where the function `dist()` can be with respect to any norm.

**The Setup** Let  $\mathcal{K} = B_\infty(1) = \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_\infty \leq 1\}$  be the unit cube. This is an appropriate choice because we can write  $\text{dist}_1(\mathbf{x}, B_1(\varepsilon/2))$  for  $\mathbf{x} \notin B_1(\varepsilon/2)$  as

$$\text{dist}_1(\mathbf{x}, B_1(\varepsilon/2)) := \min_{\mathbf{y}: \|\mathbf{y}\|_1 \leq \varepsilon/2} \|\mathbf{x} - \mathbf{y}\|_1 = -\varepsilon/2 + \|\mathbf{x}\|_1 = -\varepsilon/2 - \min_{\boldsymbol{\theta}: \|\boldsymbol{\theta}\|_\infty \leq 1} \langle -\mathbf{x}, \boldsymbol{\theta} \rangle; \quad (13)$$

The former equality was proved in Claim 1. Furthermore, we shall construct our oracle mapping  $\boldsymbol{\theta} \mapsto \mathbf{w}$  with the following guarantee:  $\langle \mathbf{u}(\mathbf{w}, y), \boldsymbol{\theta} \rangle \leq \varepsilon/2$  for any  $y$ . Using this guarantee, and if we plug in  $\mathbf{x} = \frac{1}{T} \sum_{t=1}^T \mathbf{u}(\mathbf{w}_t, y_t)$  (13), we arrive at:

$$\begin{aligned} \text{dist}_1\left(\frac{\sum_{t=1}^T \mathbf{u}(\mathbf{w}_t, y_t)}{T}, B_1(\varepsilon/2)\right) &= -\varepsilon/2 - \min_{\boldsymbol{\theta}: \|\boldsymbol{\theta}\|_\infty \leq 1} \left\langle \frac{-\sum_{t=1}^T \mathbf{u}(\mathbf{w}_t, y_t)}{T}, \boldsymbol{\theta} \right\rangle \\ &\leq \frac{1}{T} \left( \sum_{t=1}^T \langle -\mathbf{u}(\mathbf{w}_t, y_t), \boldsymbol{\theta}_t \rangle - \min_{\boldsymbol{\theta} \in \mathcal{K}} \sum_{t=1}^T \langle -\mathbf{u}(\mathbf{w}_t, y_t), \boldsymbol{\theta} \rangle \right) \end{aligned}$$

This is precisely the necessary guarantee (12).

---

**Algorithm 3** Efficient Oracle mapping  $\mathcal{O} : \mathbf{w} \mapsto \boldsymbol{\theta}$ 


---

```

Input:  $\boldsymbol{\theta}$  such that  $\|\boldsymbol{\theta}\|_\infty \leq 1$ 
if  $\boldsymbol{\theta}(0) \leq 0$  then
     $\mathbf{w} \leftarrow \delta_0$  // That is, choose  $\mathbf{w}$  to place all weight on the 0th coordinate
else if  $\boldsymbol{\theta}(m) \geq 0$  then
     $\mathbf{w} \leftarrow \delta_m$  // That is, choose  $\mathbf{w}$  to place all weight on the last coordinate
else
    Binary search  $\boldsymbol{\theta}$  to find coordinate  $i$  such that  $\boldsymbol{\theta}(i) > 0$  and  $\boldsymbol{\theta}(i+1) \leq 0$ 
     $\mathbf{w} \leftarrow \frac{\boldsymbol{\theta}(i)^{-1}}{\boldsymbol{\theta}(i)^{-1} - \boldsymbol{\theta}(i+1)^{-1}} \delta_i + \frac{-\boldsymbol{\theta}(i+1)^{-1}}{\boldsymbol{\theta}(i)^{-1} - \boldsymbol{\theta}(i+1)^{-1}} \delta_{i+1}$ 
end if
Return  $\mathbf{w}$ 

```

---

**Constructing the Oracle** We now turn our attention to designing the required oracle in an *efficient* manner. In particular, given any  $\boldsymbol{\theta}$  with  $\|\boldsymbol{\theta}\|_\infty \leq 1$  we must construct  $\mathbf{w} \in \Delta_{m+1}$  so that  $\langle \ell(\mathbf{w}, y), \boldsymbol{\theta} \rangle \leq \varepsilon/2$  for any  $y$ . The details of this oracle are given in Algorithm 3. It is straightforward why, in the final **else** condition, there must be such a pair of coordinates  $i, i+1$  satisfying the condition. We need not be concerned with the case that  $\boldsymbol{\theta}(i+1) = 0$ , where we can simply define  $\frac{0}{\infty} = 0$  and  $\frac{\infty}{\infty} = 1$  leading to  $\mathbf{w} \leftarrow \delta_{i+1}$ . It is also clear that, with the binary search, this algorithm requires at most  $O(\log m) = O(\log 1/\varepsilon)$  computation.

In order to prove that this construction is valid we need to check the condition that, for any  $y \in \{0, 1\}$ ,  $\langle \mathbf{u}(\mathbf{w}, y), \boldsymbol{\theta} \rangle \leq \varepsilon/2$ ; or more precisely,  $\sum_{i=1}^m \boldsymbol{\theta}(i) \mathbf{w}(i) \left( y - \frac{i}{m} \right) \leq \varepsilon/2$ . Recalling that  $m = 1/\varepsilon$ , this is trivially checked for the case when  $\boldsymbol{\theta}(1) \leq 0$  or  $\boldsymbol{\theta}(m) \geq 0$ . Otherwise, we have

$$\begin{aligned} \langle \mathbf{u}(\mathbf{w}, y), \boldsymbol{\theta} \rangle &= \frac{\boldsymbol{\theta}(i) \cdot \boldsymbol{\theta}(i)^{-1}}{\boldsymbol{\theta}(i)^{-1} - \boldsymbol{\theta}(i+1)^{-1}} \left( y - \frac{i}{m} \right) + \frac{\boldsymbol{\theta}(i+1) \cdot (-\boldsymbol{\theta}(i+1)^{-1})}{\boldsymbol{\theta}(i)^{-1} - \boldsymbol{\theta}(i+1)^{-1}} \left( y - \frac{i+1}{m} \right) \\ &= \frac{1}{\boldsymbol{\theta}(i)^{-1} - \boldsymbol{\theta}(i+1)^{-1}} \frac{1}{m} \leq \frac{\max(|\boldsymbol{\theta}(i)|, |\boldsymbol{\theta}(i+1)|)}{2} \varepsilon \leq \frac{\varepsilon}{2} \end{aligned}$$

**The Learning Algorithm** The final piece is to construct an efficient learning algorithm which leads to vanishing regret. That is, we need to construct a sequence of  $\boldsymbol{\theta}_t$ 's in the unit cube (denoted  $B_\infty(1)$ ) so that

$$\sum_{t=1}^T \langle \mathbf{u}_t, \boldsymbol{\theta}_t \rangle - \min_{\boldsymbol{\theta} \in B_\infty(1)} \sum_{t=1}^T \langle \mathbf{u}_t, \boldsymbol{\theta} \rangle = o(T),$$

where  $\mathbf{u}_t := \mathbf{u}(\mathbf{w}_t, y_t)$ . There are a range of possible no-regret algorithms available, but we use the one given by Zinkevich known commonly as Online Gradient Descent (Zinkevich, 2003). The details are given in Algorithm 4. This algorithm can indeed be implemented efficiently, requiring only  $O(1)$  computation on each round and  $O(\min\{m, T\})$  memory. The main advantage is that the vectors  $\mathbf{u}_t$  are generated via our oracle above, and these vectors are *sparse*, having only at most two nonzero coordinates. Hence, the Gradient Descent Step requires only  $O(1)$  computation. In addition, the Projection Step can also be performed in an efficient manner. Since we assume that  $\boldsymbol{\theta}_t \in B_\infty(1)$ , the updated point  $\boldsymbol{\theta}'_{t+1}$  can

---

**Algorithm 4** Online Gradient Descent

---

Input: convex set  $\mathcal{K} \subset \mathbb{R}^d$   
 Initialize:  $\boldsymbol{\theta}_1 = \mathbf{0}$   
 Set Parameter:  $\eta = O(T^{-1/2})$   
**for**  $t = 1, \dots, T$  **do**  
   Receive  $\mathbf{u}_t$   
    $\boldsymbol{\theta}'_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta \mathbf{u}_t$  // Gradient Descent Step  
    $\boldsymbol{\theta}_{t+1} \leftarrow \text{Project}_2(\boldsymbol{\theta}'_{t+1}, \mathcal{K})$  // L2 Projection Step  
**end for**

---

violate at most two of the  $\ell_\infty$  constraints of the ball  $B_\infty(1)$ . An  $\ell_2$  projection onto the cube requires simply rounding the violated coordinates into  $[-1, 1]$ . The number of non-zero elements in  $\boldsymbol{\theta}$  can increase by at most two every iteration, and storing  $\boldsymbol{\theta}$  is the only state that online gradient descent needs to store, hence the algorithm can be implemented with  $O(\min\{T, m\})$  memory. We thus arrive at an efficient no-regret algorithm for choosing  $\boldsymbol{\theta}_t$ .

**Putting it all Together** We can now fully specify our calibration algorithm given the subroutines defined above. The precise description is in Algorithm 5, which makes queries to Algorithms 3 and 4.

---

**Algorithm 5** Efficient Algorithm for Asymptotic Calibration

---

Input:  $\varepsilon = 1/m$  for some natural number  $m$   
 Initialize:  $\boldsymbol{\theta}_1 = \mathbf{0}$ ,  $\mathbf{w}_1 \in \Delta_{m+1}$  arbitrarily  
**for**  $t = 1, \dots, T$  **do**  
   Sample  $i_t \sim \mathbf{w}_t$ , predict  $p_t = \frac{i_t}{m}$ , observe  $y_t \in \{0, 1\}$   
   Set  $\mathbf{u}_t := \mathbf{u}(\mathbf{w}_t, y_t)$  // Vector-valued game defined in (11)  
   Query learning algorithm:  $\boldsymbol{\theta}_{t+1} \leftarrow \text{Update}(\boldsymbol{\theta}_t | \mathbf{u}_t)$  // Subroutine from Algorithm 4  
   Query halfspace oracle:  $\mathbf{w}_{t+1} \leftarrow \mathcal{O}(\boldsymbol{\theta}_{t+1})$  // Subroutine from Algorithm 3  
**end for**

---

**Proof** [of Theorem 22] Here we have bounded the distance directly by the regret, using equation (12), which tells us that the calibration rate is bounded by the regret of the online learning algorithm. Online Gradient Descent guarantees the regret to be no more than  $DG\sqrt{T}$ , where  $D$  is the  $\ell_2$  diameter of the set, and  $G$  is the  $\ell_2$ -norm of the largest cost vector. For the ball  $B_\infty(1)$ , the diameter  $D = \sqrt{\frac{1}{\varepsilon}}$ , and we can bound the norm of our loss vectors by  $G = \sqrt{2}$ . Hence:

$$C_T^\varepsilon = \text{dist}(c_T, B_1(\varepsilon/2)) \leq \frac{\text{Regret}_T}{T} \leq \frac{GD}{\sqrt{T}} = O\left(\frac{1}{\sqrt{\varepsilon T}}\right) \quad (14)$$

■

## References

- J. Abernethy, A. Agarwal, P.L. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.
- D. Blackwell. Controlled random walks. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 336–338, 1954.
- D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):18, 1956.
- Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. ISBN 0521841089, 9780521841085.
- A. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77:605–613, 1982.
- E. Even-Dar, R. Kleinberg, S. Mannor, and Y. Mansour. Online learning for global cost functions. In *22nd Annual Conference on Learning Theory (COLT)*, 2009.
- D. P Foster. A proof of calibration via blackwell’s approachability theorem. *Games and Economic Behavior*, 29(1-2):7378, 1999.
- D. P Foster and R. V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379, 1998.
- Y. Freund and R. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.
- D. Fudenberg and D. K Levine. An easier way to calibrate\*. 1. *Games and economic behavior*, 29(1-2):131137, 1999.
- J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.
- S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):11271150, 2000.
- E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007. ISSN 0885-6125.
- Elad Hazan. The convex optimization approach to regret minimization. In *To appear in Optimization for Machine Learning*. MIT Press, 2010.
- S. Mannor and N. Shimkin. Regret minimization in repeated matrix games with variable stage duration. *Games and Economic Behavior*, 63(1):227–258, 2008. ISSN 0899-8256.
- Shie Mannor and Gilles Stoltz. A Geometric Proof of Calibration. *arXiv*, Dec 2009. URL <http://arxiv.org/abs/0912.3604>.
- J. Von Neumann, O. Morgenstern, H. W Kuhn, and A. Rubinstein. *Theory of games and economic behavior*. Princeton university press Princeton, NJ, 1947.

- V. Perchet. Calibration and internal no-regret with random signals. In *Proceedings of the 20th international conference on Algorithmic learning theory*, pages 68–82. Springer-Verlag, 2009. ISBN 3642044131.
- A. Rakhlin, K. Sridharan, and A. Tewari. Online Learning: Beyond Regret. *Arxiv preprint arXiv:1011.3168*, 2010.
- A. Sandroni, R. Smorodinsky, and R.V. Vohra. Calibration with many checking rules. *Mathematics of Operations Research*, 28(1):141–153, 2003. ISSN 0364-765X.
- S. Shalev-Shwartz and Y. Singer. Convex repeated games and Fenchel duality. *Advances in Neural Information Processing Systems*, 19:1265, 2007. ISSN 1049-5258.
- M. Sion. On general minimax theorems. *Pacific J. Math*, 8(1):171–176, 1958.
- M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, volume 20, page 928, 2003.

# Competitive Closeness Testing

**Jayadev Acharya**

**Hirakendu Das**

**Ashkan Jafarpour**

**Alon Orlitsky**

**Shengjun Pan**

*University of California San Diego, La Jolla, CA 92093*

JACHARYA@UCSD.EDU

HDAS@UCSD.EDU

AJAFARPO@UCSD.EDU

ALON@UCSD.EDU

S1PAN@UCSD.EDU

## Abstract

We test whether two sequences are generated by the same distribution or by two different ones. Unlike previous work, we make no assumptions on the distributions' support size. Additionally, we compare our performance to that of the best possible test. We describe an efficiently-computable algorithm based on *pattern* maximum likelihood that is near optimal whenever the best possible error probability is  $\leq \exp(-14n^{2/3})$  using length- $n$  sequences.

## 1. Introduction

We consider the problem of testing whether two sequences are generated by the same distribution or by two different ones. There is an extensive amount of literature on this problem and several of its variants in the framework of hypothesis testing [5, 21, 7, 11, 12], which primarily considers asymptotic error performance when the sequence lengths tend to infinity.

For non-asymptotic lengths, significant progress has been made recently under distribution property testing [2, 3, 17, 19], which provide efficient algorithms for closeness testing and other problems like entropy estimation and support size estimation using a number of samples that is sublinear in the support size. Nonetheless, these algorithms and their error performance guarantees require a priori knowledge of upper bounds on the support size. In this paper, we present closeness-testing algorithms that are competitively optimal when the best possible error probability is small. The algorithms do not require knowledge of the underlying support size. Our methods extend the technique of pattern maximum likelihood (PML) used in [14, 15] for estimating large alphabet distributions in the context of universal compression.

### 1.1. Problem definition

Let  $(p_1, p_2)$  be a pair of unknown distributions over an alphabet  $\mathcal{A} = \{a_1, a_2, \dots, a_k\}$  of size  $k$ . Two length- $n$  sequences  $\bar{X}_1, \bar{X}_2$  are generated *i.i.d.* and independently of each other according to  $p_1$  and  $p_2$  respectively. The problem is to decide whether  $p_1$  and  $p_2$  are same or different given only  $\bar{X}_1$  and  $\bar{X}_2$ . A *closeness test*  $\Delta$  for sequences in  $\mathcal{A}^n$  is a mapping  $\Delta : \mathcal{A}^n \times \mathcal{A}^n \rightarrow \{\text{same}, \text{diff}\}$  that labels each sequence pair as *same* or *diff*, indicating whether the distributions that generated them are believed to be same or different. The

error probability of  $\Delta$  for any  $(p_1, p_2)$  is the probability that it labels a sequence pair they generate incorrectly, *i.e.*,

$$P_e^n(\Delta, p_1, p_2) \stackrel{\text{def}}{=} \begin{cases} \Pr(\Delta(\bar{X}_1, \bar{X}_2) = \text{diff}) & \text{if } p_1, p_2 \text{ are same,} \\ \Pr(\Delta(\bar{X}_1, \bar{X}_2) = \text{same}) & \text{if } p_1, p_2 \text{ are different.} \end{cases}$$

The goal is to design a test  $\Delta$  that uses few samples and yet has a low error probability, both when  $(p_1, p_2)$  are same, *i.e.*,  $p_1 = p_2$  and when  $(p_1, p_2)$  are sufficiently different to be distinguishable by some test.

## 1.2. A closeness test based on empirical-frequency distributions

The closeness problem is closely related to hypothesis testing. In *simple hypothesis testing* problems, one of two known distributions  $p$  and  $q$  is chosen at random and generates a random sequence  $\bar{x}$ . Based on the sequence, we are asked to determine which of the two distributions generated it. It is well known that the *likelihood ratio test (LRT)* [5, 16] which decides on  $p$  or  $q$  depending on whether  $p(\bar{x})/q(\bar{x})$  is larger or smaller than 1 has the lowest error probability.

In *composite hypothesis testing* problem [16] there are two known distribution classes  $\mathcal{P}$  and  $\mathcal{Q}$ . One of the two classes is chosen at random and an unknown distribution from that class generates the observation. Based on the observation, we need to decide which class the generating distribution came from. As noted in [2], the closeness problem can be regarded as a composite hypothesis testing problem where the two distribution classes are  $\mathcal{P}_{\text{same}}$  containing all pairs of identical distributions  $(p, p)$ , and  $\mathcal{P}_{\text{diff}}$  containing all pairs of significantly different distributions  $(p_1, p_2)$ .

For composite hypothesis testing, we do not know which distribution to select from each class, hence often the most likely distribution in each class is estimated. The actual distributions in the LRT are replaced by their maximum likelihood estimates taken from their respective classes, and the test thereby obtained is known as the *generalized likelihood ratio test (GLRT)* [16]. Since the sequences are generated *i.i.d.*, the *empirical-frequency* distribution is the *maximum likelihood* distribution, and is known to be a good estimate of the underlying distribution when the sequence length  $n$  is large relative to the alphabet size  $k$ .

Specifically for closeness testing, let  $\mu(a)$  be the number of appearances of a symbol  $a$  in  $\bar{x}$ , and let  $\hat{P}(\bar{x}) \stackrel{\text{def}}{=} \max_p p(\bar{x}) = \prod_{a \in \mathcal{A}} \left(\frac{\mu(a)}{n}\right)^{\mu(a)}$  be the maximum likelihood of a sequence  $\bar{x} \in \mathcal{A}^n$  under all possible *i.i.d.* distributions. Note that for all  $(\bar{x}_1, \bar{x}_2)$ ,

$$\hat{P}(\bar{x}_1)\hat{P}(\bar{x}_2) = \max_{p_1, p_2} p_1(\bar{x}_1)p_2(\bar{x}_2) \geq \max_{p_1=p_2} p_1(\bar{x}_1)p_2(\bar{x}_2) = \hat{P}(\bar{x}_1\bar{x}_2),$$

hence  $\hat{P}(\bar{x}_1)\hat{P}(\bar{x}_2)/\hat{P}(\bar{x}_1\bar{x}_2) \geq 1$ . A modified empirical-frequency based GLRT test was therefore used in [10], where for all  $(\bar{x}_1, \bar{x}_2) \in \mathcal{A}^n \times \mathcal{A}^n$ ,

$$\Delta^{\text{emp}}(\bar{x}_1, \bar{x}_2) \stackrel{\text{def}}{=} \begin{cases} \text{diff} & \text{if } \frac{\hat{P}(\bar{x}_1)\hat{P}(\bar{x}_2)}{\hat{P}(\bar{x}_1\bar{x}_2)} > \binom{n+k-1}{n}^2 n, \\ \text{same} & \text{otherwise.} \end{cases}$$

They showed that when  $k = o(n)$ , if  $p_1 = p_2$  then  $\hat{P}(\bar{X}_1)\hat{P}(\bar{X}_2)/\hat{P}(\bar{X}_1\bar{X}_2)$  is small and  $\leq \binom{n+k-1}{n}^2 n$  with probability  $\geq 1 - \frac{1}{n}$ . And when the  $L_1$  distance  $|p_1 - p_2| > \epsilon$  for some  $\epsilon > 0$ ,

then  $\hat{P}(\bar{X}_1)\hat{P}(\bar{X}_2)/\hat{P}(\bar{X}_1\bar{X}_2)$  is large and  $\geq 2^{n\epsilon^2/6} > \binom{n+k-1}{n}^2 n$  with probability  $1 - o(1)$ . Hence when the alphabet size  $k$  is sublinear in  $n$ , then  $\Delta^{\text{emp}}$  has low error probability, both when  $p_1 = p_2$  and when  $|p_1 - p_2| > \epsilon$  for some constant  $\epsilon > 0$ .

However, when the alphabet size is larger than  $n$ , empirical distribution may not be a good estimate of the underlying distribution and  $\Delta^{\text{emp}}$  may not have low error probability, as shown in an example in [10] and in the following, simpler, example.

**Example 1** For large  $n$  and  $k = n^3$ , let  $p_1(a_1) = 1$  and  $p_1(a_2) = \dots = p_1(a_n) = 0$ , and let  $p_2(a_1) = 1/2$  and  $p_2(a_2) = \dots = p_2(a_n) = 1/(2(k-1))$ . The two distributions are clearly very different and  $|p_1 - p_2| = 1$ . If  $\bar{X}_1$  and  $\bar{X}_2$  are length- $n$  sequences generated *i.i.d.* according to  $p_1$  and  $p_2$  respectively, then  $\bar{X}_1 = a_1^n$  and  $\bar{X}_2 = a_1^{\frac{n}{2}}a_2a_3 \cdots a_{\frac{n}{2}+1}$  are typical sequences. In particular, by the Birthday problem, with high probability no symbol in  $\{a_2, a_3, \dots, a_k\}$  appears more than once in  $\bar{X}_2$ . It follows that typically,

$$\frac{\hat{P}(\bar{X}_1)\hat{P}(\bar{X}_2)}{\hat{P}(\bar{X}_1\bar{X}_2)} = \frac{\hat{P}(a_1^n)\hat{P}(a_1^{\frac{n}{2}}a_2a_3 \cdots a_{\frac{n}{2}+1})}{\hat{P}(a_1^{\frac{3n}{2}}a_2a_3 \cdots a_{\frac{n}{2}+1})} = \frac{1^n \times (\frac{1}{2})^{\frac{n}{2}}(\frac{1}{n})^{\frac{n}{2}}}{(\frac{3}{4})^{\frac{3n}{2}}(\frac{1}{2n})^{\frac{n}{2}}} = \left(\frac{4}{3}\right)^{\frac{3n}{2}} \approx 1.54^n,$$

suggesting as it should that the sequences were generated by different distributions.

However, when both  $\bar{X}_1$  and  $\bar{X}_2$  are generated according to the same distribution,  $p_2$ , then typically  $\bar{X}_1 = a_1^{\frac{n}{2}}a_2a_3 \cdots a_{\frac{n}{2}+1}$  and  $\bar{X}_2 = a_1^{\frac{n}{2}}a_{\frac{n}{2}+2} \cdots a_{n+1}$  where no symbol in  $\{a_2, a_3, \dots, a_k\}$  appears more than once in  $\bar{X}_1\bar{X}_2$ . Then,

$$\frac{\hat{P}(\bar{X}_1)\hat{P}(\bar{X}_2)}{\hat{P}(\bar{X}_1\bar{X}_2)} = \frac{\hat{P}(a_1^{\frac{n}{2}}a_2a_3 \cdots a_{\frac{n}{2}+1})\hat{P}(a_1^{\frac{n}{2}}a_{\frac{n}{2}+2} \cdots a_{n+1})}{\hat{P}(a_1^n a_2a_3 \cdots a_{n+1})} = \frac{(\frac{1}{2})^{\frac{n}{2}}(\frac{1}{n})^{\frac{n}{2}} \times (\frac{1}{2})^{\frac{n}{2}}(\frac{1}{n})^{\frac{n}{2}}}{(\frac{1}{2})^n(\frac{1}{2n})^n} = 2^n,$$

an even higher ratio than when the distributions were different.

Therefore, for any choice of the threshold  $t$ , the GLRT test  $\hat{P}(\bar{X}_1)\hat{P}(\bar{X}_2)/\hat{P}(\bar{X}_1\bar{X}_2)$  will have a high error for at either  $(p_1, p_2)$  or  $(p_2, p_2)$ . Furthermore, note that when  $\bar{X}_1, \bar{X}_2$  are both generated according to  $p_2$ , the sequences  $\bar{X}_1, \bar{X}_2$  have very different empirical distribution estimates than  $\bar{X}_1\bar{X}_2$ .  $\square$

### 1.3. Related work on estimating large alphabet distributions

Batu et al [2] developed a test that distinguishes between two distributions that are close and those that are well separated in  $L_1$  distance using sequences whose length is sublinear in size of the underlying alphabet. Using sequences of length  $n = \mathcal{O}(k^{2/3} \log k \cdot \epsilon^{-4} \cdot \log \frac{1}{\delta})$ , their algorithm outputs *same* when  $|p_1 - p_2| \leq \max(\frac{\epsilon}{32k^{1/3}}, \frac{\epsilon}{4k^{1/2}})$  and *diff* when  $|p_1 - p_2| > \epsilon$  with error probability  $\leq \delta$  for both cases. Since the empirical frequency is a good estimate for large probabilities, the algorithm estimates the  $L_1$  distance contribution of only the high-probability symbols using their empirical frequencies. The contribution of low probability symbols is estimated using a test for  $L_2$  distance that relies on the number of collisions (also known as coincidences or repetitions) in the sequences. They establish a corresponding lower bound by showing pairs of distributions  $(p_1, p_2)$  such that  $|p_1 - p_2| > \epsilon$  and that no algorithm can distinguish it from the identical pair  $(p_1, p_1)$  using  $n = o(k^{2/3} \cdot \epsilon^{-2/3})$  samples. Valiant

[19] further showed that distinguishing distribution pairs with  $L_1$  distance less than  $\alpha$  from those with distance greater than  $\beta$  for  $0 < \alpha < \beta < 2$  requires  $n = k^{1-o(1)}$  samples and can be done using  $n = \tilde{\mathcal{O}}(k)$  samples by [2] or by another test shown in [19]. Although no assumptions are made on the structure of distributions, the tests in [2, 19] and their sample complexities still depend on the knowledge of an upper bound on the alphabet size  $k$  of the unknown underlying distributions. Moreover, as in Example 1, there are many distribution pairs that can be tested for closeness in much less than  $\tilde{\mathcal{O}}(k^{2/3})$  samples.

The related problem of classification was considered by many researchers, including recently by Kelly et al [10]. Here, one is given training sequences  $\bar{X}_1$  and  $\bar{X}_2 \in \mathcal{A}^n$  generated *i.i.d.* and independently according to unknown distributions  $p_1$  and  $p_2$  that are separated in  $L_1$  distance. A third sequence  $\bar{Y} \in \mathcal{A}^n$  is generated *i.i.d.* and independently of each other according to either  $p_1$  or  $p_2$  with equal probability and the problem is to decide whether  $\bar{Y}$  is generated according  $p_1$  or  $p_2$ . They show a test that has low error probability when  $(p_1, p_2)$  belong to a restricted class of distributions such that the probabilities of all symbols are  $\Theta(\frac{1}{k})$  and  $k = \Theta(n^\alpha)$ , for any fixed  $\alpha \in [0, 2)$ . Their test uses the  $L_2$  distance between the empirical frequency distributions, of the sequences to determine which one of the pairs  $(\bar{X}_1, \bar{Y})$  or  $(\bar{X}_2, \bar{Y})$  are closer and classify accordingly.

The problem of estimating the probability multiset of large alphabet distributions was also studied in the context of universal compression of large alphabet sources in [14, 15]. The main idea is to consider the *pattern* of a sequence, which conveys only the structure of the sequence and the order in which symbols appear in the sequence, and not the identities of the actual symbols. The pattern contains all the information that is needed to test symmetric properties like entropy that depend only on the probability multiset and not on the way in which the probabilities are associated with the symbols of the alphabet. In [14], several estimators based on the maximum likelihood of patterns were shown that estimate the pattern probabilities (that are usually exponentially small in  $n$ ) to within a factor that is subexponential in the sequence length  $n$ , regardless of the alphabet size and the structure of the underlying distribution. Preliminary results on application of such estimators to the problem of classification were shown in [18]. Partial results on classifiers based on maximum likelihood estimation of the *joint pattern* of two or more sequences were shown in [1]. In this paper, we show closeness tests based on maximum likelihood of joint patterns that perform almost as good as any test can, without making any assumptions on the underlying distributions. These tests can be used as good classifiers as well.

#### 1.4. Closeness tests based on pattern maximum likelihood

The pattern of a sequence is defined as follows. Let  $\bar{x} = x_1 x_2 \cdots x_n = x_1^n \in \mathcal{A}^n$  be a sequence of length  $n$  and  $\mathcal{A}(\bar{x})$  denote the set of symbols that appear in  $\bar{x}$ . The index  $\iota_{\bar{x}}(a)$  of a symbol  $a \in \mathcal{A}(\bar{x})$  is

$$\iota_{\bar{x}}(a) \stackrel{\text{def}}{=} \min\{|\mathcal{A}(x_1^i)| : 1 \leq i \leq n \text{ and } x_i = a\},$$

*i.e.*, one more than the number of distinct symbols that have appeared before the first appearance of  $a$  in  $\bar{x}$ . The *pattern* of  $\bar{x}$  is the sequence

$$\Psi(\bar{x}) \stackrel{\text{def}}{=} \iota_{\bar{x}}(x_1) \iota_{\bar{x}}(x_2) \cdots \iota_{\bar{x}}(x_n)$$

obtained by replacing the symbols in  $\bar{x}$  by their respective indices. For example, if  $\bar{x} = \text{abracadabra}$ , then  $\iota_{\bar{x}}(\text{a}) = 1$ ,  $\iota_{\bar{x}}(\text{b}) = 2$ ,  $\iota_{\bar{x}}(\text{r}) = 3$ ,  $\iota_{\bar{x}}(\text{c}) = 4$  and  $\iota_{\bar{x}}(\text{d}) = 5$ . Hence,  $\Psi(\text{abracadabra}) = 12314151231$ . The set of all possible patterns of different length- $n$  sequences is represented by  $\Psi^n$ . For example,  $\Psi^1 = \{1\}$ ,  $\Psi^2 = \{11, 12\}$  and  $\Psi^3 = \{111, 112, 121, 122, 123\}$ .

We extend the definition of patterns to two or more sequences. The *joint pattern* of a pair of sequences  $(\bar{x}_1, \bar{x}_2) \in \mathcal{A}^{n_1} \times \mathcal{A}^{n_2}$  is  $\Psi(\bar{x}_1, \bar{x}_2) \stackrel{\text{def}}{=} (\bar{\psi}_1, \bar{\psi}_2)$ , where  $\bar{\psi}_1 = \Psi(\bar{x}_1)$  and  $\bar{\psi}_1 \bar{\psi}_2 = \Psi(\bar{x}_1 \bar{x}_2)$ . For example, for  $\text{bab}$  and  $\text{abca}$ , the first pattern is  $\Psi(\text{bab}) = 121$  and that of the concatenated sequence is  $\Psi(\text{bababca}) = 1212132$ , hence the joint pattern is  $\Psi(\text{bab}, \text{abca}) = (121, 2132)$ . Clearly, the joint pattern conveys the patterns of the individual sequences and the association between the symbols of the sequences. The joint pattern of a list of three or more sequences is defined similarly. We use  $\Psi^{n_1, n_2}$  to denote the set of all possible joint patterns of pairs of sequences of length  $(n_1, n_2)$ . For example,  $\Psi^{2,1} = \{(11, 1), (11, 2), (12, 1), (12, 2), (12, 3)\}$ .

The probability of a single pattern  $\bar{\psi} \in \Psi^n$  under a distribution  $p$  is the probability that a length- $n$  sequence  $\bar{X}$  generated *i.i.d.* according to  $p$  has pattern  $\bar{\psi}$ , *i.e.*,

$$p(\bar{\psi}) \stackrel{\text{def}}{=} p(\Psi(\bar{X}) = \bar{\psi}) = \sum_{\bar{x}: \Psi(\bar{x}) = \bar{\psi}} p(\bar{x}).$$

Similarly, the probability of a joint pattern  $(\bar{\psi}_1, \bar{\psi}_2) \in \Psi^{n_1, n_2}$  under a pair of distributions  $(p_1, p_2)$  is the probability that two sequences  $\bar{X}_1$  and  $\bar{X}_2$  of length  $n_1$  and  $n_2$  generated *i.i.d.* according to  $p_1$  and  $p_2$  respectively have joint pattern  $(\bar{\psi}_1, \bar{\psi}_2)$  and is denoted by

$$p_{1,2}(\bar{\psi}_1, \bar{\psi}_2) = p_{1,2}(\Psi(\bar{X}_1, \bar{X}_2) = (\bar{\psi}_1, \bar{\psi}_2)) = \sum_{\substack{(\bar{x}_1, \bar{x}_2): \\ \Psi(\bar{x}_1, \bar{x}_2) = (\bar{\psi}_1, \bar{\psi}_2)}} p_1(\bar{x}_1)p_2(\bar{x}_2).$$

For example, if  $\mathcal{A} = \{\text{a, b, c, d}\}$  and  $p = (p_{\text{a}}, p_{\text{b}}, p_{\text{c}}, p_{\text{d}})$ , then the probability of the pattern 1213 is

$$p(1213) = p(\text{abac}) + p(\text{abad}) + p(\text{acab}) + \dots = p_{\text{a}}^2 p_{\text{b}} p_{\text{c}} + p_{\text{a}}^2 p_{\text{b}} p_{\text{d}} + p_{\text{a}}^2 p_{\text{c}} p_{\text{b}} + \dots.$$

Similarly, if  $p_1 = (p_{\text{a}}, p_{\text{b}}, p_{\text{c}}, p_{\text{d}})$  and  $p_2 = (p'_{\text{a}}, p'_{\text{b}}, p'_{\text{c}}, p'_{\text{d}})$ , then probability of the pattern (12, 13) is

$$p_{1,2}(12, 13) = p_{1,2}(\text{ab, ac}) + p_{1,2}(\text{ab, ad}) + p_{1,2}(\text{ba, bc}) + \dots = p_{\text{a}} p_{\text{b}} p'_{\text{a}} p'_{\text{c}} + p_{\text{a}} p_{\text{b}} p'_{\text{a}} p'_{\text{d}} + \dots.$$

Notice that if  $(\bar{\psi}_1, \bar{\psi}_2) \in \Psi^{n_1, n_2}$ , then  $\bar{\psi}_1 \bar{\psi}_2 \in \Psi^{n_1+n_2}$ . Also, if  $p_1 = p_2 = p$ , then  $p_{1,2}(\bar{\psi}_1, \bar{\psi}_2) = p_{1,1}(\bar{\psi}_1, \bar{\psi}_2) = p_1(\bar{\psi}_1 \bar{\psi}_2)$ .

The maximum likelihood of a pattern  $\bar{\psi}$  under all *i.i.d.* distributions is  $\hat{P}(\bar{\psi}) \stackrel{\text{def}}{=} \max_p p(\bar{\psi})$ .

Similarly, the maximum likelihood of a joint pattern  $(\bar{\psi}_1, \bar{\psi}_2)$  under all pairs of *i.i.d.* and independent distributions is denoted by  $\hat{P}(\bar{\psi}) \stackrel{\text{def}}{=} \max_{p_1, p_2} p_{1,2}(\bar{\psi})$ .

Since joint patterns contain all the relevant information for closeness testing, consider a simple hypothesis testing problem where a sequence pair  $(\bar{X}_1, \bar{X}_2) \in \mathcal{A}^n \times \mathcal{A}^n$

is generated according to either  $(p_1, p_2)$  or  $(p, p)$ , but we are given only the joint pattern  $\Psi(\bar{X}_1, \bar{X}_2)$  and not the actual sequences. In this case, the likelihood ratio test  $p_{1,2}(\Psi(\bar{X}_1, \bar{X}_2)) \stackrel{\text{diff}}{\underset{\text{same}}{\gtrless}} p(\Psi(\bar{X}_1 \bar{X}_2))$  is a test with minimum error probability. Hence, similar to Subsection 1.2, viewing closeness testing as a composite hypothesis testing problem with the joint pattern of the sequences given as the observations, we consider the test  $\Delta^{\hat{P}(\Psi)} \stackrel{\text{def}}{=} \Delta_{n,\delta}^{\hat{P}(\Psi)}$  defined as

$$\Delta_{n,\delta}^{\hat{P}(\Psi)}(\bar{x}_1, \bar{x}_2) \stackrel{\text{def}}{=} \begin{cases} \text{diff} & \text{if } \frac{\hat{P}(\Psi(\bar{x}_1, \bar{x}_2))}{\hat{P}(\Psi(\bar{x}_1 \bar{x}_2))} > \frac{1}{\sqrt{\delta}}, \\ \text{same} & \text{otherwise,} \end{cases}$$

for all  $(\bar{x}_1, \bar{x}_2) \in \mathcal{A}^n \times \mathcal{A}^n$  and for some  $\delta < \exp(-12n^{2/3})$ . In other words, the test outputs *diff* if the maximum likelihood of the pattern of the two sequences under two different distributions is much higher than that under two identical distributions.

Without loss of generality, we consider only *symmetric* tests, namely those whose output depends only on joint pattern of the sequences and not the specific symbols that have appeared, since the property of closeness depends only on the probability multiset and not the associated symbols. (See also Appendix C for a discussion along the lines of [4].) We say that a pair of distributions  $(p_1, p_2)$  is  $(n, \delta)$ -*different* if there exists a symmetric test that can distinguish with error probability  $< \delta$ , pairs of length  $n$  sequences generated according to  $(p_1, p_2)$  from those generated by any pair of identical distributions  $(p, p)$ . In other words, there exists a test  $\Delta$  such that for all  $p$ ,

$$P_e^n(\Delta, p_1, p_2) < \delta \quad \text{and} \quad P_e^n(\Delta, p, p) < \delta.$$

Our first main result, Theorem 7, states that for all  $\delta \leq \exp(-12n^{2/3})$ , the test  $\Delta^{\hat{P}(\Psi)}$  has error probability  $\leq \sqrt{\delta} \exp(6n^{2/3})$  both when the two distributions are identical and when they are  $(n, \delta)$ -different.

Revisiting Example 1, in the case when  $(\bar{X}_1, \bar{X}_2) \sim (p_1, p_2)$ , consider the typical sequence pair  $(\bar{X}_1, \bar{X}_2) = (a_1^n, a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1})$ . Then,  $\hat{P}(\Psi(\bar{X}_1, \bar{X}_2)) = \hat{P}(1^n, 1^{\frac{n}{2}} 23 \cdots (\frac{n}{2} + 1)) \geq 1 \cdot (\frac{1}{2})^{\frac{n}{2}} (\frac{1}{2})^{\frac{n}{2}} = (\frac{1}{2})^n$ , since the distributions  $(p'_1, p'_2)$  assign  $\Psi(\bar{X}_1, \bar{X}_2)$  such a likelihood, where  $p'_1(a_1) = 1$ ,  $p'_2(a_1) = \frac{1}{2}$ , and the remaining probability  $\frac{1}{2}$  of  $p'_2$  is spread over a continuous alphabet or a large tail, similar to  $p_2$ . Also, from [13],  $\hat{P}(\Psi(\bar{X}_1 \bar{X}_2)) = \hat{P}(1^{\frac{3n}{2}} 23 \cdots (\frac{n}{2}+1)) = (\frac{3}{4})^{\frac{3n}{2}} (\frac{1}{4})^{\frac{n}{2}}$ , which is attained by the distribution  $p$  such that  $p(a_1) = \frac{3}{4}$  and has the remaining probability  $\frac{1}{4}$  spread over a continuous alphabet. Hence,

$$\frac{\hat{P}(\Psi(\bar{X}_1, \bar{X}_2))}{\hat{P}(\Psi(\bar{X}_1 \bar{X}_2))} \geq \frac{(\frac{1}{2})^n}{(\frac{3}{4})^{\frac{3n}{2}} (\frac{1}{4})^{\frac{n}{2}}} = \left(\frac{4}{3}\right)^{\frac{3n}{2}} > 1.53^n,$$

and the test  $\Delta^{\hat{P}(\Psi)}$  outputs *diff* for  $\delta = \exp(-14n^{2/3})$ . When  $(\bar{X}_1, \bar{X}_2) \sim (p_2, p_2)$ , for the typical sequence pair  $(\bar{X}_1, \bar{X}_2) = (a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1}, a_1^{\frac{n}{2}} a_{\frac{n}{2}+2} \cdots a_{n+1})$ , again as shown by [13],  $\hat{P}(\Psi(\bar{X}_1, \bar{X}_2)) \leq \hat{P}(\Psi(\bar{X}_1) \hat{P}(\Psi(\bar{X}_2)) = \hat{P}(1^{\frac{n}{2}} 23 \cdots (\frac{n}{2} + 1))^2 = ((\frac{1}{2})^{\frac{n}{2}} (\frac{1}{2})^{\frac{n}{2}})^2 = (\frac{1}{2})^{2n}$ , and  $\hat{P}(\Psi(\bar{X}_1 \bar{X}_2)) = \hat{P}(1^n 23 \cdots (n + 1)) = (\frac{1}{2})^n (\frac{1}{2})^n = (\frac{1}{2})^{2n}$ . Hence, in this case

$$\frac{\hat{P}(\Psi(\bar{X}_1, \bar{X}_2))}{\hat{P}(\Psi(\bar{X}_1 \bar{X}_2))} = 1,$$

and the output of  $\Delta^{\hat{P}(\Psi)}$  is *same*. We note that the maximum likelihood distributions of  $\Psi(\bar{X}_1, \bar{X}_2)$  and of  $\Psi(\bar{X}_1 \bar{X}_2)$  are consistent, *i.e.*, same, unlike in the case of  $\Delta^{\text{emp}}$ .

As evident from the previous example, the computation of pattern maximum likelihood (PML) is difficult in general and hence we show an efficient test based on pattern probability estimators that also has low error probability. Several such estimators were shown in [14] which can compute maximum likelihood of patterns to within a subexponential factor. In particular, we consider the following estimator. The *profile* of a pattern or a sequence conveys the number of symbols appearing a given number of times in it. For example, the profile of **abdb** is  $\varphi(\text{abdb}) = (\varphi_1, \varphi_2, \varphi_3, \varphi_4) = (2, 1, 0, 0)$ , indicating that there are  $\varphi_1 = 2$  symbols that appear once in **abdb** and  $\varphi_2 = 1$  symbol that appears 2 times and so on. The sequences **abdb** and **dcca** for example have the same profile, though their patterns are different. The definition of a profile can be similarly extended to joint patterns or pairs of sequences and consists of entries  $\varphi_{\mu_1, \mu_2}$  that are the number of symbols that have appeared  $\mu_1$  times in first sequence and  $\mu_2$  times in the second sequence. For example,

$$\varphi(\text{dac}, \text{adbda}) = \varphi(123, 21412) = \begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 2 \end{array},$$

where the prevalances  $\varphi_{\mu_1, \mu_2}$  are arranged in a matrix with the rows indexed with  $\mu_1$  and columns with  $\mu_2$ . As seen in the matrix,  $\varphi_{1,2} = 2$ , since there are 2 symbols, namely **d** and **a** that appear  $\mu_1 = 1$  times in **dac** and  $\mu_2 = 2$  times in **adbda**. By convention, we set  $\varphi_{0,0} \equiv 0$ .

Let  $N(\varphi)$  be the number of patterns with the same profile  $\varphi$  and  $\Phi^n$  be the set of all distinct profiles of sequences of length  $n$ . It was shown in [14] that the probability estimator for  $\bar{\psi} \in \Psi^n$ ,

$$q(\bar{\psi}) \stackrel{\text{def}}{=} \frac{1}{|\Phi^n|} \frac{1}{N(\varphi(\bar{\psi}))},$$

which assigns equal probability estimate to all profiles and equal estimate to all patterns within a profile, is a good estimate for pattern maximum likelihood, *i.e.*,  $q(\bar{\psi}) \geq \hat{p}(\bar{\psi}) \exp(-\pi\sqrt{2n/3})$ .

We consider a similar estimator for maximum likelihood of joint patterns. Namely, denoting the number of joint patterns with the same profile  $\varphi$  by  $N(\varphi)$  and the set of all distinct profiles of length- $(n, n)$  sequences by  $\Phi^{n,n}$ , we estimate the probability of a joint pattern  $(\bar{\psi}_1, \bar{\psi}_2) \in \Psi^{n,n}$  as

$$q_{\text{jp}}(\bar{\psi}_1, \bar{\psi}_2) \stackrel{\text{def}}{=} \frac{1}{|\Phi^{n,n}|} \frac{1}{N(\varphi(\bar{\psi}_1, \bar{\psi}_2))}.$$

We use the estimators  $q$  and  $q_{\text{jp}}$  instead of the pattern maximum likelihoods in  $\Delta^{\hat{P}(\Psi)}$  and consider the test  $\Delta_{n,\delta}^{N(\varphi)}$  defined for  $\delta < \exp(-14n^{2/3})$  and  $(\bar{x}_1, \bar{x}_2) \in \mathcal{A}^n \times \mathcal{A}^n$  by

$$\Delta_{n,\delta}^{N(\varphi)}(\bar{x}_1, \bar{x}_2) \stackrel{\text{def}}{=} \begin{cases} \text{diff} & \text{if } \frac{N(\varphi(\bar{x}_1 \bar{x}_2))}{N(\varphi(\bar{x}_1, \bar{x}_2))} > \frac{1}{\sqrt{\delta}}, \\ \text{same} & \text{otherwise.} \end{cases}$$

Our second main result, Theorem 12, shows that  $(p_1, p_2)$  are identical and when they are  $(n, \delta)$ -different, the test  $\Delta^{N(\varphi)}$  error probability is upper bounded by

$$P_{\text{e,sym}}(\Delta^{N(\varphi)}, p_1, p_2) \leq \sqrt{\delta} \exp(7n^{2/3}).$$

In the process, we show a convexity result for profile probabilities, that resembles the convexity of KL-divergence.

For  $\varphi \in \Phi^n$ ,  $N(\varphi)$  can be calculated by the expressions [14]

$$N(\varphi) = \frac{n!}{\prod_{\mu=1}^n (\mu!)^{\varphi_\mu} \varphi_\mu!}.$$

As shown in Appendix B, for  $\varphi \in \Phi^{n,n}$ ,

$$N(\varphi) = \frac{(n!)^2}{\prod_{\mu_1, \mu_2} (\mu_1! \mu_2!)^{\varphi_{\mu_1, \mu_2}} \varphi_{\mu_1, \mu_2}!},$$

Hence, for  $(\bar{\psi}_1, \bar{\psi}_2) \in \Psi^{n,n}$ , the quantity  $\frac{N(\varphi(\bar{\psi}_1, \bar{\psi}_2))}{N(\varphi(\bar{\psi}_1, \bar{\psi}_2))}$  can be evaluated efficiently with time and space complexity both  $\mathcal{O}(n)$ .

Consider Example 1 again, this time using the test  $\Delta^{N(\varphi)}$ . When  $(\bar{X}_1, \bar{X}_2) \sim (p_1, p_2)$  and  $\Psi(\bar{X}_1, \bar{X}_2) = (\bar{\psi}_1, \bar{\psi}_2) = (1^n, 1^{\frac{n}{2}} 2 3 \cdots (\frac{n}{2} + 1))$ , the profile  $\varphi = \varphi(\bar{\psi}_1, \bar{\psi}_2)$  has  $\varphi_{0,1} = \frac{n}{2}$ ,  $\varphi_{n, \frac{n}{2}} = 1$  and all other  $\varphi_{\mu_1, \mu_2} = 0$ . And the profile  $\varphi' = \varphi(\bar{\psi}_1 \bar{\psi}_2)$  has  $\varphi'_1 = \frac{n}{2}$ ,  $\varphi'_{\frac{3n}{2}} = 1$  and all other  $\varphi'_\mu = 0$ . Hence, by Stirling approximation,

$$\frac{N(\varphi(\bar{X}_1, \bar{X}_2))}{N(\varphi(\bar{X}_1, \bar{X}_2))} = N(\varphi')/N(\varphi) = \frac{(2n)!}{(\frac{3n}{2})! \cdot (\frac{n}{2})!} / \frac{(n!)^2}{n!(\frac{n}{2})! \cdot (\frac{n}{2})!} \approx \left(\frac{4}{3}\right)^{\frac{3}{2n}} > 1.53^n,$$

and the test  $\Delta_{n, \delta}^{N(\varphi)}$  outputs *diff* for a suitable  $\delta$  as in the case of  $\Delta_{n, \delta}^{\hat{P}(\Psi)}$ , say  $\delta = \exp(-16n^{2/3})$ . When  $(\bar{X}_1, \bar{X}_2) \sim (p_2, p_2)$ , and  $\Psi(\bar{X}_1, \bar{X}_2) = (\bar{\psi}_1, \bar{\psi}_2) = (1^{\frac{n}{2}} 2 3 \cdots a_{\frac{n}{2}+1}, 1^{\frac{n}{2}} (\frac{n}{2} + 2) \cdots (n + 1))$ ,

$$\frac{N(\varphi(\bar{X}_1, \bar{X}_2))}{N(\varphi(\bar{X}_1, \bar{X}_2))} = \frac{(2n)!}{n! n!} / \frac{(n!)^2}{(\frac{n}{2})! (\frac{n}{2})! \cdot (\frac{n}{2})! \cdot (\frac{n}{2})!} \approx \frac{\sqrt{\pi n}}{2}$$

and the output of  $\Delta_{n, \delta}^{N(\varphi)}$  is *same* for  $\delta = \exp(-16n^{2/3})$ .

While the error probability results that we show for the tests  $\Delta_{n, \delta}^{N(\varphi)}$  and  $\Delta_{n, \delta}^{\hat{P}(\Psi)}$  are useful only when  $\delta < \exp(-14n^{2/3})$ , for higher values of  $\delta$ , we can characterize their performance in terms of *sample complexity*. Corollary 14 shows that if  $(p_1, p_2)$  are  $(n, \delta)$ -different for some  $\delta < \frac{1}{4}$ , then for  $\delta' = \delta^2 \exp(-14n'^{2/3})$ , the test  $\Delta_{n', \delta'}^{N(\varphi)}$  also has error probability less than  $\delta$  when given sequences of length

$$n' = \max \left\{ 19n, \frac{120000n^3}{(\log_2 \frac{1}{4\delta})^3} \right\}.$$

In particular, if  $\delta < \exp(-19n^{2/3})$ , the error probability of  $\Delta^{N(\varphi)}$  is less than  $\delta$  when given  $n' = 19n$  samples.

## 2. Error analysis of the test $\Delta^{\hat{P}(\Psi)}$

In order to analyze the error probability of  $\Delta_{n,\delta}^{\hat{P}(\Psi)}$ , we show some ancillary results on profiles of joint patterns and their probabilities.

We begin by showing that  $|\Phi^{n,n}|$ , the number of profiles of joint patterns, is subexponential in the sequence length. To count the number of profiles  $|\Phi^{n_1,n_2,\dots,n_d}|$ , we relate it to *partitions* of  $(n_1, n_2, \dots, n_d)$ . We say that a multiset of  $d$ -tuples of non-negative integers  $\{(\mu_{1,i}, \mu_{2,i}, \dots, \mu_{d,i})\}_{i=1}^m$  is an (unordered) partition of  $(n_1, n_2, \dots, n_d)$  if  $\sum_{i=1}^m \mu_{j,i} = n_j$  for  $j = 1, 2, \dots, d$ . The sum of two  $d$ -tuples denotes their component-wise sum, *i.e.*,  $(\mu_1, \mu_2, \dots, \mu_d) + (\mu'_1, \mu'_2, \dots, \mu'_d) \stackrel{\text{def}}{=} (\mu_1 + \mu'_1, \mu_2 + \mu'_2, \dots, \mu_d + \mu'_d)$ . The product of a scalar with a  $d$ -tuple is component-wise product with the scalar, *i.e.*,  $\alpha \cdot (\mu_1, \mu_2, \dots, \mu_d) \stackrel{\text{def}}{=} (\alpha \cdot \mu_1, \alpha \cdot \mu_2, \dots, \alpha \cdot \mu_d)$ . For example,  $\{(0,1), (0,1), (2,1)\}$  is an unordered partition of  $(2,3)$ , because  $2 \cdot (0,1) + (2,1) = (2,3)$ .

We denote the number of partitions of  $(n_1, n_2, \dots, n_d)$  by the *joint partition function*  $P(n_1, n_2, \dots, n_d)$ . For example,  $P(2,1) = 4$ , since

$$(2,1) = (1,0) + (1,1) = (2,0) + (0,1) = 2 \cdot (1,0) + (0,1).$$

**Observation 1** For all  $d \geq 1$  and non-negative integers  $n_1, n_2, \dots, n_d$ ,

$$|\Phi^{n_1,n_2,\dots,n_d}| = P(n_1, n_2, \dots, n_d).$$

□

It is a well known result due to Hardy and Ramanujan [8, 9] that for all  $n$ , the partition function  $P(n)$  is bounded as

$$\exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{n}(1-o(1))\right) \leq P(n) < \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{n}\right).$$

The following lemma shows an upper bound on  $P(n_1, n_2, \dots, n_d)$ , similar to [20, 6].

**Lemma 2** For all  $d \geq 1$  and all  $n_1, n_2, \dots, n_d \geq 2^{d+1}$ ,

$$P(n_1, n_2, \dots, n_d) \leq \exp\left(2\left(1 + \frac{1}{d}\right)\sum_{j=1}^d n_j^{d/(d+1)}\right).$$

**Proof** See for example Appendix A. □

**Corollary 3** For all  $d \geq 1$  and  $n \geq 2^{d+1}$ ,

$$|\Phi^{n,n,\dots \text{ (d times)}}| = P(\underbrace{n, \dots, n}_d) < \exp\left(2(d+1)n^{d/(d+1)}\right).$$

□

Let  $(\bar{X}_1, \bar{X}_2) \in \mathcal{A}^{n_1} \times \mathcal{A}^{n_2}$  be generated *i.i.d.* and independently according to  $(p_1, p_2)$  respectively. The probability of a profile  $\varphi \in \Phi^{n_1, n_2}$  under  $(p_1, p_2)$  is the probability of observing a pair of sequences with that profile, *i.e.*,

$$p_{1,2}(\varphi) \stackrel{\text{def}}{=} p_{1,2}\left(\varphi(\bar{X}_1, \bar{X}_2) = \varphi\right) = \sum_{(\bar{\psi}_1, \bar{\psi}_2): \varphi(\bar{\psi}_1, \bar{\psi}_2) = \varphi} p_{1,2}(\bar{\psi}_1, \bar{\psi}_2).$$

Joint patterns with the same profile have the same probability when the sequences are generated by *i.i.d.* distributions. Hence, for all  $(\bar{\psi}_1, \bar{\psi}_2)$ ,

$$p_{1,2}(\varphi(\bar{\psi}_1, \bar{\psi}_2)) = N(\varphi(\bar{\psi}_1, \bar{\psi}_2)) \cdot p_{1,2}(\bar{\psi}_1, \bar{\psi}_2).$$

The following lemma provides a simple bound on the probability of generating sequences whose profile has low probability.

**Observation 4** Let  $(\bar{X}_1, \bar{X}_2) \in \mathcal{A}^{n_1} \times \mathcal{A}^{n_2}$  be generated *i.i.d.* according to  $(p_1, p_2)$  respectively, where  $n_1, n_2, \geq 8$ . Then, for all  $0 < \delta \leq 1$ ,

$$\Pr(p_{1,2}(\varphi(\bar{X}_1, \bar{X}_2)) < \delta) < \delta \exp(3(n_1^{2/3} + n_2^{2/3})).$$

**Proof** From Lemma 15,

$$\Pr(p_{1,2}(\varphi(\bar{X}_1, \bar{X}_2)) < \delta) = \sum_{\varphi: p_{1,2}(\varphi) < \delta} p_{1,2}(\varphi) < |\Phi^{n_1, n_2}| \cdot \delta \leq \delta \exp(3(n_1^{2/3} + n_2^{2/3})). \quad \square$$

**Observation 5** Let  $(\bar{X}_1, \bar{X}_2) \in \mathcal{A}^n \times \mathcal{A}^n$  be generated *i.i.d.* according to  $(p_1, p_2)$ , where  $n \geq 8$ . Then, for all  $0 < \delta \leq 1$ ,

$$\Pr(p_{1,2}(\varphi(\bar{X}_1, \bar{X}_2)) < \delta) < \delta \exp(6n^{2/3}). \quad \square$$

We make the following observation on  $(n, \delta)$ -different distributions before we proceed to analyze the error probability of  $\Delta^{\hat{P}(\Psi)}$ .

**Observation 6** Let  $(p_1, p_2)$  be  $(n, \delta)$ -different distributions over  $\mathcal{A}$ , and let  $\varphi \in \Phi^{n,n}$  be a profile such that  $p_{1,2}(\varphi) \geq \delta$ . Then, for all distributions  $p_3$  over  $\mathcal{A}$ ,  $p_{3,3}(\varphi) < \delta$ .

**Proof** Suppose on the contrary, there exists a distribution  $p_3$  such that  $p_{3,3}(\varphi) \geq \delta$ . Any symmetric test  $\Delta$  labels all sequence pairs with profile  $\varphi$  either *same* or *diff*. If it labels them *same*, then  $P_e^n(\Delta, p_1, p_2) \geq \delta$  and if it labels them *diff*, then  $P_e^n(\Delta, p_3, p_3) \geq \delta$ , *i.e.*, one of the error probabilities is  $\geq \delta$ , which contradicts the fact that  $(p_1, p_2)$  are  $(n, \delta)$ -different.  $\square$

The following theorem upper bounds the error probability of the test  $\Delta_{n,\delta}^{\hat{P}(\Psi)}$ .

**Theorem 7** For all  $n \geq 8$ , all  $0 < \delta < \exp(-12n^{2/3})$ , and all pairs distributions  $(p_1, p_2)$  that are either *same* or  $(n, \delta)$ -different,

$$P_e^n(\Delta_{n,\delta}^{\hat{P}(\Psi)}, p_1, p_2) < \sqrt{\delta} \exp(6n^{2/3}).$$

**Proof** Let  $(\bar{X}_1, \bar{X}_2) \sim p_1^n \times p_2^n$ . Consider the case when the  $(p_1, p_2)$  are same, i.e.,  $p_1 = p_2$ . Then,

$$\begin{aligned} P_e^n(\Delta^{\hat{P}(\Psi)}, p_1, p_1) &= \Pr\left(\frac{\hat{p}(\Psi(\bar{X}_1, \bar{X}_2))}{\hat{p}(\Psi(\bar{X}_1 \bar{X}_2))} > \frac{1}{\sqrt{\delta}}\right) \\ &\stackrel{(a)}{=} \Pr\left(\frac{\hat{p}(\Psi(\bar{X}_1, \bar{X}_2))}{p_{3,3}(\Psi(\bar{X}_1, \bar{X}_2))} > \frac{1}{\sqrt{\delta}}\right) \\ &\stackrel{(b)}{=} \Pr\left(\frac{\hat{p}(\varphi(\bar{X}_1, \bar{X}_2))}{p_{3,3}(\varphi(\bar{X}_1, \bar{X}_2))} > \frac{1}{\sqrt{\delta}}\right) \\ &\stackrel{(c)}{\leq} \Pr\left(\frac{1}{p_{1,1}(\varphi(\bar{X}_1, \bar{X}_2))} > \frac{1}{\sqrt{\delta}}\right) \\ &\stackrel{(d)}{<} \sqrt{\delta} \exp(6n^{2/3}), \end{aligned}$$

where in (a),  $p_3 = \arg \max_p p(\Psi(\bar{X}_1 \bar{X}_2))$  and in (b), we convert pattern probabilities to profile probabilities by multiplying and dividing by  $N(\varphi(\bar{X}_1, \bar{X}_2))$  and using  $p(\varphi) = N(\varphi)p(\bar{\psi}_1, \bar{\psi}_2)$ . For (c), we use that  $\hat{p}(\varphi(\bar{X}_1, \bar{X}_2)) \leq 1$  and we use Observation 5 for (d).

Now consider the case when  $(p_1, p_2)$  are  $(n, \delta)$ -different. For a sequence pair  $(\bar{X}_1, \bar{X}_2)$ , let  $p_3 = \arg \max_p p(\Psi(\bar{X}_1 \bar{X}_2))$ . Then,

$$\begin{aligned} \Pr\left(\frac{\hat{p}(\Psi(\bar{X}_1, \bar{X}_2))}{\hat{p}(\Psi(\bar{X}_1 \bar{X}_2))} \leq \frac{1}{\sqrt{\delta}}\right) &\leq \Pr\left(\frac{p_{1,2}(\Psi(\bar{X}_1, \bar{X}_2))}{p_{3,3}(\Psi(\bar{X}_1, \bar{X}_2))} \leq \frac{1}{\sqrt{\delta}}\right) \\ &= \Pr\left(\frac{p_{1,2}(\varphi(\bar{X}_1, \bar{X}_2))}{p_{3,3}(\varphi(\bar{X}_1, \bar{X}_2))} \leq \frac{1}{\sqrt{\delta}}\right) \\ &< \sqrt{\delta} \exp(6n^{2/3}). \end{aligned}$$

For the last step, in the case when  $p_{1,2}(\varphi) \geq \sqrt{\delta}$ , there is no error since Observation 6 implies that for all  $p_3$ ,  $p_{3,3}(\varphi) < \delta$  and hence  $\frac{p_{1,2}(\varphi(\bar{X}_1, \bar{X}_2))}{p_{3,3}(\varphi(\bar{X}_1, \bar{X}_2))} > \frac{\sqrt{\delta}}{\delta} = \frac{1}{\sqrt{\delta}}$ . Hence, the error probability is bounded by the probability of the case when  $p_{1,2}(\varphi) < \sqrt{\delta}$ , which by Observation 5 is  $< \sqrt{\delta} \exp(6n^{2/3})$ .  $\square$

### 3. Error analysis of the test $\Delta^{N(\varphi)}$

As mentioned in Section 1, direct computation of maximum likelihood of patterns in the test  $\Delta^{\hat{P}(\Psi)}$  may be difficult and hence we look at a computationally easier test  $\Delta^{N(\varphi)}$ . We now show a few more useful results for analyzing the error probability of  $\Delta^{N(\varphi)}$ , which relate the quantities  $N(\varphi)$ , the number of patterns in a profile and  $\hat{P}(\varphi)$ , the maximum likelihood of the profile under *i.i.d.* distributions.

The *type* of a sequence  $\bar{x} \in \mathcal{A}^n$  is the vector of multiplicities  $\tau(\bar{x}) \stackrel{\text{def}}{=} (\mu(a_1), \mu(a_2), \dots, \mu(a_k))$ , where  $\mu(a_i)$  is the number of appearances of  $a_i$  in  $\bar{x}$  for  $i = 1, 2, \dots, k$ . Similarly, the *joint type*\* of a pair of sequences  $(\bar{x}_1, \bar{x}_2) \in \mathcal{A}^{n_1} \times \mathcal{A}^{n_2}$  is the vector of multiplicity

---

\*. This definition of joint type is different from that used in the *method of types* in information theory.

pairs  $\tau(\bar{x}_1, \bar{x}_2) \stackrel{\text{def}}{=} ((\mu_1(a_1), \mu_2(a_1)), (\mu_1(a_2), \mu_2(a_2)), \dots, (\mu_1(a_k), \mu_2(a_k)))$ , where  $\mu_1(a_i)$  and  $\mu_2(a_i)$  are the number of appearances of  $a_i$  in  $\bar{x}_1$  and  $\bar{x}_2$  for  $i = 1, 2, \dots, k$ . The set of all possible distinct types of sequences in  $\mathcal{A}^n$  is denoted by  $\mathcal{T}^n$  and the set of all possible distinct joint types of sequences in  $\mathcal{A}^{n_1} \times \mathcal{A}^{n_2}$  is denoted by  $\mathcal{T}^{n_1, n_2}$ .

The probability of a type  $\tau = (\mu(a_i))_{i=1}^k \in \mathcal{T}^n$  under a distribution  $p$  over  $\mathcal{A}$  is

$$p(\tau) \stackrel{\text{def}}{=} \sum_{\tau(\bar{x})=\tau} p(\bar{x}) = \binom{n}{\mu(a_1), \mu(a_2), \dots, \mu(a_k)} \prod_{i=1}^k p(a_i)^{\mu(a_i)},$$

i.e., the probability of observing a sequence whose type is  $\tau$ . Similarly, the probability of a joint type  $\tau = ((\mu_1(a_i), \mu_2(a_i)))_{i=1}^k \in \mathcal{T}^{n_1, n_2}$  under a pair of distributions  $(p_1, p_2)$  over  $\mathcal{A}$  is

$$\begin{aligned} p_{1,2}(\tau) &\stackrel{\text{def}}{=} \sum_{\tau(\bar{x}_1, \bar{x}_2)=\tau} p_{1,2}(\bar{x}_1, \bar{x}_2) \\ &= \binom{n_1}{\mu_1(a_1), \mu_1(a_2), \dots, \mu_1(a_k)} \binom{n_2}{\mu_2(a_1), \mu_2(a_2), \dots, \mu_2(a_k)} \prod_{i=1}^k p_1(a_i)^{\mu_1(a_i)} p_2(a_i)^{\mu_2(a_i)}. \end{aligned}$$

The *sum type* of a joint type  $\tau = ((\mu_1(a_i), \mu_2(a_i)))_{i=1}^k \in \mathcal{T}^{n,n}$  is  $\tau_s(\tau) \stackrel{\text{def}}{=} (\mu(a_i))_{i=1}^k \in \mathcal{T}^{2n}$ , where  $\mu(a_i) \stackrel{\text{def}}{=} \mu_1(a_i) + \mu_2(a_i)$  for  $i = 1, 2, \dots, k$ . The probability of a (sum) type  $\tau' \in \mathcal{T}^{2n}$  under a pair of distributions  $p_{1,2} = (p_1, p_2)$  is the probability of the set of all types  $\tau \in \mathcal{T}^{n,n}$  such that  $\tau_s(\tau) = \tau'$ , i.e.,

$$p_{1,2}(\tau') \stackrel{\text{def}}{=} \sum_{\substack{\tau \in \mathcal{T}^{n,n}: \\ \tau_s(\tau) = \tau'}} p_{1,2}(\tau).$$

For any pair of distributions  $(p_1, p_2)$  over  $\mathcal{A} \times \mathcal{A}$ ,  $p_{1/2} \stackrel{\text{def}}{=} (p_1 + p_2)/2$  denotes the distribution over  $\mathcal{A}$  such that  $p_{1/2}(a_i) = (p_1(a_i) + p_2(a_i))/2$  for  $i = 1, 2, \dots, k$ .

**Observation 8** For all types  $\tau' \in \mathcal{T}^{2n}$  and all  $(p_1, p_2)$ ,

$$\sum_{\substack{\tau \in \mathcal{T}^{n,n}: \\ \tau_s(\tau) = \tau'}} p_{1,2}(\tau) = p_{1,2}(\tau') \leq p_{1/2}(\tau') \frac{(n!)^2 2^{2n}}{(2n)!} < p_{1/2}(\tau') \sqrt{\pi n} e^{\frac{1}{6n}}.$$

**Proof** Let  $\tau' = (\mu(a_i))_{i=1}^k$ . Then,

$$\begin{aligned} p_{1,2}(\tau') &= \sum_{\substack{\tau \in \mathcal{T}^{n,n}: \\ \tau_s(\tau) = \tau'}} p_{1,2}(\tau) \\ &= \sum_{\substack{(\mu_1(a_1), \dots, \mu_1(a_k)): \\ 0 \leq \mu_1(a_i) \leq \mu(a_i) \text{ for } i=1, \dots, k, \\ \text{and } \mu_1(a_1) + \dots + \mu_1(a_k) = n}} n! n! \prod_{i=1}^k \frac{1}{\mu_1(a_i)! (\mu(a_i) - \mu_1(a_i))!} p_1(a_i)^{\mu_1(a_i)} p_2(a_i)^{\mu(a_i) - \mu_1(a_i)} \\ &= \frac{n! n!}{\prod_{i=1}^k \mu(a_i)!} \sum_{\substack{(\mu_1(a_1), \dots, \mu_1(a_k)): \\ 0 \leq \mu_1(a_i) \leq \mu(a_i) \text{ for } i=1, \dots, k, \\ \text{and } \mu_1(a_1) + \dots + \mu_1(a_k) = n}} \prod_{i=1}^k \binom{\mu(a_i)}{\mu_1(a_i)} p_1(a_i)^{\mu_1(a_i)} p_2(a_i)^{\mu(a_i) - \mu_1(a_i)} \end{aligned}$$

$$\begin{aligned}
 &\leq \frac{n!n!}{\prod_{i=1}^k \mu(a_i)!} \sum_{\substack{(\mu_1(a_1), \dots, \mu_1(a_k)): \\ 0 \leq \mu_1(a_i) \leq \mu(a_i) \text{ for } i=1, \dots, k}} \prod_{i=1}^k \binom{\mu(a_i)}{\mu_1(a_i)} p_1(a_i)^{\mu_1(a_i)} p_2(a_i)^{\mu(a_i) - \mu_1(a_i)} \\
 &= \frac{n!n!}{\prod_{i=1}^k \mu(a_i)!} \prod_{i=1}^k \left( \sum_{\mu_1(a_i)=0}^{\mu(a_i)} \binom{\mu(a_i)}{\mu_1(a_i)} p_1(a_i)^{\mu_1(a_i)} p_2(a_i)^{\mu(a_i) - \mu_1(a_i)} \right) \\
 &= \frac{n!n!}{\prod_{i=1}^k \mu(a_i)!} \prod_{i=1}^k (p_1(a_i) + p_2(a_i))^{\mu(a_i)} \\
 &= \frac{(n!)^2 2^{2n}}{(2n)!} \binom{2n}{\mu(a_1), \mu(a_2), \dots, \mu(a_k)} \prod_{i=1}^k \left( \frac{p_1(a_i) + p_2(a_i)}{2} \right)^{\mu(a_i)} \\
 &= \frac{(n!)^2 2^{2n}}{(2n)!} p_{1/2}(\tau'). \tag*{$\square$}
 \end{aligned}$$

The profile of a type  $\tau \in \mathcal{T}^n$  is  $\varphi(\tau) = \varphi(\bar{x})$ , where  $\bar{x}$  is any sequence whose type is  $\tau(\bar{x}) = \tau$ . Similarly, for any  $\tau \in \mathcal{T}^{n_1, n_2}$ ,  $\varphi(\tau) \stackrel{\text{def}}{=} \varphi(\bar{x}_1, \bar{x}_2)$ , where  $(\bar{x}_1, \bar{x}_2)$  is any sequence pair such that  $\tau(\bar{x}_1, \bar{x}_2) = \tau$ .

**Observation 9** For all profiles  $\varphi \in \Phi^n$  and all distributions  $p$ ,

$$p(\varphi) = \sum_{\tau \in \mathcal{T}^n: \varphi(\tau) = \varphi} p(\tau).$$

Likewise, for all profiles  $\varphi \in \Phi^{n_1, n_2}$  and all pairs of distributions  $(p_1, p_2)$ ,

$$p_{1,2}(\varphi) = \sum_{\tau \in \mathcal{T}^{n_1, n_2}: \varphi(\tau) = \varphi} p_{1,2}(\tau). \tag*{$\square$}$$

The *sum profile* of a profile  $\varphi \in \Phi^{n,n}$  is  $\varphi_s(\varphi) \stackrel{\text{def}}{=} \varphi(\bar{\psi}_1 \bar{\psi}_2) \in \Phi^{2n}$  where  $(\bar{\psi}_1, \bar{\psi}_2)$  is any joint pattern having profile  $\varphi(\bar{\psi}_1, \bar{\psi}_2) = \varphi$ . Hence, if  $\varphi = [\varphi_{\mu_1, \mu_2}]$ , where  $\mu_1 = 0, 1, \dots, n$  and  $\mu_2 = 0, 1, \dots, n$ , then  $\varphi_s(\varphi) = (\varphi_1, \varphi_2, \dots, \varphi_{2n})$  is given by  $\varphi_\mu = \sum_{\mu_1 + \mu_2 = \mu} \varphi_{\mu_1, \mu_2}$ . The probability of a (sum) profile  $\varphi' \in \Phi^{2n}$  under a pair of distributions  $p_{1,2}$  is the probability  $p_{1,2}$  assigns to the set of all profiles  $\varphi \in \Phi^{n,n}$  such that  $\varphi_s(\varphi) = \varphi'$ , i.e.,

$$p_{1,2}(\varphi') \stackrel{\text{def}}{=} \sum_{\substack{\varphi \in \Phi^{n,n}: \\ \varphi_s(\varphi) = \varphi'}} p_{1,2}(\varphi).$$

The following lemma on profile probabilities is analogous to the convexity of KL-divergence.

**Lemma 10** For all  $\varphi' \in \Phi^{2n}$  and all  $(p_1, p_2)$ ,

$$\sum_{\substack{\varphi \in \Phi^{n,n}: \\ \varphi_s(\varphi) = \varphi'}} p_{1,2}(\varphi) = p_{1,2}(\varphi') \leq p_{1/2}(\varphi') \frac{(n!)^2 2^{2n}}{2n!} < p_{1/2}(\varphi') \sqrt{\pi n} e^{\frac{1}{6n}}.$$

**Proof** Using Observations 8 and 9,

$$\begin{aligned}
 p_{1,2}(\varphi') &= \sum_{\substack{\varphi \in \Phi^{n,n}: \\ \varphi_s(\varphi) = \varphi'}} p_{1,2}(\varphi) \\
 &= \sum_{\substack{\tau \in \mathcal{T}^{n,n}: \\ \varphi_s(\varphi(\tau)) = \varphi(\tau_s(\tau)) = \varphi'}} p_{1,2}(\tau) \\
 &= \sum_{\substack{\tau' \in \mathcal{T}^{2n}: \\ \varphi(\tau') = \varphi'}} p_{1,2}(\tau') \\
 &\leq \sum_{\substack{\tau' \in \mathcal{T}^{2n}: \\ \varphi(\tau') = \varphi'}} \frac{(n!)^2 2^{2n}}{(2n)!} p_{1/2}(\tau') \\
 &= p_{1/2}(\varphi') \frac{(n!)^2 2^{2n}}{(2n)!}.
 \end{aligned}$$

□

The following Lemma 11 relates the ratio of maximum likelihoods of any joint pattern  $(\bar{\psi}_1, \bar{\psi}_2)$  and its concatenated pattern  $\bar{\psi}_1 \bar{\psi}_2$  which appear in the test  $\Delta^{\hat{P}(\Psi)}$ , to the ratio of counts of patterns in their respective profiles, i.e.,  $N(\varphi(\bar{\psi}_1, \bar{\psi}_2))$  and  $N(\varphi(\bar{\psi}_1 \bar{\psi}_2))$  that appear in the test  $\Delta^{N(\varphi)}$ .

**Lemma 11** For all joint patterns  $(\bar{\psi}_1, \bar{\psi}_2) \in \Psi^{n,n}$ ,

$$\frac{N(\varphi(\bar{\psi}_1 \bar{\psi}_2))}{N(\varphi(\bar{\psi}_1, \bar{\psi}_2))} \geq \frac{\hat{p}(\bar{\psi}_1, \bar{\psi}_2)}{\hat{p}(\bar{\psi}_1 \bar{\psi}_2)} \frac{(2n)!}{(n!)^2 2^{2n}} > \frac{\hat{p}(\bar{\psi}_1, \bar{\psi}_2)}{\hat{p}(\bar{\psi}_1 \bar{\psi}_2)} \frac{1}{\sqrt{\pi n e^{\frac{1}{6n}}} \cdot \sqrt{\pi n e^{\frac{1}{6n}}}}.$$

**Proof** Let  $p_{1,2} = (p_1, p_2)$  be such that  $\hat{p}(\bar{\psi}_1, \bar{\psi}_2) = p_{1,2}(\bar{\psi}_1, \bar{\psi}_2)$ . Note that  $\varphi_s(\varphi(\bar{\psi}_1, \bar{\psi}_2)) = \varphi(\bar{\psi}_1 \bar{\psi}_2)$ . Using Lemma 10, we have

$$\begin{aligned}
 N(\varphi(\bar{\psi}_1, \bar{\psi}_2)) \hat{p}(\bar{\psi}_1, \bar{\psi}_2) &= N(\varphi(\bar{\psi}_1, \bar{\psi}_2)) p_{1,2}(\bar{\psi}_1, \bar{\psi}_2) \\
 &= p_{1,2}(\varphi(\bar{\psi}_1, \bar{\psi}_2)) \\
 &\leq p_{1,2}(\varphi_s(\varphi(\bar{\psi}_1, \bar{\psi}_2))) \\
 &\leq p_{1/2}(\varphi_s(\varphi(\bar{\psi}_1, \bar{\psi}_2))) \frac{(n!)^2 2^{2n}}{(2n)!} \\
 &= p_{1/2}(\varphi(\bar{\psi}_1 \bar{\psi}_2)) \frac{(n!)^2 2^{2n}}{(2n)!} \\
 &\leq \hat{p}(\varphi(\bar{\psi}_1 \bar{\psi}_2)) \frac{(n!)^2 2^{2n}}{(2n)!} \\
 &= N(\varphi(\bar{\psi}_1 \bar{\psi}_2)) \hat{p}(\bar{\psi}_1 \bar{\psi}_2) \frac{(n!)^2 2^{2n}}{(2n)!}.
 \end{aligned}$$

□

**Theorem 12** For all  $n \geq 8$ , all  $0 < \delta < \frac{1}{4\pi n e^{1/3n}} \exp(-12n^{2/3})$ , and all pairs distributions  $(p_1, p_2)$  that are either same or  $(n, \delta)$ -different,

$$P_e^n(\Delta^{N(\varphi)}, p_1, p_2) < \sqrt{\delta} \exp(6n^{2/3}) \sqrt{\pi n e^{\frac{1}{6n}}}.$$

**Proof** Let  $(\bar{X}_1, \bar{X}_2) \sim p_1^n \times p_2^n$ . Consider the case when  $p_1 = p_2$ . Then,

$$\begin{aligned} P_e^n(\Delta^{N(\varphi)}, p_1, p_1) &= \Pr\left(\frac{N(\varphi(\bar{X}_1 \bar{X}_2))}{N(\varphi(\bar{X}_1, \bar{X}_2))} > \frac{1}{\sqrt{\delta}}\right) \\ &\stackrel{(a)}{=} \Pr\left(\frac{p_1(\varphi(\bar{X}_1 \bar{X}_2))}{p_{1,1}(\varphi(\bar{X}_1, \bar{X}_2))} > \frac{1}{\sqrt{\delta}}\right) \\ &\leq \Pr\left(\frac{1}{p_{1,1}(\varphi(\bar{X}_1, \bar{X}_2))} > \frac{1}{\sqrt{\delta}}\right) \\ &< \sqrt{\delta} \exp(6n^{2/3}), \end{aligned}$$

where in (a), we used  $\frac{N(\varphi(\bar{\psi}_1 \bar{\psi}_2))}{N(\varphi(\bar{\psi}_1, \bar{\psi}_2))} = \frac{N(\varphi(\bar{\psi}_1 \bar{\psi}_2))p_1(\bar{\psi}_1 \bar{\psi}_2)}{N(\varphi(\bar{\psi}_1, \bar{\psi}_2)p_{1,1}(\bar{\psi}_1, \bar{\psi}_2))} = \frac{p_1(\varphi(\bar{\psi}_1 \bar{\psi}_2))}{p_{1,1}(\varphi(\bar{\psi}_1, \bar{\psi}_2))}$  and the last inequality is due to Observation 5.

Consider the case when  $(p_1, p_2)$  are  $(n, \delta)$ -different. For a sequence pair  $(\bar{X}_1, \bar{X}_2)$ , let  $p_3 = \arg \max_p p(\Psi(\bar{X}_1 \bar{X}_2))$ . Then,

$$\begin{aligned} P_e^n(\Delta^{N(\varphi)}, p_1, p_2) &= \Pr\left(\frac{N(\varphi(\bar{X}_1 \bar{X}_2))}{N(\varphi(\bar{X}_1, \bar{X}_2))} \leq \frac{1}{\sqrt{\delta}}\right) \\ &\stackrel{(a)}{\leq} \Pr\left(\frac{1}{\sqrt{\pi n e^{\frac{1}{6n}}}} \frac{\hat{p}(\Psi(\bar{X}_1, \bar{X}_2))}{\hat{p}(\Psi(\bar{X}_1 \bar{X}_2))} \leq \frac{1}{\sqrt{\delta}}\right) \\ &\leq \Pr\left(\frac{p_{1,2}(\Psi(\bar{X}_1, \bar{X}_2))}{p_{3,3}(\Psi(\bar{X}_1 \bar{X}_2))} \leq \frac{\sqrt{\pi n e^{\frac{1}{6n}}}}{\sqrt{\delta}}\right) \\ &< \sqrt{\delta} \exp(6n^{2/3}) \sqrt{\pi n e^{\frac{1}{6n}}}, \end{aligned}$$

where in (a), we used Lemma 11. For the last inequality, we again consider the cases  $p_{1,2}(\varphi(\bar{X}_1, \bar{X}_2)) \geq \sqrt{\delta} \sqrt{\pi n e^{\frac{1}{6n}}}$  and  $< \sqrt{\delta} \sqrt{\pi n e^{\frac{1}{6n}}}$  separately similar to the proof of Theorem 12. In the case when  $p_{1,2}(\varphi(\bar{X}_1, \bar{X}_2)) \geq \sqrt{\delta} \sqrt{\pi n e^{\frac{1}{6n}}} > \delta$ , Observation 6 implies  $p_{3,3}(\varphi(\bar{X}_1, \bar{X}_2)) < \delta$ . Hence,  $\frac{p_{1,2}(\varphi(\bar{X}_1, \bar{X}_2))}{p_{3,3}(\varphi(\bar{X}_1, \bar{X}_2))} > \frac{\sqrt{\delta} \sqrt{\pi n e^{\frac{1}{6n}}}}{\delta} = \frac{\sqrt{\pi n e^{\frac{1}{6n}}}}{\sqrt{\delta}}$  and hence, this case does not contribute to error probability. The error probability is therefore bounded by the probability of the other case, which by Observation 5 is bounded as  $\Pr(p_{1,2}(\varphi(\bar{X}_1, \bar{X}_2)) < \sqrt{\delta} \sqrt{\pi n e^{\frac{1}{6n}}}) < \sqrt{\delta} \exp(6n^{2/3}) \sqrt{\pi n e^{\frac{1}{6n}}}$ .  $\square$

#### 4. Sample complexity of closeness testing

The error analysis results of Theorems 7 and 12 can be rephrased in terms of sample complexity. Also, Theorems 7 and 12 are applicable only when  $\delta \leq \exp(-14n^{2/3})$ , and this section partially addresses the general case when  $\delta < \frac{1}{2}$ .

**Observation 13** If  $(p_1, p_2)$  are  $(n, \delta)$ -different distributions for some  $0 < \delta < \frac{1}{2}$ , then they are also  $(n', \delta')$ -different, where

$$n' = \min \left\{ 20n, \frac{15000n^3}{D(\frac{1}{2}||\delta)^3} \right\} \text{ and } \delta' \leq \delta^2 \exp(14n'^{2/3}),$$

where  $D(\delta_1 || \delta_2) \stackrel{\text{def}}{=} \delta_1 \log \frac{\delta_1}{\delta_2} + (1 - \delta_1) \log \frac{1 - \delta_1}{1 - \delta_2}$ .

**Proof sketch** Since  $(p_1, p_2)$  are  $(n, \delta)$ -different, for any  $p_3$  there is a test that can distinguish  $(p_1, p_2)$  and  $(p_3, p_3)$  with error probability  $< \delta$ . We can obtain another test for sequences of length  $n' = (2r+1)n$  such that the error probability of this test is  $\delta' = \sum_{i=r+1}^{2r+1} \delta^i \binom{2r+1}{i} (1 - \delta)^{2r+1-i}$  by using the original test on  $(2r+1)$  pairs of length- $n$  sequences and outputting the majority decision. It can be verified that  $(2r+1) \geq \min\{19, \frac{15000n^2}{D(\frac{1}{2}||\delta)^3}\}$  suffices to guarantee that  $\sum_{i=r+1}^{2r+1} \delta^i \binom{2r+1}{i} (1 - \delta)^{2r+1-i} \leq \delta^2 \exp(14((2r+1)n)^{2/3})$ .  $\square$

**Corollary 14** If  $(p_1, p_2)$  are  $(n, \delta)$ -different distributions for some  $0 < \delta < \frac{1}{4}$ , then they are also  $(n', \delta')$ -different where  $\delta' \leq \delta^2 \exp(14n'^{2/3})$  for  $n' = \max \left\{ 19n, \frac{120000n^3}{(\log_2 \frac{1}{4\delta})^3} \right\}$ . Furthermore if  $\delta < \exp(-19n^{2/3})$ , then  $n' = 19n$  suffices.  $\square$

Hence, using Theorem 12 and Corollary 14, it follows that whenever  $(p_1, p_2)$  are identical or  $(n, \delta)$ -different, the error probability of the test  $\Delta_{n', \delta'}^{N(\varphi)}$  is less than  $\delta$ , using sequences of length  $n' = \max \left\{ 19n, \frac{120000n^3}{(\log_2 \frac{1}{4\delta})^3} \right\}$ , where  $\delta' = \delta^2 \exp(14n'^{2/3})$ .

## 5. Related and open problems

For the problem of classification described in 1.3, our results imply that whenever the distributions of the classes,  $p_1$  and  $p_2$ , are  $(n, \delta)$ -different, the closeness tests  $\Delta_{n', \delta'}^{\tilde{P}(\Psi)}$  or  $\Delta_{n', \delta'}^{N(\varphi)}$  can be used to construct classifiers whose error probability is  $\leq \sqrt{\delta} \exp(7n^{2/3})$ . We define two distributions  $(p_1, p_2)$  to be  $(n, \delta)$ -classifiable if length- $n$  sequence pairs generated by  $(p_1, p_2)$  can be distinguished with error probability  $< \delta$  from those generated by  $(p_1, p_1)$  and  $(p_2, p_2)$  by a symmetric test. While  $(n, \delta)$ -different implies  $(n, \delta)$ -classifiable, it remains to answer if the opposite is also true.

As mentioned earlier, our results are applicable when the error probabilities  $\delta \leq \exp(-14n^{2/3})$ , and while we partially address the case of general  $\delta < \frac{1}{2}$ , and it remains to perform a better analysis. We also hope to reduce the subexponential factor of  $\exp(7n^{2/3})$  in the right hand side of Theorems 7 and 12 using a tighter analysis.

Lastly, it remains to fully answer the question of when two distributions  $(p_1, p_2)$  are  $(n, \delta)$ -different. In many cases such as Example 1, the quantity  $\frac{N(\varphi(\bar{X}_1 \bar{X}_2))}{N(\varphi(\bar{X}_1, \bar{X}_2))}$  in the test  $\Delta_{n', \delta'}^{N(\varphi)}$  can be shown to be exponentially large in  $n$  with high probability, that implies  $(n, \delta)$ -difference for a suitable  $\delta$ . This question is also answered in part by [2] and [19] where distributions are parametrized in terms of alphabet size.

## References

- [1] J. Acharya, H. Das, A. Orlitsky, S. Pan, and N.P. Santhanam. Classification using pattern probability estimators. In *Proceedings of IEEE Symposium on Information Theory*, pages 1493–1497, 2010.
- [2] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, page 259, Washington, DC, USA, 2000. IEEE Computer Society. ISBN 0-7695-0850-2.
- [3] T. Batu, L. Fortnow, E. Fischer, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. *FOCS '01: Proceedings of the 42nd Annual Symposium on Foundations of Computer Science*, page 442, 2001.
- [4] Tugkan Batu. *Testing properties of distributions*. PhD thesis, Cornell University, 2001.
- [5] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Interscience, 2nd edition, 2006.
- [6] A.K. Dhulipala and A. Orlitsky. Universal compression of markov and related sources over arbitrary alphabets. *IEEE Transactions on Information Theory*, 53:4182–4190, 2006.
- [7] M. Gutman. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Transactions on Information Theory*, 35:401–408, 1989.
- [8] G.H. Hardy and S. Ramanujan. Asymptotic formulae in combinatory analysis. *Proceedings of London Mathematics Society*, 17(2):75–115, 1918.
- [9] G.H. Hardy and E.M. Wright. *An introduction to the theory of numbers*. Oxford University Press, 1985.
- [10] B. Kelly, T. Tularak, A. B. Wagner, and P. Viswanath. Universal hypothesis testing in the learning-limited regime. In *Proceedings of IEEE Symposium on Information Theory*, pages 1478–1482, 2010.
- [11] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.
- [12] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.
- [13] A. Orlitsky and Shengjun Pan. The maximum likelihood probability of skewed patterns. In *Proceedings of IEEE Symposium on Information Theory*, pages 1130–1134, 2009.
- [14] A. Orlitsky, N.P. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, 50:1469–1481, 2004.

- [15] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Limit results on pattern entropy. *IEEE Transactions on Information Theory*, 52:2954–2964, 2006.
- [16] H. V. Poor. *An introduction to signal detection and estimation*. New York: Springer-Verlag, 2nd edition, 1994.
- [17] Sofya Raskhodnikova. *Property Testing: Theory and Applications*. PhD thesis, Massachusetts Institute of Technology, 2003.
- [18] N.P. Santhanam, A. Orlitsky, and K. Viswanathan. New tricks for old dogs: Large alphabet probability estimation. In *Information Theory Workshop*, pages 638–643, 2007.
- [19] Paul Valiant. Testing symmetric properties of distributions. In *STOC '08: Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 383–392, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-047-0.
- [20] J.H. van Lint and R.M. Wilson. *A course in combinatorics*. Cambridge University Press, 2001.
- [21] J. Ziv. On classification with empirically observed statistics and universal data compression. *IEEE Transactions on Information Theory*, 34:278–286, 1988.

## Appendix A. Number of profiles of a given length

**Lemma 15** For all  $d \geq 1$  and all  $n_1, n_2, \dots, n_d \geq 2^{d+1}$ ,

$$P(n_1, n_2, \dots, n_d) \leq \exp \left( 2 \left( 1 + \frac{1}{d} \right) \sum_{j=1}^d n_j^{d/(d+1)} \right).$$

**Proof** The (ordinary) generating function of  $P(n_1, n_2, \dots, n_d)$  is

$$G(x_1, x_2, \dots, x_d) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \cdots \sum_{n_d=0}^{\infty} P(n_1, n_2, \dots, n_d) x_1^{n_1} x_2^{n_2} \cdots x_d^{n_d} = \prod_{\substack{(\mu_1, \mu_2, \dots, \mu_d) \\ \in \mathbb{N}^d \setminus (0, 0, \dots, 0)}} \frac{1}{1 - x_1^{\mu_1} x_2^{\mu_2} \cdots x_d^{\mu_d}},$$

where  $\mathbb{N} = \{0, 1, 2, \dots\}$  and  $0 < x_1, x_2, \dots, x_d < 1$ . Hence,

$$\begin{aligned} \log G(x_1, x_2, \dots, x_d) &= \sum_{\substack{(\mu_1, \mu_2, \dots, \mu_d) \\ \in \mathbb{N}^d \setminus (0, 0, \dots, 0)}} -\log \left( 1 - \prod_{j=1}^d x_j^{\mu_j} \right) \\ &= \sum_{\substack{(\mu_1, \mu_2, \dots, \mu_d) \\ \in \mathbb{N}^d \setminus (0, 0, \dots, 0)}} \sum_{l=1}^{\infty} \frac{1}{l} \left( \prod_{j=1}^d x_j^{\mu_j} \right)^l \\ &= \sum_{l=1}^{\infty} \frac{1}{l} \sum_{\substack{(\mu_1, \mu_2, \dots, \mu_d) \\ \in \mathbb{N}^d \setminus (0, 0, \dots, 0)}} \prod_{j=1}^d (x_j^l)^{\mu_j} \\ &= \sum_{l=1}^{\infty} \frac{1}{l} \left( \frac{1}{\prod_{j=1}^d (1 - x_j^l)} - 1 \right) \\ &= \sum_{l=1}^{\infty} \frac{1}{l} \frac{1 - \prod_{j=1}^d (1 - x_j^l)}{\prod_{j=1}^d ((1 - x_j) (\sum_{i=0}^{l-1} x_j^i))} \\ &< \sum_{l=1}^{\infty} \frac{1}{l} \frac{1 - \prod_{j=1}^d (1 - x_j^l)}{\left( \prod_{j=1}^d (1 - x_j) \right) \left( 1 + \sum_{j=1}^d \sum_{i=1}^{l-1} x_j^i \right)} \\ &\stackrel{(a)}{<} \frac{1}{\prod_{j=1}^d (1 - x_j)} \left( 1 + \sum_{l=2}^{\infty} \frac{1}{l(l-1)} \right) \\ &= \frac{2}{\prod_{j=1}^d (1 - x_j)}. \end{aligned}$$

In the Inequality (a), we consider the cases  $l = 1$  and  $l > 1$  separately. When  $l > 1$ , in the denominator,  $\left( 1 + \sum_{j=1}^d \sum_{i=1}^{l-1} x_j^i \right) > (l-1) \sum_{j=1}^d x_j^i > (l-1) \left( 1 - \prod_{j=1}^d (1 - x_j^l) \right)$ . Since  $G(x_1, x_2, \dots, x_d) > P(n_1, n_2, \dots, n_d) x_1^{n_1} x_2^{n_2} \cdots x_d^{n_d}$ , we have

$$\log P(n_1, n_2, \dots, n_d) < \log G(x_1, x_2, \dots, x_d) - \sum_{j=1}^d n_j \log x_j < \frac{2}{\prod_{j=1}^d (1 - x_j)} - \sum_{j=1}^d n_j \log x_j.$$

Substituting  $x_j = 1 - n_j^{-1/(d+1)}$  for  $j = 1, 2, \dots, d$ , we get

$$\log P(n_1, n_2, \dots, n_d) < 2 \prod_{j=1}^d n_j^{1/(d+1)} + \sum_{j=1}^d n_j \log(1 - n_j^{-1/(d+1)}) \leq 2 \left(1 + \frac{1}{d}\right) \sum_{j=1}^d n_j^{d/(d+1)}.$$

In the last step, we used AM-GM inequality, i.e.,  $\prod_{j=1}^d n_j^{1/(d+1)} = (\prod_{j=1}^d n_j^{d/(d+1)})^{1/d} \leq \frac{1}{d} \sum_{j=1}^d n_j^{d/(d+1)}$ , and  $\log(1 - \epsilon) < 2\epsilon$  for  $\epsilon \leq \frac{1}{2}$ , hence  $\log(1 - n_j^{-1/(d+1)}) \leq 2n_j^{-1/(d+1)}$  for  $n_j > 2^{d+1}$  and  $j = 1, 2, \dots, d$ .  $\square$

## Appendix B. Number of patterns of a given profile

The number of joint patterns with the same profile  $\varphi$  is denoted by  $N(\varphi)$ . For example, consider the profile  $\varphi = \varphi(1232, 13)$  which has  $\varphi_{1,1} = 2$ ,  $\varphi_{2,0} = 1$  and all other  $\varphi_{\mu_1, \mu_2} = 0$ . Then,  $N(\varphi) = 12$  since the set of all joint patterns that have this profile is  $\{(1123, 23), (1123, 32), (1213, 23), (1213, 32), (1223, 13), (1223, 31), (1231, 23), (1231, 32), (1232, 13), (1232, 31), (1233, 13), (1233, 21)\}$ . The following lemma gives an expression for  $N(\varphi)$  and extends Lemma 3 in [14].

**Lemma 16** *For all  $d \geq 1$  and all  $\varphi \in \Phi^{n_1, n_2, \dots, n_d}$ ,*

$$N(\varphi) = \frac{\prod_{j=1}^d n_j!}{\prod_{\mu_1=0}^{n_1} \prod_{\mu_2=0}^{n_2} \cdots \prod_{\mu_d=0}^{n_d} (\mu_1! \mu_2! \cdots \mu_d!)^{\varphi_{\mu_1, \mu_2, \dots, \mu_d}} \varphi_{\mu_1, \mu_2, \dots, \mu_d}!}.$$

**Proof** We show the lemma for  $d = 2$ , and the proof is similar for any  $d \geq 1$ . Let  $\varphi \in \Phi^{n_1, n_2}$ . Any joint pattern  $(\bar{\psi}_1, \bar{\psi}_2)$  that has profile  $\varphi$  is a pair of sequences with symbols from  $\{1, 2, \dots, m\}$ , where  $m = \sum_{\mu_1=0}^{n_1} \sum_{\mu_2=0}^{n_2} \varphi_{\mu_1, \mu_2}$  is the total number of symbols in  $\bar{\psi}_1 \bar{\psi}_2$ . Let  $\{\mu_1(i)\}_{i=1}^m$  and  $\{\mu_2(i)\}_{i=1}^m$  be non-negative integers such that  $\sum_{i=1}^m \mu_1(i) = n_1$  and  $\sum_{i=1}^m \mu_2(i) = n_2$ . The number of sequence pairs whose alphabet is  $\{1, 2, \dots, m\}$ , and the number of appearances of  $i$  in first sequence is  $\mu_1(i)$  and in second sequence is  $\mu_2(i)$ , for  $i = 1, 2, \dots, m$ , is

$$\binom{n_1}{\mu_1(1), \mu_1(2), \dots, \mu_1(m)} \binom{n_2}{\mu_2(1), \mu_2(2), \dots, \mu_2(m)} = \frac{n_1! n_2!}{\prod_{i=1}^m \mu_1(i)! \mu_2(i)!}.$$

The number of different ways of choosing  $\{\mu_1(i)\}_{i=1}^m$  and  $\{\mu_2(i)\}_{i=1}^m$  such it conforms to profile is  $\varphi$  is

$$\binom{m}{\varphi_{0,0}, \varphi_{0,1}, \dots, \varphi_{n_1, n_2}} = \frac{m!}{\prod_{\mu_1=0}^{n_1} \prod_{\mu_2=0}^{n_2} \varphi_{\mu_1, \mu_2}!}.$$

Thus, the number of sequence pairs whose alphabet is  $\{1, 2, \dots, m\}$  and profile is  $\varphi$  is

$$N^*(\varphi) = \frac{n_1! n_2!}{\prod_{i=1}^m \mu_1(i)! \mu_2(i)!} \frac{m!}{\prod_{\mu_1=0}^{n_1} \prod_{\mu_2=0}^{n_2} \varphi_{\mu_1, \mu_2}!} = \frac{n_1! n_2! m!}{\prod_{\mu_1=0}^{n_1} \prod_{\mu_2=0}^{n_2} (\mu_1! \mu_2!)^{\varphi_{\mu_1, \mu_2}} \varphi_{\mu_1, \mu_2}!}.$$

Clearly,  $N^*(\varphi) = m! \cdot N(\varphi)$ , since

- $\geq$ : For each joint pattern having profile  $\varphi$ , the labels  $\{1, 2, \dots, m\}$  can be permuted in  $m!$  ways to generate  $m!$  different sequence pairs whose alphabet is  $\{1, 2, \dots, m\}$  and profile is  $\varphi$ . Furthermore, the sets of sequence pairs generated in this way by different joint patterns are disjoint. So  $N^*(\varphi) \geq m! \cdot N(\varphi)$ .
- $\leq$ : Given any pair of sequences  $(\bar{x}_1, \bar{x}_2)$  having alphabet  $\{1, 2, \dots, m\}$  and profile  $\varphi$ , their symbols can be permuted keeping the positions same to obtain a joint pattern with profile  $\varphi$ , which is in fact  $\Psi(\bar{x}_1, \bar{x}_2)$ . There are exactly  $m!$  sequence pairs having alphabet  $\{1, 2, \dots, m\}$  and profile  $\varphi$  that have the same joint pattern. Hence,  $N^*(\varphi) \leq m! \cdot N(\varphi)$ .

Thus,

$$N(\varphi) = \frac{N^*(\varphi)}{m!} = \frac{n_1!n_2!}{\prod_{\mu_1=0}^{n_1} \prod_{\mu_2=0}^{n_2} (\mu_1!\mu_2!)^{\varphi_{\mu_1,\mu_2}} \varphi_{\mu_1,\mu_2}!}.$$

□

## Appendix C. Symmetric tests

We provide a formal treatment to the intuition that joint patterns of sequences contain sufficient information for the problem of closeness testing, similar to [4].

We define the *symmetric error probability* of a test  $\Delta$  for  $(p_1, p_2)$  as its worst case error probability over all possible permutations of the alphabet, *i.e.*,

$$P_{e,\text{sym}}^n(\Delta, p_1, p_2) \stackrel{\text{def}}{=} \max_{\sigma \in S_{\mathcal{A}}} P_e^n(\Delta, p_1^\sigma, p_2^\sigma),$$

where  $S_{\mathcal{A}}$  is the set of all permutations of  $\mathcal{A}$ . Clearly, since separation between distributions does not depend on the actual symbols, and depends only the probability multiset, it is appropriate to look at the symmetric error probability.

A *symmetric* test is a test whose output does not change when the alphabet is permuted and gives the same output for all sequence pairs which have the same joint pattern, *i.e.*,  $\Delta(\bar{x}_1, \bar{x}_2) = \tilde{\Delta}(\Psi(\bar{x}_1, \bar{x}_2))$  for all  $(\bar{x}_1, \bar{x}_2)$ , where  $\tilde{\Delta} : \Psi^{n,n} \rightarrow \{\text{same}, \text{diff}\}$ . Hence, a symmetric test depends only the joint pattern of the sequences. Note that for a symmetric test  $\Delta$ ,  $P_{e,\text{sym}}(\Delta, p_1, p_2) = P_e(\Delta, p_1, p_2)$  for all distribution pairs  $(p_1, p_2)$ . The following observation shows that we may limit ourselves to considering only symmetric closeness tests.

**Observation 17** *Let  $\Delta : \mathcal{A}^n \times \mathcal{A}^n \rightarrow \{\text{same}, \text{diff}\}$  be any test for closeness, possibly not symmetric. Then, there exists a symmetric test  $\tilde{\Delta} : \mathcal{A}^n \times \mathcal{A}^n \rightarrow \{\text{same}, \text{diff}\}$  such that for all pairs of distributions  $(p_1, p_2)$  over  $\mathcal{A}$ ,  $P_{e,\text{sym}}^n(\tilde{\Delta}, p_1, p_2) \leq 2 \cdot P_{e,\text{sym}}^n(\Delta, p_1, p_2)$ .*

**Proof** Let  $\tilde{\Delta}$  be the test whose output for a sequence pair is same as that made by  $\Delta$  for the majority of sequence pairs with the same joint pattern, *i.e.*,  $\tilde{\Delta}(\bar{x}_1, \bar{x}_2) =$

majority $\{\Delta(\bar{x}'_1, \bar{x}'_2) : \Psi(\bar{x}'_1, \bar{x}'_2) = \Psi(\bar{x}_1, \bar{x}_2)\}$ . Clearly,  $P_e^n(\tilde{\Delta}, p_{1,2}^\sigma)$  is same for all permutations  $\sigma$  of  $\mathcal{A}$ . Thus, if  $p_1, p_2$  are similar,

$$\begin{aligned}
 P_{e,\text{sym}}^n(\tilde{\Delta}, p_1, p_2) &= P_e^n(\tilde{\Delta}, p_1, p_2) \\
 &= \frac{1}{|\mathcal{A}|!} \sum_{\sigma \in S_{\mathcal{A}}} P_e^n(\tilde{\Delta}, p_{1,2}^\sigma) \\
 &= \frac{1}{|\mathcal{A}|!} \sum_{\sigma \in S_{\mathcal{A}}} \sum_{\substack{(\bar{x}_1, \bar{x}_2): \\ \tilde{\Delta}(\bar{x}_1, \bar{x}_2) = \text{diff}}} p_1^\sigma(\bar{x}_1) p_2^\sigma(\bar{x}_2) \\
 &= \sum_{\substack{(\bar{x}_1, \bar{x}_2): \\ \tilde{\Delta}(\bar{x}_1, \bar{x}_2) = \text{diff}}} \frac{1}{|\mathcal{A}|!} \sum_{\sigma \in S_{\mathcal{A}}} p_1^\sigma(\bar{x}_1) p_2^\sigma(\bar{x}_2) \\
 &\stackrel{(a)}{\leq} 2 \sum_{\substack{(\bar{x}_1, \bar{x}_2): \\ \tilde{\Delta}(\bar{x}_1, \bar{x}_2) = \text{diff}}} \frac{1}{|\mathcal{A}|!} \sum_{\sigma \in S_{\mathcal{A}}} p_1^\sigma(\bar{x}_1) p_2^\sigma(\bar{x}_2) \\
 &= 2 \cdot \frac{1}{|\mathcal{A}|!} \sum_{\sigma \in S_{\mathcal{A}}} \sum_{\substack{(\bar{x}_1, \bar{x}_2): \\ \tilde{\Delta}(\bar{x}_1, \bar{x}_2) = \text{diff}}} p_1^\sigma(\bar{x}_1) p_2^\sigma(\bar{x}_2) \\
 &\leq 2 \cdot \max_{\sigma \in S_{\mathcal{A}}} \sum_{\substack{(\bar{x}_1, \bar{x}_2): \\ \tilde{\Delta}(\bar{x}_1, \bar{x}_2) = \text{diff}}} p_1^\sigma(\bar{x}_1) p_2^\sigma(\bar{x}_2) \\
 &= 2 \cdot P_{e,\text{sym}}^n(\Delta, p_1, p_2),
 \end{aligned}$$

where in (a), we note that all  $(\bar{x}_1, \bar{x}_2)$  having the same joint pattern have the same probability  $\frac{1}{|\mathcal{A}|!} \sum_{\sigma \in S_{\mathcal{A}}} p_1^\sigma(\bar{x}_1) p_2^\sigma(\bar{x}_2)$ . A similar argument can be shown for the case  $p_1 \neq p_2$ .  $\square$

# Oracle inequalities for computationally budgeted model selection

**Alekh Agarwal**

*University of California, Berkeley*

ALEKH@CS.BERKELEY.EDU

**John C. Duchi**

*University of California, Berkeley*

JDUCHI@CS.BERKELEY.EDU

**Peter L. Bartlett**

*University of California, Berkeley and Queensland University of Technology*

BARTLETT@CS.BERKELEY.EDU

**Clement Levraud**

*École Normale Supérieure*

CLEMENT@ENS.FR

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We analyze general model selection procedures using penalized empirical loss minimization under computational constraints. While classical model selection approaches do not consider computational aspects of performing model selection, we argue that any practical model selection procedure must not only trade off estimation and approximation error, but also the effects of the computational effort required to compute empirical minimizers for different function classes. We provide a framework for analyzing such problems, and we give algorithms for model selection under a computational budget. These algorithms satisfy oracle inequalities that show that the risk of the selected model is not much worse than if we had devoted all of our computational budget to the best function class.

**Keywords:** Model selection, oracle inequalities, computational budget

## 1. Introduction

In the standard statistical prediction setting, one receives samples  $\{z_1, \dots, z_n\} \subseteq \mathcal{Z}$  drawn i.i.d. from some unknown distribution  $P$  over a sample space  $\mathcal{Z}$ , and given a loss function  $\ell$ , seeks a function  $f$  to minimize the risk

$$R(f) := \mathbb{E}[\ell(z, f)]. \quad (1)$$

Since  $R(f)$  is unknown, the typical approach is to (approximately) minimize the empirical risk,  $\hat{R}_n(f) := \frac{1}{n} \sum_{i=1}^n \ell(z_i, f)$ , over a function class  $\mathcal{F}$ . We seek a function  $f_n$  with a risk close to the Bayes risk, the minimal risk over all measurable functions, which is  $R_0 := \inf_f R(f)$ . There is a natural tradeoff based on the class  $\mathcal{F}$  one chooses, since

$$R(f_n) - R_0 = \left( R(f_n) - \inf_{f \in \mathcal{F}} R(f) \right) + \left( \inf_{f \in \mathcal{F}} R(f) - R_0 \right),$$

which decomposes the excess risk of  $f_n$  into estimation error (left) and approximation error (right).

A common approach to addressing this tradeoff is to express  $\mathcal{F}$  as a union of classes  $\mathcal{F}_1, \dots, \mathcal{F}_k$ . The *model selection problem* is to choose a class  $\mathcal{F}_i$  and a function  $f \in \mathcal{F}_i$  to give the best tradeoff between estimation error and approximation error.<sup>1</sup> A common approach to the model selection problem is the now classical idea of *complexity regularization*, which arose out of early works by Mallows (1973) and Akaike (1974). The complexity regularization approach balances two competing objectives: the minimum empirical risk of a model class  $\mathcal{F}_i$  (approximation error) and a complexity penalty (to control estimation error) for the class. Different choices of the complexity penalty give rise to different model selection criteria and algorithms (see e.g. Massart, 2003, and the references therein). Results of several authors (e.g. Bartlett et al., 2002; Lugosi and Wegkamp, 2004; Massart, 2003) show that given a dataset of size  $n$ , the output  $\hat{f}_n$  of the procedure roughly satisfies

$$\mathbb{E}R(\hat{f}_n) - R_0 \leq \min_i \left[ \inf_{f \in \mathcal{F}_i} R(f) - R_0 + \gamma_i(n) \right] + \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right), \quad (2)$$

where  $\gamma_i(n)$  is a complexity penalty for class  $i$ , which is usually decreasing to zero in  $n$  and increasing in  $i$ . (Several approaches to complexity regularization are possible, and an incomplete bibliography includes Vapnik and Chervonenkis, 1974; Geman and Hwang, 1982; Rissanen, 1983; Barron, 1991; Bartlett et al., 2002; Lugosi and Wegkamp, 2004).

These oracle inequalities show that, for a given sample size, the model selection procedure gives the best trade-off between the approximation and estimation errors. A drawback with the above mentioned approaches is that we need to be able to optimize over each model in the hierarchy on the entire data, in order to prove guarantees on the result of the model selection procedure. This is natural when the sample size is the key limitation, and it is computationally feasible when the sample size is small and the samples are low-dimensional. However, the cost of training  $K$  different model classes on the entire data sequence can be prohibitive when the datasets become large and high-dimensional as is common in modern settings. In these cases, it is computational resources—rather than the sample size—that are the key constraint. In this paper, we consider model selection from this computational perspective, viewing the amount of computation, rather than the sample size, as the parameter which will enter our oracle inequalities. Specifically, we consider model selection methods that work within a given computational budget.

An interesting and difficult aspect of the problem that we must address is the interaction between model class complexity and computation time. It is natural to assume that for a fixed sample size, it is more expensive to estimate a model from a complex class than a simple class. Put inversely, given a computational bound, a simple model class can fit a model to a much larger sample size than a rich model class. So any strategy for model selection under a computational budget constraint should trade off two criteria: (i) the relative training cost of different model classes, which allows simpler classes to receive far more data (thus making them resilient to overfitting), and (ii) lower approximation error in the more complex model classes.

In addressing these computational and statistical issues, this paper makes three main contributions. First, we propose a novel computational perspective on the model selection problem, which we believe should be a natural consideration in statistical learning problems.

---

1. In general, the number of classes  $K$  can be infinite, though we restrict attention to finitely many classes for this paper.

Secondly, within this framework, we provide an algorithm—exploiting algorithms for multi-armed bandit problems—that uses confidence bounds based on concentration inequalities to select a good model under a given computational budget. We also prove a minimax optimal oracle inequality on the performance of the selected model. Our third main contribution is another algorithm based on a coarse-grid search, for model hierarchies that are structured by inclusion, that is,  $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subseteq \mathcal{F}_K$ . Under natural assumptions regarding the growth of the complexity penalties as we go to more complex classes, the coarse-grid search procedure satisfies better oracle inequalities than the earlier bandit algorithm. Both of our algorithms are computationally simple and efficient.

The remainder of this paper is organized as follows. In the next section, we formalize our setting and present the algorithms. Section 3 presents our main results as well as some consequences for specific problems and examples. We provide proofs in Sections 4 through 5; Section 4 contains the proof of the result for unstructured model selection problems, while Sec. 5 contains the proofs of oracle inequalities for model selection problems with nested classes  $\mathcal{F}_i$ .

## 2. Setup and algorithms

In this section, we will describe our statistical and computational assumptions about the problem, giving examples of classes of problems and statistical procedures that satisfy the assumptions. We will follow this with descriptions of our algorithms, including intuitive explanations of the procedures.

### 2.1. Setup and Goals

Recall from the introduction that we have a collection of  $K$  model classes  $\mathcal{F}_1, \dots, \mathcal{F}_K$ . Let us begin by describing our computational assumptions. First, we assume as our basic unit of measure a computational quantum; within this quantum, a model can be trained on any single class  $\mathcal{F}_i$  using  $n_i$  samples. That is, we associate with each class  $\mathcal{F}_i$  a number of samples  $n_i \in \mathbb{N}$ , where  $n_i$  is chosen so that training a model from class  $\mathcal{F}_i$  on  $n_i$  examples requires the same amount of time as training a model from class  $\mathcal{F}_j$  on  $n_j$  samples. We assume an overall time budget of  $T$  quanta, so that if we devote the entire computational budget to class  $i$ , we could use  $Tn_i$  samples to train a model.<sup>2</sup> Our high level goal is to derive algorithms that perform nearly as well as if an oracle gave the best model class  $i^*$  in advance, and we could devote the entire computational budget  $T$  to class  $i^*$ .

For the statistical assumptions in our problem, we take an approach similar to that of Bartlett et al. (2002), restricting our attention to complexity penalties based on concentration inequalities. Each of our model selection procedures uses a black-box algorithm  $\mathcal{A}$  for fitting functions from the model class  $\mathcal{F}_i$  to the data. We require that these algorithms be statistically well-behaved, in the sense that the empirical risk of  $\mathcal{A}$ 's output model  $\hat{f}$  is near the true risk of  $\hat{f}$ . Recalling the definitions of  $R, \widehat{R}$  from the introduction, and defining  $[K] = \{1, \dots, K\}$ , we state our main concentration assumption:

---

2. The linearity assumption is essentially no loss of generality. In addition, several algorithms satisfy it. We can work with general non-linear scalings too, at the cost of significant notational burden which we choose to avoid here.

**Assumption A** Let  $\mathcal{A}(i, n) \in \mathcal{F}_i$  denote the output of algorithm  $\mathcal{A}$  on a sample of  $n$  data points.

- (a) For each  $i \in [K]$ , there is a function  $\gamma_i$  and constants  $\kappa_1, \kappa_2 > 0$  such that for any  $n \in \mathbb{N}$ ,

$$\mathbb{P}\left(|\widehat{R}_n(\mathcal{A}(i, n)) - R(\mathcal{A}(i, n))| > \gamma_i(n) + \kappa_2\epsilon\right) \leq \kappa_1 \exp(-4n\epsilon^2). \quad (3)$$

- (b) The output  $\mathcal{A}(i, n)$  is a  $\gamma_i(n)$ -minimizer of  $\widehat{R}_n$ , that is,

$$\widehat{R}_n(\mathcal{A}(i, n)) - \inf_{f \in \mathcal{F}_i} \widehat{R}_n(f) \leq \gamma_i(n).$$

- (c) The function  $\gamma_i(n) \leq c_i n^{-\alpha_i}$  for some  $\alpha_i > 0$ .

- (d) For any fixed function  $f \in \mathcal{F}_i$ ,  $\mathbb{P}(|\widehat{R}_n(f) - R(f)| > \kappa_2\epsilon) \leq \kappa_1 \exp(-4n\epsilon^2)$ .

There are many classes of functions and corresponding algorithms that satisfy Assumption A. For one simple example, let  $\{\mathcal{F}_i\}$  be VC-classes of functions, where each  $\mathcal{F}_i$  has VC-dimension  $d_i$ , and  $\ell$  be the hinge loss, where  $\ell(z, f) = [1 - yf(x)]_+$ . Assuming that  $\ell(z, f) \leq B$  for all  $f \in \mathcal{F}$ , Dudley's entropy integral in this case gives (Dudley, 1978)

$$\gamma_i(n) = \mathcal{O}\left(\sqrt{\frac{d_i}{n}}\right) \quad \text{and} \quad \kappa_1 \leq 2, \quad \kappa_2 = \mathcal{O}(B). \quad (4)$$

Similar results hold for other convex losses and problems, for example regression and density estimation problems with squared or log losses. For function classes of bounded complexity, such as VC, Sobolev, or Besov classes, penalty functions  $\gamma_i(n)$  can be computed that satisfy Assumption A using many techniques; some relevant approaches include Rademacher and Gaussian complexities of the function classes  $\mathcal{F}_i$ , metric entropy, Dudley's entropy integral, or localization techniques (e.g. Pollard, 1984; Bartlett and Mendelson, 2002; Dudley, 1967). In many concrete cases, such as parametric models or VC classes, Assumption A(c) is satisfied with  $\alpha_i = \frac{1}{2}$ .

Our approach, similar to the idea of complexity regularization, is to perform a kind of penalized model selection. If we knew the true risk functional  $R$ , we could minimize a combination of the risk and complexity penalty based on the number of samples our computational budget allows for the class. In particular, given penalty functions  $\gamma_i$ , we define the best class in hindsight as

$$i^* := \operatorname{argmin}_{i \in [K]} \left\{ \inf_{f \in \mathcal{F}_i} R(f) + \gamma_i(Tn_i) \right\}. \quad (5)$$

The idea is that an algorithm performing model selection—while taking into account its computational limitations—should choose the best class considering the total number of samples it could possibly have seen for the class. We note that this is also closely related to the criterion (2) minimized in the absence of a computational budget, but in the classical case it is assumed that each function class can be evaluated on an identical and fixed number of samples  $n$ .

```

FOREACH  $i \in [K]$  query  $n_i$  examples from class  $\mathcal{F}_i$ 
FOR  $t = K + 1$  to  $T$ 
    SET  $n_i(t)$  to be the number of examples seen for class  $i$  at time  $t$ 
    LET  $i_t = \operatorname{argmin}_{i \in [K]} \bar{R}(j, n_i(t)) - \sqrt{\frac{\log T}{n_i(t)}}$ 
    QUERY  $n_{i_t}$  examples for class  $i_t$ 
OUTPUT  $\hat{i}$ , the index of the most frequently selected class.

```

**Algorithm 1:** Multi-armed bandit algorithm for selection of best class  $\hat{i}$ .

## 2.2. Upper-confidence bound algorithm without structure

We now turn to outlining the first of the two main scenarios analyzed in this paper. For now, we do not assume any structure relating the collection of model classes  $\mathcal{F}_1, \dots, \mathcal{F}_K$ . The main idea of our algorithm in this case is to incrementally allocate our computational quota amongst the function classes, where we trade off receiving samples for classes that have good risk performance against exploring classes for which we have received few data points. We view the budgeted model selection problem as a repeated game with  $T$  rounds. At iteration  $t$ , the procedure allocates one additional quantum of computation to a (to be specified) function class  $i$ . We assume that the computational complexity of fitting a model grows linearly and incrementally with the number of samples, which means that allocating an additional quantum of training time allows the black-box training algorithm  $\mathcal{A}$  to process an additional  $n_i$  samples for class  $\mathcal{F}_i$ . The linear growth assumption is satisfied, for instance, when the loss function  $\ell$  is convex and the black-box learning algorithm  $\mathcal{A}$  is a stochastic or online convex optimization procedure (e.g. Cesa-Bianchi and Lugosi, 2006; Nemirovski et al., 2009).

Using our previously defined notation, we now define the criterion we use in our procedure to select the class  $i$  to which we allocate a quantum. The optimistic selection criterion for class  $i$ , assuming that  $\mathcal{F}_i$  has seen  $n$  samples at this point in the game, is

$$\bar{R}(i, n) = \hat{R}_n(\mathcal{A}(i, n)) - \gamma_i(n) - \sqrt{\frac{\log K}{n}} + \gamma_i(Tn_i). \quad (6)$$

The intuition behind the definition of  $\bar{R}(i, n)$  is that we would like the algorithm to choose functions  $f$  and classes  $i$  that minimize  $\hat{R}_n(f) + \gamma_i(Tn_i) \approx R(f) + \gamma_i(Tn_i)$ , but the negative  $\gamma_i(n)$  and  $\sqrt{\log K/n}$  terms lower the criterion significantly when  $n$  is small and thus encourage initial exploration. The criterion (6) essentially combines the penalized model-selection objective used by Bartlett et al. (2002) (though we use a  $\log K$  term, as we assume a finite number of classes) with an optimistic criterion similar to those used in multi-armed bandit algorithms (Auer et al., 2002). Algorithm 1 contains our bandit procedure for model selection. We run Alg. 1 for  $T$  rounds, where  $T$  is such that the entire computational budget is exhausted. Our results in Section 3.1 show that Alg. 1 satisfies our twofold goals of selecting the class  $i^*$  with high probability and outputting a function  $f$  with good risk performance.

### 2.3. Coarse-grid search algorithm for inclusion hierarchy

In practice, the classes  $\mathcal{F}_i$  are rarely completely unrelated; perhaps the most common scenario in model selection is structural risk minimization, where the model classes  $\mathcal{F}_i$  are subsets ordered in increasing complexity of a larger model space. To that end, our second main scenario involves studying computationally constrained model selection procedures under the following assumption.

**Assumption B** *The function classes  $\mathcal{F}_i$  satisfy an inclusion hierarchy:*

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_K \quad (7)$$

One simple example satisfying above assumption is classes of functions of the form  $x \mapsto f(x) = \langle \theta, x \rangle$ , where each function class  $\mathcal{F}$  is identified with an increasing bound on  $\|\theta\|$ . A second simple family of examples consists of scenarios in which  $f \in \mathcal{F}_i$  is of the form  $x \mapsto f(x) = \langle \theta, \phi_i(x) \rangle$  where  $\phi_i$  is a feature mapping of the input data  $\mathcal{Z}$  and  $\phi_i$  is a projection of  $\phi_{i+1}$ . For example, functions in class  $i + 1$  observe more features than those in class  $i$  or the different classes  $\mathcal{F}_i$  may consist of an increasing sequence of wavelet bases.

Intuitively, we expect the structure assumed above to help our model selection procedure because the minimum expected risks of different function classes are no longer independent of each other. It is easy to see that under our assumption,

$$R_i^* \leq R_j^* \quad \text{for } i \geq j. \quad (8)$$

Clearly, under Assumption B, the penalties can always be chosen to be increasing as a function of the class complexity:

$$\gamma_i(n) \geq \gamma_j(n) \quad \text{for } i \geq j. \quad (9)$$

Since our approach involves giving a different number of samples to each class, we require a slightly stronger ordering than the above equation. We assume that for any budget  $T$ , we have

$$\gamma_i(Tn_i) \geq \gamma_j(Tn_j) \quad \text{for } i \geq j. \quad (10)$$

This assumption is reasonable since we expect that  $\gamma_i(n)$  is a decreasing function of  $n$  and  $n_i \leq n_j$  for  $i \geq j$ , so that  $\gamma_i(Tn_i) \geq \gamma_i(Tn_j) \geq \gamma_j(Tn_j)$ .

We now show a simple grid-search based algorithm that gives oracle inequalities depending only logarithmically on the number of classes for this inclusion hierarchy under natural conditions on the growth of the complexity penalties as a function of the class index  $i$ . The method takes inspiration from the naïve strategy that splits the budget  $T$  uniformly across the  $K$  classes and finds the class with the smallest penalized empirical risk, using  $Tn_i/K$  samples for class  $i$ . Of course, the naïve approach has the drawback that the computational budget available to each class is reduced by a factor of  $K$ , which yields very poor scaling with the number  $K$  of classes.

The key observation we exploit is that under the nesting structure (7), we do not need to find the smallest regularized empirical risk for each class. We can instead pick a small subset  $S$  of classes and perform model selection only over the classes in  $S$ , then use the inclusion assumption B to reason about the classes not in  $S$  for appropriate choices of  $S$ . With this intuition, we now define a good choice for  $S$ :

**Definition 1 (Coarse grid)** For a set  $S \subseteq [K]$ , we say that  $S$  satisfies the coarse grid conditions with parameters  $s \in \mathbb{N}$  and  $\lambda > 0$  if  $|S| = s$  and for each  $i \in [K]$  there is an index  $j \in S$  such that

$$\gamma_i\left(\frac{Tn_i}{s}\right) \leq \gamma_j\left(\frac{Tn_j}{s}\right) \leq (1 + \lambda)\gamma_i\left(\frac{Tn_i}{s}\right). \quad (11)$$

We define  $s(\lambda)$  to be the size of the smallest set  $S$  satisfying condition (11), noting that  $s(\lambda) \leq K$ . In general, for a given  $\lambda$  there may be no small set  $S$  satisfying Definition 1; however, we are interested in settings where a set  $S$  of size  $s(\lambda) = \mathcal{O}(\log K)$  exists.

**Example 1** Let  $\{\mathcal{F}_i\}$  be an increasing collection of VC-classes, say  $f \in \mathcal{F}_i$  is of the form  $x \mapsto f(x) = \langle \theta, \phi_i(x) \rangle$  where  $\phi_i$  is a  $d_i$ -dimensional mapping and  $\mathcal{F}_i$  has VC-dimension  $d_i$ . In this case, recalling the VC-bound (4), we know that (up to constant factors)  $\gamma_i(n) \leq \sqrt{d_i/n}$ . Making the reasonable assumption that training time is linearly dependent on the VC dimension, we have  $n_i = n_K(d_K/d_i)$  for  $i \in [K]$ , so

$$\gamma_i(Tn_i) = \gamma_i\left(\frac{Tn_K d_K}{d_i}\right) \leq \sqrt{\frac{d_i^2}{Tn_K d_K}} = d_i \cdot \frac{1}{\sqrt{Tn_K d_K}}.$$

Example 1 is suggestive of a pattern common to many hierarchies of function classes—including parametric and VC-classes with  $i$  indexing VC-dimension—where the penalty functions interact with the sample sizes  $n_i$  so that  $\gamma_i$  splits naturally into a product  $\gamma_i(Tn_i) = g(T)h(i)$  for some functions  $g$  and  $h$  (which may depend on  $K$ ). For such cases, the condition (11) reduces to ensuring

$$g\left(\frac{T}{s(\lambda)}\right)h(i) \leq g\left(\frac{T}{s(\lambda)}\right)h(j) \leq (1 + \lambda)g\left(\frac{T}{s(\lambda)}\right)h(i),$$

which amounts to showing  $h(i) \leq h(j) \leq (1 + \lambda)h(i)$  independent of the setting of  $s(\lambda)$  (since  $h$  is non-increasing, we need only show the latter inequality). Let  $S = \{j_1, \dots, j_{s(\lambda)}\}$ . We construct  $S$  by setting  $j_{s(\lambda)} = K$  and recursively defining  $j_i$  to be the smallest index  $j_i < j_{i+1}$  such that

$$h(j_{i+1}) \leq (1 + \lambda)h(j_i).$$

Then the number of classes can be bounded by using the relation

$$h(K) = h(j_{s(\lambda)}) \leq (1 + \lambda)h(j_{s(\lambda)-1}) \leq \dots \leq (1 + \lambda)^{s(\lambda)}h(1),$$

so that so long as  $s(\lambda) \geq \frac{\log(h(K)/h(1))}{\log(1+\lambda)}$ , we can choose a set  $S$  satisfying condition (11) with  $|S| = s(\lambda)$ . In particular,  $s(\lambda)$  is logarithmic in  $K$  as long as the function  $h$  grows sub-exponentially. Other natural examples of function classes satisfying such growth conditions include Besov or Sobolev function classes nested by degree or smoothness as well as wavelet bases. We refer the reader to the work of Barron et al. (1999) for a compendium of results where  $\gamma_i(Tn_i) = g(T)h(i)$ .

Given the above, our algorithm has a simple description. We fix a desired accuracy  $\lambda$  and find the smallest set  $S$  satisfying Definition 1. We then pick the class  $\hat{i}$  satisfying

$$\hat{i} \in \operatorname{argmin}_{i \in S} \left\{ \hat{R}_{Tn_i/s(\lambda)}(i) + \gamma_i(Tn_i/s(\lambda)) \right\}, \quad (12)$$

where  $|S| = s(\lambda)$ . We observe that the penalty functions are typically known in closed form (with the exception of data-dependent complexity penalties), and hence computation of the set  $S$  can be efficient and is (generally) much cheaper than training the models. In Section 3.2, we give an oracle inequality on the performance of the estimate  $\hat{i}$  from the procedure (12) that has only mild dependence on the number of classes so long as  $s(\lambda)$  does not grow too fast with  $K$ .

### 3. Main results and their consequences

In this section, we come to the description of the performance guarantees for Algorithms 1 and (12). To build intuition, we also specialize the theorems to specific statistical problems and model classes.

#### 3.1. Oracle inequalities for unstructured model classes

In this section we give performance guarantees on the class picked by Algorithm 1. We define the excess penalized risk

$$\Delta_i := R_i^* + \gamma_i(Tn_i) - R_{i^*}^* - \gamma_{i^*}(Tn_{i^*}) \geq 0. \quad (13)$$

Essentially without loss of generality, we assume that the infimum in the equation  $R_i^* = \inf_{f \in \mathcal{F}_i} R(f)$  is attained by a function  $f_i^*$ . If the infimum is not attained we simply choose some fixed  $f_i^*$  such that  $R(f_i^*) \leq \inf_{f \in \mathcal{F}_i} R(f) + \delta$  for an arbitrarily small  $\delta > 0$ . We first perform analysis under the assumption that  $\Delta_i > 0$  strictly for  $i \neq i^*$ , but we will then relax to allow non-unique  $i^*$ .

The gains of a computationally adaptive strategy over naïve strategies are best seen when the gap (13) is non-zero. Under this assumption, we can follow the ideas of Auer et al. (2002) and show that the fraction of the computational budget allocated to any suboptimal class  $i \neq i^*$  goes quickly to zero as  $T$  grows. We provide the proof of the following theorem in Section 4.

**Theorem 2** *Let Alg. 1 be run for  $T$  rounds and  $T_i(T)$  be the number of times class  $i$  is queried. Let  $\Delta_i$  be defined as in (13), the conditions of Assumption A hold, and assume that  $T \geq K$ . Define  $\beta_i = \max\{1/\alpha_i, 2\}$ . There is a constant  $C$  such that*

$$\mathbb{E}[T_i(T)] \leq \frac{C}{n_i} \left( \frac{c_i + \kappa_2 \sqrt{\log T}}{\Delta_i} \right)^{\beta_i} \quad \text{and} \quad \mathbb{P}\left(T_i(T) > \frac{C}{n_i} \left( \frac{c_i + \kappa_2 \sqrt{\log T}}{\Delta_i} \right)^{\beta_i}\right) \leq \frac{\kappa_1}{TK^4},$$

where  $c_i$  and  $\alpha_i$  come from the definition of the concentration function  $\gamma_i$  in Assumption A(c).

At a high level, this result shows that the fraction of budget allocated to any suboptimal class goes to 0 at the rate  $\frac{1}{n_i T} \left( \frac{\sqrt{\log T}}{\Delta_i} \right)^{\beta_i}$ . Hence, asymptotically in  $T$ , we will receive exponentially more samples for  $i^*$  than any other class and will perform almost as well as if we had known  $i^*$  in advance. To see an example of concrete rates that can be concluded from the above result, let  $\mathcal{F}_1, \dots, \mathcal{F}_K$  be model classes with finite VC-dimension,<sup>3</sup> so that Assumption A is satisfied with  $\alpha_i = \frac{1}{2}$ . Then we have

---

3. Similar corollaries hold for any model class whose metric entropy grows as  $\text{polylog}(\frac{1}{\epsilon})$ .

**Corollary 3** Under the conditions of Theorem 2, assume  $\mathcal{F}_1, \dots, \mathcal{F}_K$  are model classes of finite VC-dimension, where  $\mathcal{F}_i$  has dimension  $d_i$ . Then there is a constant  $C$  such that

$$\mathbb{E}[T_i(T)] \leq C \frac{\max\{d_i, \kappa_2^2 \log T\}}{\Delta_i^2 n_i} \quad \text{and} \quad \mathbb{P}\left(T_i(T) > C \frac{\max\{d_i, \kappa_2^2 \log T\}}{\Delta_i^2 n_i}\right) \leq \frac{\kappa_1}{TK^4}.$$

The result of Corollary 3 is nearly optimal in general due to a lower bound for the special case of multi-armed bandit problems (Lai and Robbins, 1985). To see the connection, let  $\mathcal{F}_i$  correspond to the  $i$ th arm in a multi-armed bandit problem and the risk  $R_i^*$  be the expected reward of arm  $i$ . In this case, the complexity penalty  $\gamma_i$  for each class is 0. Lai and Robbins give a lower bound that shows that the expected number of pulls of any suboptimal arm is at least  $\mathbb{E}[T_i(T)] = \Omega\left(\frac{\log T}{\text{KL}(p_i \| p_{i^*})}\right)$ , where  $p_i$  and  $p_{i^*}$  are the reward distributions for the  $i$ th and optimal arms, respectively.

Unfortunately, the condition that  $\Delta_i > 0$  may not always be satisfied, or  $\Delta_i$  may be so small as to render the bound in Theorem 2 vacuous. Nevertheless, we intuitively believe that our algorithm can quickly find a small set of “good” classes—those with small penalized risk—and spend its computational budget to try to distinguish amongst them. In this case, though, Algorithm 1 will not visit suboptimal classes and so can still output a function  $f$  satisfying good oracle bounds. In order to prove a result quantifying this intuition, we first upper bound the *regret* of Algorithm 1, that is, the average excess risk suffered by the algorithm over all iterations, and then show how to use this bound for obtaining a model with a small risk. We state our results for the case where  $\alpha_i \equiv \alpha$  and define  $\beta = \max\{1/\alpha, 2\}$ .

**Proposition 4** Use the same assumptions as Theorem 2, but further assume that  $\alpha_i \equiv \alpha$  for all  $i$ . With probability at least  $1 - \kappa_1/TK^3$ , the regret (average excess risk) of Algorithm 1 is bounded as

$$\sum_{i=1}^K \Delta_i T_i(T) \leq 2eT^{1-1/\beta} \left( C \sum_{i=1}^K \frac{(c_i + \kappa_2 \sqrt{\log T})^\beta}{n_i} \right)^{1/\beta}$$

for a constant  $C$  dependent on  $\alpha$ .

In order to obtain a model with a small risk, we need to make an additional assumption that the models are compatible in the sense that one can define the addition operator  $f + g$  for  $f \in \mathcal{F}_i, g \in \mathcal{F}_j$  meaningfully. We also assume that the risk functional  $R(f)$  is convex in  $f$ . In such a setting, we can average the functions minimizing the objective  $\bar{R}(i, n)$ , that is,  $f_t = \operatorname{argmin}_{f \in \mathcal{F}_i} \hat{R}_{n_i(t)}(f)$ , to obtain a function satisfying the desired oracle inequality. For this theorem, we also assume that the constants  $c_i$  from Assumption A(c) satisfy  $c_i = \mathcal{O}(\sqrt{\log T})$ .

**Theorem 5** Use the same assumptions as Proposition 4. Let  $f_t$  be the function chosen by algorithm  $\mathcal{A}$  at round  $t$  of Alg. 1 and define the average function  $\hat{f}_T = \frac{1}{T} \sum_{t=1}^T f_t$ . If the risk functional  $R$  is convex, there are constants  $C, C'$  (dependent on  $\alpha$ ) such that with

probability greater than  $1 - 2\kappa_2/(TK^3)$ ,

$$\begin{aligned} R(\hat{f}_T) &\leq R^* + \gamma_{i^*}(Tn_{i^*}) + 2e\kappa_2 T^{-\beta} \sqrt{\log T} \left( \sum_{i=1}^K \frac{C}{n_i} \right)^{1/\beta} \\ &\quad + C' T^{-1/\beta} \left( \sum_{i=1}^K \left[ c_i n_i^{-\alpha} + \kappa_2 n_i^{-\frac{1}{2}} \sqrt{\log K} + \kappa_2 n_i^{-\frac{1}{2}} \sqrt{\log T} \right]^\beta \right)^{1/\beta}. \end{aligned}$$

Let us interpret the above bound and discuss its optimality. When  $\alpha = \frac{1}{2}$  (e.g., for VC classes), we have  $\beta = 2$ ; moreover, it is clear that  $\sum_{i=1}^K \frac{C}{n_i} = \mathcal{O}(K)$ . Thus, to within constant factors, we have

$$R(\hat{f}_T) = R^* + \gamma_{i^*}(Tn_{i^*}) + \mathcal{O}\left(\frac{\sqrt{K} \max\{\log T, \log K\}}{\sqrt{T}}\right).$$

Ignoring logarithmic factors, the above bound is minimax optimal, which follows by a reduction of our model selection problem to the special case of a multi-armed bandit problem. In this case, Theorem 5.1 of Auer et al. (2003) shows that for any set of  $K, T$  values, there is a distribution over the rewards of arms which forces  $\Omega(\sqrt{KT})$  regret, that is, the average excess risk of the classes chosen by Alg. 1 must be  $\Omega(\sqrt{KT})$ . We provide proofs of Proposition 4 and Theorem 5 in the long version of the paper.

### 3.2. Oracle inequalities for nested hierarchies

In this section we provide an oracle inequality on the output of the procedure (12) that has a more favorable dependence on the number of classes  $K$  than our bounds for unstructured function classes  $\mathcal{F}_i$ . The main idea is to use Assumption B along with Definition 1 to show that performing a coarse grid search over  $S$  is sufficient to deduce an oracle inequality over the entire hierarchy. The next theorem provides an oracle inequality for the risk of the function  $f = \mathcal{A}(\hat{i}, \tilde{n}_{\hat{i}} T / s(\lambda))$ , which is the output of the learning algorithm  $\mathcal{A}$  applied to the class  $\hat{i}$  picked by our algorithm.

**Theorem 6** *Let  $f = \mathcal{A}(\hat{i}, \tilde{n}_{\hat{i}} T / s(\lambda))$  be the output of the algorithm  $\mathcal{A}$  for class  $\hat{i}$  specified by the procedure (12). Let Assumptions A–B hold. With probability at least  $1 - 3\kappa_1 \exp(-4m)$*

$$R(f) \leq \min_{i \in [K]} \left\{ R_i^* + 2(1 + \lambda)\gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) \right\} + \kappa_2 \sqrt{\frac{s(\lambda) \log K}{2Tn_K}} + \kappa_2 \sqrt{\frac{ms(\lambda)}{Tn_K}}.$$

**Remark:** It is possible to reduce the  $\kappa_2 \sqrt{s(\lambda) \log K / 2Tn_K}$  term in the bound above to a  $(1 + \lambda) \sqrt{s(\lambda) \log i / 2Tn_i}$  term appearing inside the minimum over classes  $i \in [K]$  by requiring the coarse grid condition (11) to hold over terms of the form  $\gamma_i(Tn_i / s(\lambda)) + \sqrt{s(\lambda) \log i / 2Tn_i}$ . This stronger bound applies, for example, to sequences of VC-classes as described in Example 1.

The above result makes it clear that the excess risk of the algorithm—outside of the minimum over all the classes—scales as  $\mathcal{O}(T^{-1/2})$ . It is of interest to contrast Theorem 6

with results of the previous section. In the completely general case, we have a dependence on  $K$  better than  $\sqrt{K}$  only when there is constant separation between the penalized risks of different classes. Since  $s(\lambda) \leq K$ , the result of the above theorem is essentially as strong as any of the results from the previous section, as we would hope when we know  $\mathcal{F}_i \subseteq \mathcal{F}_{i+1}$ .

Nonetheless, the main strength of Theorem 6 is in scenarios where  $s(\lambda) = \mathcal{O}(\log K)$ , such as VC-classes (e.g. Example 1) with at most polynomial growth in VC-dimension. In such scenarios, the function  $f$  that the procedure outputs is competitive (up to logarithmic factors) with an oracle that devotes the entire computation budget to the optimal class. We note that model selection procedures suffer a penalty of  $\sqrt{\log K}$  (or  $\sqrt{\log i}$ ) even in computationally unconstrained settings (see, e.g., Bartlett et al., 2002), so our computationally restricted procedure suffers at most an additional penalty of  $\mathcal{O}(\sqrt{\log K})$ . We conclude by recalling that many common model selection scenarios satisfy  $s(\lambda) = \mathcal{O}(\log K)$ , as noted in Section 2.3.

#### 4. Proof of Theorem 2

At a high level, the proof of this theorem involves combining the techniques for analysis of multi-armed bandits developed by Auer et al. (2002) with Assumption A. We start by giving a lemma which will be useful to prove the theorem. The lemma states that after a sufficient number of initial iterations  $\tau$ , the probability Algorithm 1 chooses to receive samples for a sub-optimal function class  $i \neq i^*$  is extremely small. Recall also our notational convention that  $\beta_i = \max\{1/\alpha_i, 2\}$ .

**Lemma 7** *For any class  $i$ , any  $s_i \in [1, T]$  and  $s_{i^*} \in [\tau, T]$  where  $\tau > 0$  satisfies*

$$\tau > \frac{2^{\beta_i}(c_i + \kappa_2\sqrt{\log T} + \kappa_2\sqrt{\log K})^{\beta_i}}{n_i\Delta_i^{\beta_i}},$$

*under Assumption A we have*

$$\mathbb{P}\left(\bar{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \leq \bar{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}}\right) \leq \frac{2\kappa_1}{(TK)^4}.$$

We defer the proof of the lemma to Appendix A, though at a high level the proof works as follows. The “bad event” in Lemma 7, that is, Algorithm 1 selects a sub-optimal class  $i \neq i^*$ , occurs only if one of the following three errors occurs: the empirical risk of class  $i$  is much lower than its true risk, the empirical risk of class  $i^*$  is higher than its true risk, or  $s_i$  is not large enough to actually separate the true penalized risks from one another. Under the assumptions of the lemma, however, coupled with the uniform convergence properties in Assumption A, each of these three sub-events is quite unlikely. Now we turn to the proof of Theorem 2 assuming the lemma.

Let  $i_t$  denote the model class index  $i$  chosen by Algorithm 1 at time  $t$ , and let  $s_i(t)$  denote the number of times class  $i$  has been selected at round  $t$  of the algorithm. When no time index is needed,  $s_i$  will denote the same thing. Note that if  $i_t = i$  and the number of

times class  $i$  is queried exceeds  $\tau > 0$ , then by the definition of the selection criterion (6) and choice of  $i_t$  in Alg. 1, for some  $s_i \in \{\tau, \dots, t-1\}$  and  $s_{i^*} \in \{1, \dots, t-1\}$  we have

$$\bar{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \leq \bar{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}}.$$

Here we interpret  $\bar{R}(i, n_i s_i)$  to mean a random realization of the observed risk consistent with the samples we observe. Using the above implication, we thus have

$$\begin{aligned} T_i(n) &= 1 + \sum_{t=K+1}^T \mathbb{I}(i_t = i) \leq \tau + \sum_{t=K+1}^T \mathbb{I}(i_t = i, T_i(t-1) \geq \tau) \\ &\leq \tau + \sum_{t=K+1}^T \mathbb{I}\left(\min_{\tau \leq s_i < t} \bar{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \leq \max_{0 < s < t} \bar{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}}\right) \\ &\leq \tau + \sum_{t=1}^T \sum_{s_{i^*}=1}^{t-1} \sum_{s_i=\tau}^{t-1} \mathbb{I}\left(\bar{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \leq \bar{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}}\right). \quad (14) \end{aligned}$$

To control the last term, we invoke Lemma 7 and obtain that

$$\tau > \frac{2^{\beta_i} (c_i + \kappa_2 \sqrt{\log T} + \kappa_2 \sqrt{\log K})^{\beta_i}}{n_i \Delta_i^{\beta_i}} \Rightarrow \mathbb{E}[T_i(n)] \leq \tau + \sum_{t=1}^T \sum_{s=1}^{t-1} \sum_{s_i=\tau}^{t-1} 2 \frac{\kappa_1}{(TK)^4} \leq \tau + \frac{\kappa_1}{TK^4}.$$

Hence for any suboptimal class  $i \neq i^*$ ,  $\mathbb{E}[T_i(n)] \leq \tau_i + \kappa_1/(TK^4)$ , where  $\tau_i$  satisfies the lower bound of Lemma 7 and is thus logarithmic in  $T$ . Under the assumption that  $T \geq K$ , for  $i \neq i^*$ ,

$$\mathbb{E}[T_i(T)] \leq C \frac{(c_i + \kappa_2 \sqrt{\log T})^{\max\{1/\alpha_i, 2\}}}{n_i \Delta_i^{\max\{1/\alpha_i, 2\}}} \quad (15)$$

for a constant  $C \leq 2 \cdot 4^{\max\{1/\alpha_i, 2\}}$ . Now we prove the high-probability bound. For this part, we need only concern ourselves with the sum of indicators from (14). Markov's inequality shows that

$$\mathbb{P}\left(\sum_{t=K+1}^T \mathbb{I}(i_t = i, T_i(t-1) \geq \tau) \geq 1\right) \leq \frac{\kappa_1}{TK^4}.$$

Thus we can assert that the bound (15) on  $T_i(T)$  holds with high probability.

**Remark:** By examining the proof of Theorem 2, it is straightforward to see that if we modify the multipliers on the  $\sqrt{\cdot}$  terms in the criterion (6) by  $m\kappa_2$  instead of  $\kappa_2$ , we get that the probability bound is of the order  $T^{3-4m^2} K^{-4m^2}$ , while the bound on  $T_i(T)$  is scaled by  $m^{1/\alpha_i}$ .

## 5. Model selection over nested hierarchies

In this section, we prove Theorem 6. The following proposition states that the class returned by the output of the procedure (12) satisfies an oracle inequality over the set  $S$ .

**Proposition 8** Let  $f = \hat{\mathcal{A}}(\hat{i}, n_{\hat{i}} T / s(\lambda))$  be the output of the algorithm  $\mathcal{A}$  for class  $\hat{i}$  specified in Equation 12. Under the conditions of Theorem 6, with probability at least  $1 - 3\kappa_1 \exp(-4m)$

$$R(f) \leq \min_{i \in S} \left\{ R_i^* + \gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{s(\lambda) \log i}{2Tn_i}} \right\} + \kappa_2 \sqrt{\frac{ms(\lambda)}{Tn_K}}.$$

The proof of the proposition follows from an argument similar to that given by Bartlett et al. (2002). We present a proof at the end of this section, since our setting is slightly different: each class receives a different number of independent samples. First, however, we complete the proof of Theorem 6 using the proposition.

**Proof of Theorem 6** Let  $i \in [K]$  be any class (not necessarily in  $S$ ), and let  $j \in S$  be the smallest class satisfying  $j \geq i$ . Then by construction of  $S$ , we know that

$$\gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) \leq \gamma_j \left( \frac{Tn_j}{s(\lambda)} \right) \leq (1 + \lambda) \gamma_i \left( \frac{Tn_i}{s(\lambda)} \right).$$

Thus we can lower bound the penalized risk of class  $i$  as

$$R_i^* + 2(1 + \lambda) \gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) \geq R_j^* + 2\gamma_j \left( \frac{Tn_j}{s(\lambda)} \right),$$

where we used the nesting assumption B to conclude that  $j \geq i$  implies  $R_j^* \leq R_i^*$ .

Now combining the above lower bound with the inequality in Proposition 8 yields that with probability at least  $1 - 3\kappa_1 \exp(-m)$

$$\begin{aligned} R(f) &\leq \min_{i \in S} \left\{ R_i^* + \gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{s(\lambda) \log i}{2Tn_i}} \right\} + \sqrt{\frac{ms(\lambda)}{Tn_K}} \\ &\leq \min_{i \in [K]} \left\{ R_i^* + 2(1 + \lambda) \gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) \right\} + \kappa_2 \sqrt{\frac{s(\lambda) \log K}{2Tn_K}} + \kappa_2 \sqrt{\frac{ms(\lambda)}{Tn_K}} \end{aligned}$$

since  $K \geq i$  and  $n_i \geq n_K$ . ■

**Proof of Proposition 8** To prove the proposition, we would like to control the probability

$$\begin{aligned} &\mathbb{P} \left[ R(f) > \min_{i \in S} \left\{ R_i^* + 2\gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{s(\lambda) \log i}{2Tn_i}} \right\} + \epsilon \right] \\ &\leq \underbrace{\mathbb{P} \left[ R(f) > \min_{i \in S} \left\{ \hat{R}_{Tn_i/s(\lambda)}(i) + \gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) \right\} + \epsilon/2 \right]}_{\mathcal{T}_1} \\ &\quad + \underbrace{\mathbb{P} \left[ \min_{i \in S} \left\{ \hat{R}_{Tn_i/s(\lambda)}(i) + \gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) \right\} > \min_{i \in S} \left\{ R_i^* + 2\gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{s(\lambda) \log i}{2Tn_i}} \right\} + \epsilon/2 \right]}_{\mathcal{T}_2} \end{aligned} \tag{16}$$

where the inequality follows from a union bound.

We now bound the terms  $\mathcal{T}_1$  and  $\mathcal{T}_2$  separately. To bound the terms, we first observe that by the construction (12), the minimum over the penalized empirical risk is attained for the class  $\hat{i}$ . We thus simplify  $\mathcal{T}_1$  as

$$\begin{aligned} \mathbb{P} \left[ R(f) > \min_{i \in S} \left\{ \widehat{R}_{Tn_i/s(\lambda)}(i) + \gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) \right\} + \epsilon/2 \right] &= \mathbb{P} \left[ R(f) > \left\{ \widehat{R}(f) + \gamma_{\hat{i}} \left( \frac{Tn_{\hat{i}}}{s(\lambda)} \right) \right\} + \epsilon/2 \right] \\ &\leq \kappa_1 \exp \left( -\frac{Tn_{\hat{i}}\epsilon^2}{\kappa_2^2 s(\lambda)} \right), \end{aligned}$$

where the inequality follows by application of Assumption A(a). To bound  $\mathcal{T}_2$  in the sum (16), we define  $f_i^* = \operatorname{argmin}_{f \in \mathcal{F}_i} R(f)$  so that  $R_i^* = R(f)$ . Noting that the event in  $\mathcal{T}_2$  implies that

$$\max_{i \in S} \left\{ \widehat{R}_{Tn_i/s(\lambda)}(i) - R_i^* - \gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) - \kappa_2 \sqrt{\frac{s(\lambda) \log i}{2Tn_i}} \right\} > \frac{\epsilon}{2},$$

we can use the union bound to see

$$\begin{aligned} \mathcal{T}_2 &\leq \mathbb{P} \left[ \sup_{i \in S} \left\{ \widehat{R}_{Tn_i/s(\lambda)}(i) - R_i^* - \gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) - \kappa_2 \sqrt{\frac{s(\lambda) \log i}{2Tn_i}} \right\} > \frac{\epsilon}{2} \right] \\ &\leq \sum_{i \in S} \mathbb{P} \left[ \widehat{R}_{Tn_i/s(\lambda)}(i) - R_i^* - \gamma_i \left( \frac{Tn_i}{s(\lambda)} \right) - \kappa_2 \sqrt{\frac{s(\lambda) \log i}{2Tn_i}} > \frac{\epsilon}{2} \right] \\ &\leq \sum_{i \in S} \mathbb{P} \left[ \widehat{R}(f_i^*) - R_i^* > \frac{\epsilon}{2} + \kappa_2 \sqrt{\frac{s(\lambda) \log i}{2Tn_i}} \right], \end{aligned}$$

where the final inequality uses Assumption A(b), which states that  $\mathcal{A}$  outputs a  $\gamma_i$ -minimizer of the empirical risk. Now we can bound the deviations using Assumption A(d), since  $f_i^*$  is non-random:

$$\mathcal{T}_2 \leq \sum_{i \in S} \kappa_1 \exp \left( -\frac{Tn_i \epsilon^2}{s(\lambda) \kappa_2^2} \right) \exp(-2 \log i).$$

Setting  $\epsilon = \kappa_2 \sqrt{\frac{ms(\lambda)}{Tn_K}}$ , see that the first term in bounding  $\mathcal{T}_2$  reduces to  $\exp(-mn_i/n_K) \leq \exp(-m)$  since  $n_i \geq n_K$ . Then we get

$$\begin{aligned} \mathcal{T}_2 &\leq \sum_{i \in S} \kappa_1 \exp(-m) \exp(-2 \log i) \\ &\leq 2\kappa_1 \exp(-m), \end{aligned}$$

where the last step uses  $\sum_{i=1}^{\infty} 1/i^2 = \pi^2/6 \leq 2$ . Finally, plugging the stated setting of  $\epsilon$  into the bound on  $\mathcal{T}_1$  completes the proof. ■

## Acknowledgments

In performing this research, AA was supported by a Microsoft Research Fellowship, and JCD was supported by the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program. AA and PB gratefully acknowledge the support of the NSF under award DMS-0830410.

## References

- H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974.
- P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003.
- Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002. ISSN 0885-6125.
- A. R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric functional estimation and related topics*, pages 561–576. Kluwer Academic, 1991.
- Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.
- P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
- P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.
- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:290–330, 1967.
- R. M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6(6):899–929, 1978.
- S. Geman and C. R. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics*, 10:401–414, 1982.
- T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.
- G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *Annals of Statistics*, 32(4):1679–1697, 2004.
- C. L. Mallows. Some comments on  $C_p$ . *Technometrics*, 15(4):661–675, 1973.
- P. Massart. Concentration inequalities and model selection. In J. Picard, editor, *Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003 Series*. Springer, 2003.
- A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
- David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.
- Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983.

V. N. Vapnik and A. Ya. Chervonenkis. *Theory of pattern recognition.* Nauka, Moscow, 1974. (In Russian).

## Appendix A. Proof of Lemma 7

Following Auer et al. (2002), we show that the event in the lemma occurs with very low probability by breaking it up into smaller events more amenable to analysis. Recall that we're interested in controlling the probability of the event

$$\bar{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \leq \bar{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}} \quad (17)$$

For this bad event to happen, at least one of the following three events must happen:

$$\hat{R}_{n_i s_i}(\mathcal{A}(i, n_i s_i)) - \inf_{f \in \mathcal{F}_i} R(f) \leq -\gamma_i(n_i s_i) - \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \quad (18a)$$

$$\hat{R}_{n_{i^*} s_{i^*}}(\mathcal{A}(i^*, n_{i^*} s_{i^*})) - \inf_{f \in \mathcal{F}_{i^*}} R(f) \geq \gamma_{i^*}(n_{i^*} s_{i^*}) + \kappa_2 \sqrt{\frac{\log K}{n_{i^*} s_{i^*}}} + \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}} \quad (18b)$$

$$R_i^* + \gamma_i(T n_i) \leq R^* + \gamma_{i^*}(T n_{i^*}) + 2 \left( \gamma_i(n_i s_i) + \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} + \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \right). \quad (18c)$$

Temporarily use the shorthand  $f_i = \mathcal{A}(i, n_i s_i)$  and  $f_{i^*} = \mathcal{A}(i^*, n_{i^*} s_{i^*})$ . The relationship between Eqs. (18a)–(18c) and the event in (17) follows from the fact that if none of (18a)–(18c) occur, then

$$\begin{aligned} & \bar{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \\ &= \hat{R}_{n_i s_i}(f_i) + \gamma_i(T n_i) - \gamma_i(n_i s_i) - \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \\ &\stackrel{(18a)}{>} \inf_{f \in \mathcal{F}_i} R(f) + \gamma_i(T n_i) - 2 \left( \gamma_i(n_i s_i) + \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} + \kappa_2 \sqrt{\frac{\log t}{n_i s_i}} \right) \\ &\stackrel{(18c)}{>} \inf_{f \in \mathcal{F}_{i^*}} R(f) + \gamma_{i^*}(T n_{i^*}) + 2 \left( \gamma_{i^*}(n_{i^*} s_{i^*}) + \kappa_2 \sqrt{\frac{\log K}{n_{i^*} s_{i^*}}} + \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}} \right) \\ &\quad - 2 \left( \gamma_i(n_i s_i) + \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} + \kappa_2 \sqrt{\frac{\log n}{n_i s_i}} \right) \\ &\stackrel{(18b)}{>} \hat{R}_{n_{i^*} s_{i^*}}(f_{i^*}) + \gamma_{i^*}(T n_{i^*}) - \gamma_i(n_i s_i) - \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} - \kappa_2 \sqrt{\frac{\log t}{n_i s_i}} \\ &= \bar{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log t}{n_{i^*} s_{i^*}}}. \end{aligned}$$

From the above string of inequalities, to show that the event (17) has low probability, we need simply show that each of (18a), (18b), and (18c) have low probability.

To prove that each of the bad events have low probability, we note the following consequences of Assumption A. Recall the definition of  $f_i^*$  as the minimizer of  $R(f)$  over the class  $\mathcal{F}_i$ . Then by Assumption A(a),

$$R(f_i^*) - \gamma_i(n) - \kappa_2\epsilon \leq R(\mathcal{A}(i, n)) - \gamma_i(n) - \kappa_2\epsilon < \hat{R}_n(\mathcal{A}(i, n)),$$

while Assumptions A(b) and A(d) imply

$$\hat{R}_n(\mathcal{A}(i, n)) \leq \hat{R}_n(f_i^*) + \gamma_i(n) \leq R(f_i^*) + \gamma_i(n) + \kappa_2\epsilon,$$

each with probability at least  $1 - \kappa_1 \exp(-4n\epsilon^2)$ . In particular, we see that the events (18a) and (18b) have low probability:

$$\begin{aligned} \mathbb{P} \left[ \hat{R}_{n_i s_i}(\mathcal{A}(i, n_i s_i)) - R(f_i^*) \leq -\gamma_i(n_i s_i) - \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \right] \\ \leq \kappa_1 \exp \left( -4n_i s_i \left( \frac{\log K}{n_i s_i} + \frac{\log T}{n_i s_i} \right) \right) = \frac{\kappa_1}{(tK)^4} \\ \mathbb{P} \left[ \hat{R}_{n_{i^*} s_{i^*}}(\mathcal{A}(i^*, n_{i^*} s_{i^*})) - R^* \geq \gamma_{i^*}(n_{i^*} s_{i^*}) + \kappa_2 \sqrt{\frac{\log K}{n_{i^*} s_{i^*}}} + \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}} \right] \\ \leq \kappa_1 \exp \left( -4n_{i^*} s_{i^*} \left( \frac{\log K}{n_{i^*} s_{i^*}} + \frac{\log T}{n_{i^*} s_{i^*}} \right) \right) = \frac{\kappa_1}{(tK)^4}. \end{aligned}$$

What remains is to show that for large enough  $\tau$ , (18c) does not happen. Recalling the definition that  $R^* + \gamma_{i^*}(Tn_{i^*}) = R_i^* + \gamma_i(Tn_i) - \Delta_i$ , we see that for (18c) to fail it is sufficient that

$$\Delta_i > 2\gamma_i(\tau n_i) + 2\kappa_2 \sqrt{\frac{\log K}{n_i \tau}} + 2\kappa_2 \sqrt{\frac{\log T}{n_i \tau}}.$$

Let  $x \wedge y := \min\{x, y\}$  and  $x \vee y := \max\{x, y\}$ . Since  $\gamma_i(n) \leq c_i n^{-\alpha_i}$ , the above is satisfied when

$$\frac{\Delta_i}{2} > c_i(\tau n_i)^{-(\alpha_i \wedge \frac{1}{2})} + \kappa_2 \sqrt{\log K}(\tau n_i)^{-(\alpha_i \wedge \frac{1}{2})} + \kappa_2 \sqrt{\log T}(\tau n_i)^{-(\alpha_i \wedge \frac{1}{2})} \quad (19)$$

We can solve (19) above and see immediately that if

$$\tau_i > \frac{2^{1/\alpha_i \vee 2} (c_i + \kappa_2 \sqrt{\log T} + \kappa_2 \sqrt{\log K})^{1/\alpha_i \vee 2}}{n_i \Delta_i^{1/\alpha_i \vee 2}},$$

then

$$R_i^* > R^* + 2 \left( \gamma_i(n_i \tau_i) + \kappa_2 \sqrt{\frac{\log K}{n_i \tau_i}} + \kappa_2 \sqrt{\frac{\log T}{n_i \tau_i}} \right). \quad (20)$$

Thus the event in (18c) fails to occur, completing the proof of the lemma.

AGARWAL DUCHI BARTLETT LEVRARD

# Bandits, Query Learning, and the Haystack Dimension

**Kareem Amin**

**Michael Kearns**

**Umar Syed**

*Department of Computer and Information Science  
University of Pennsylvania  
Philadelphia, PA 19104*

AKAREEM@CIS.UPENN.EDU

MKEARNS@CIS.UPENN.EDU

USYED@CIS.UPENN.EDU

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

Motivated by multi-armed bandits (MAB) problems with a very large or even infinite number of arms, we consider the problem of finding a maximum of an unknown target function by querying the function at chosen inputs (or arms). We give an analysis of the query complexity of this problem, under the assumption that the payoff of each arm is given by a function belonging to a known, finite, but otherwise arbitrary function class. Our analysis centers on a new notion of function class complexity that we call the *haystack dimension*, which is used to prove the approximate optimality of a simple greedy algorithm. This algorithm is then used as a subroutine in a functional MAB algorithm, yielding provably near-optimal regret.

We provide a generalization to the infinite cardinality setting, and comment on how our analysis is connected to, and improves upon, existing results for query learning and generalized binary search.

**Keywords:** Multi-armed bandits, learning theory

## 1. Introduction

A *multi-armed bandit* (MAB) problem proceeds over several rounds, and in each round a decision-maker chooses an action, or *arm*, and receives a random payoff from an unknown distribution associated with the chosen action. MAB problems have been a focus of intensive study in the statistics and machine learning literature because they are an excellent model of many real-world sequential decision making problems that contain an “exploration vs. exploitation” trade-off, such as problems in clinical trials, sponsored web search, quantitative finance, and many other areas. The performance of a MAB algorithm is measured by its *regret*, which is the difference in expected total payoff received by the algorithm and by a highest-payoff action.

The classical formulation of the MAB problem assumes that the set of arms or actions is finite, and regret guarantees typically depend linearly on the number of actions. But many interesting applications have an extremely large, or even infinite, number of actions. As just one example, in (organic or sponsored) web search, a core goal is to select web pages in response to user queries in order to maximize click-throughs; even for a fixed query, the number of possible response web sites may be sufficiently large as to be effectively infinite, thus requiring some notion of similarity or generalization across actions.

Achieving regret that is sublinear in the number of actions clearly requires assumptions. A natural way to make the problem feasible is to make some specific functional assumptions about the action payoffs, i.e., to assume that the expected payoff of each action  $x \in \mathcal{X}$  is given by  $f^*(x)$ , where  $f^* \in \mathcal{F}$  is an unknown function belonging to a function class  $\mathcal{F}$ . One approach along these lines was pioneered by Kleinberg et al. (2008), who assumed that the set of actions  $\mathcal{X}$  is endowed with a metric, and that each  $f \in \mathcal{F}$  is Lipschitz continuous with respect to this metric. In this way, the observed payoff of any action provides information about the payoffs of “nearby” actions (with respect to the metric). Kleinberg et al. (2008) described efficient algorithms that exploit the structure of  $\mathcal{F}$  to achieve no-regret. Nearly all existing algorithms for functional MAB problems rely on some kind of smoothness assumption (Bubeck et al., 2008; Lu et al., 2010; Slivkins, 2011; Flaxman et al., 2005; Dani and Hayes., 2006; Abernethy et al., 2008; Srinivas et al., 2008).

In this paper we consider the MAB problem for a *general* function class  $\mathcal{F}$ , and focus on the number of rounds required to achieve low regret (ignoring computational efficiency). We give a characterization in terms of a new measure of the complexity of the function class  $\mathcal{F}$  that we call the *haystack dimension*,

which intuitively captures the extent to which maximizing a function via queries requires a search for a small number of items (needles) amongst a much larger number of otherwise undifferentiated possibilities (a haystack). We then give upper and lower bounds involving the haystack dimension of  $\mathcal{F}$  that are within a  $\log |\mathcal{F}|$  factor. Note that for the hardest MAB problems — where the haystack dimension can be as large as  $|\mathcal{F}|$  — this logarithmic factor is relatively benign. Our main results are graphically summarized in Figure 1.

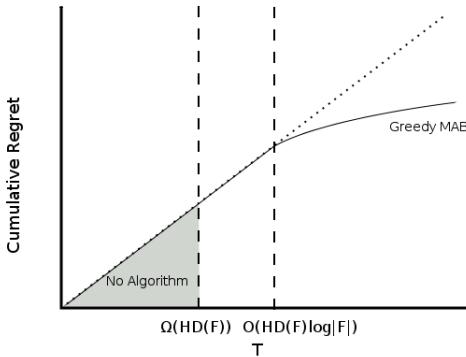


Figure 1: A graphical summary of the main results of this paper. This figure illustrates that *any* MAB algorithm for a function class  $\mathcal{F}$  must suffer linear regret for a number of rounds on the order of the haystack dimension (denoted  $HD(\mathcal{F})$ ), while the MAB algorithm presented in this paper, called Greedy MAB, begins to suffer sublinear regret after roughly  $HD(\mathcal{F}) \log |\mathcal{F}|$  rounds.

An interesting aspect of our methods is the connection drawn between MAB problems and the problem of exact learning of functions from queries. We observe that functional MAB problems implicitly embed the problem of finding a maximum of an unknown function in  $\mathcal{F}$  from only input-output queries (generalizations of membership queries), which may or may not be much easier than exact learning. Our analysis shows that any functional MAB algorithm must implicitly be willing to trade off between two distinct types of queries:

*max queries* (which attempt to directly guess the maximum of  $f^*$ ) and *information queries* (which attempt to make progress by reducing the version space, as is traditional in many query learning models). We show that either one of these query types (and essentially no others) can result in progress towards finding the maximum. The haystack dimension can then be viewed as a measure of the extent to which progress can be made at any step via either one of these query types.

Our characterization holds for any finite-cardinality  $\mathcal{F}$  (though even the number of *actions* may still be infinite), but we also describe a generalization to the case of infinite  $\mathcal{F}$  via covering techniques (which in general does not provide as tight bounds as its finite-cardinality specialization). We also stress that our results only apply to query complexity and regret; we make no claims about computational efficiency (necessarily, due to the generality of our setting). In this sense, the haystack dimension can be seen as playing a role in the study of functional MAB problems analogous to that played by quantities such as VC dimension and teaching dimension in other learning models, which also characterize sample or informational complexity, but not computational complexity. In separate work (Amin et al., 2011), we have developed computationally efficient algorithms for functional *contextual* MAB problems (Langford and Zhang, 2007), in which the payoff function depends on both the chosen action and the current context, both of which may be drawn from very large spaces.

## 2. Related Work

Many authors (at least since Thompson (1933)) have studied finite MAB problems where the action payoffs are assumed to be correlated; see Mersereau et al. (2009, p. 4) for an excellent survey. As explained in Section 1, more recent work has focused on infinite MAB problems where the action payoffs are related via an unknown function belonging to a known function class, such as the set of all Lipschitz continuous functions (Kleinberg et al., 2008; Bubeck et al., 2008; Slivkins, 2011). Compared to previous work, our results provide a complete analysis for significantly more general function classes.

Obviously, maximizing an unknown function via queries is no harder than exactly learning the function. Hegedüs (1995) characterized the query complexity of exact learning in terms of the *extended teaching dimension* of the function class. For some restricted function classes the haystack dimension and extended teaching dimension coincide<sup>1</sup>, and in these cases our analysis approximately recovers the bounds due to Hegedüs (1995), but with a significant advantage: our lower bound holds for all *randomized* algorithms, while the earlier bound only applied to deterministic algorithms.

A variant of exact learning (but not maximization) of functions has been considered under the name of *generalized binary search*. Nowak (2009) provided an analysis that only applies under a certain technical condition. In the language of this paper, the condition implies that the haystack dimension is a constant independent of the structure of the function class. In contrast, our analysis applies to any  $\mathcal{F}$  and considers the maximization problem and its relationship to MAB directly.

---

1. Essentially just those classes for which maximizing an unknown function is as difficult as exactly learning it; see Example 2.

### 3. Functional Bandits (MAB) and Maximizing From Queries (MAX)

A *functional MAB problem* is defined by a set of *actions*  $\mathcal{X}$  and a set of possible *payoff functions*  $\mathcal{F}$ . A target payoff function  $f^* \in \mathcal{F}$  is selected. In each round  $t = 1, 2, \dots$ , an algorithm selects an action  $x_t$ , and receives an independent payoff from a distribution whose support is contained in  $[-b, b]$ , and has mean  $f^*(x_t)$ . The goal of the algorithm is to receive nearly as much cumulative payoff as an algorithm that selects the best action every round. More precisely, the worst-case expected *regret* of algorithm  $A$  in round  $T$  is  $R_A(T) \triangleq \sup_{f^* \in \mathcal{F}} E \left[ T \cdot \sup_{x \in \mathcal{X}} f^*(x) - \sum_{t=1}^T f^*(x_t) \right]$ , where the expectation is with respect to the random payoffs and any internal randomization of the algorithm. We say that algorithm  $A$  is *no-regret* if  $\lim_{T \rightarrow \infty} R_A(T)/T = 0$ .

For any functional MAB problem, we can define a corresponding and (as we shall see) closely related functional *maximizing from queries* (or MAX) problem: In each round  $t = 1, 2, \dots$  an algorithm  $A$  selects a *query*  $x_t \in \mathcal{X}$ , and then observes  $y_t = f^*(x_t)$ . Letting  $X^f$  be the set of maxima of the function  $f$  (assumed to be non-empty), the goal of the algorithm is to eventually select an  $x \in X^{f^*}$ . Let  $T^{A,f^*} = \min\{t : x_t \in X^{f^*}\}$  be the first round such that  $x_t$  is a maximum of  $f^*$ . We are interested in bounding the worst-case expected *query complexity*  $Q_A \triangleq \sup_{f^* \in \mathcal{F}} E[T^{A,f^*}]$ , where the expectation is with respect to any internal randomization of the algorithm.

While the definition of query complexity says that algorithm  $A$  will *select* a query in  $X^{f^*}$  within  $Q_A$  rounds, it does not require that the algorithm be able to *identify* this query. However, if  $A$  is deterministic and an upper bound  $B$  on  $Q_A$  is known, then the latter problem easily reduces to the former: If  $Q_A \leq B$  then  $x_{t^*} \in X^{f^*}$ , where  $t^* \in \arg \max_{1 \leq t \leq B} y_t$ . We will describe a deterministic algorithm with near-optimal query complexity in Section 6.

It is important to note that, in a functional MAB problem, in each round  $t$  the algorithm only observes a *sample* from a distribution with mean  $f^*(x_t)$ , while in a functional MAX problem, the algorithm observes  $f^*(x_t)$  *directly*. In Sections 4–7, we characterize the query complexity of the MAX problem for  $\mathcal{F}$ , and then apply these results in Sections 9–10 to characterize the optimal regret for the corresponding MAB problem for  $\mathcal{F}$ . Consequently, the analysis in Sections 4–7 will not deal with stochasticity, which is addressed afterwards. Also, we refer to elements of  $\mathcal{X}$  as *actions* in the MAB context, but as *queries* in the MAX context — this difference exists only to agree with historical usage.

### 4. The Haystack Dimension

In this section we give the definition of the haystack dimension for function classes  $\mathcal{F}$  of *finite cardinality*; generalization to the infinite case is given later.

The formal definition of the haystack definition requires some notation and machinery, but the intuition behind it is rather simple, so we first describe it informally. In words, the haystack dimension identifies the “worst” subset of  $\mathcal{F}$ , in the sense that on that subset, no matter what query is made and no matter what response is received, only a small fraction  $\theta$  of the functions in the subset are eliminated due to inconsistency with the query, or are maximized by the query. It turns out mathematically that the right definition of the

haystack dimension is the inverse quantity  $1/\theta$  for this worst subset. We now proceed with the formal definition.

In the context of a MAX problem, a query  $x \in \mathcal{X}$  can be thought of as providing information about the identity of  $f^*$ . In particular,  $f^*$  cannot be any of the functions in  $\mathcal{F}$  inconsistent with the value  $f^*(x)$  observed at  $x$ . So one strategy for finding an element of  $X^{f^*}$  is to first issue a sequence of *information queries*, that uniquely identify  $f^*$ , and then select any  $x \in X^{f^*}$ .<sup>2</sup>

However, sometimes identifying the true function  $f^*$  exactly requires many more queries than necessary for maximization. For example, if many functions  $f \in \mathcal{F}$  are maximized by one particular query, it may be useful to play such a *max query*, even if it is not particularly useful for learning  $f^*$ .

In the extreme case, there might exist an  $x^* \in \mathcal{X}$  such that  $x^* \in X^f$  for all  $f \in \mathcal{F}$ . In this case, an element of  $X^{f^*}$  can be selected in one query, without ever needing to identify  $f^*$ . On the other hand, there are also  $\mathcal{F}$  for which exact learning is the fastest route to maximization.<sup>3</sup>

Any general algorithm for maximization from queries thus needs to be implicitly able to consider queries that would eliminate many candidate functions, as well as queries that might be an actual maximum.

Before continuing, we define some convenient notation that we will use throughout the rest of the paper. For any set  $F \subseteq \mathcal{F}$ , define the *inconsistent set*  $F(\langle x, y \rangle) \triangleq \{f \in F : f(x) \neq y\}$  to be the functions in  $F$  that are inconsistent with the query-value pair  $\langle x, y \rangle$ . Also, for any set  $F \subseteq \mathcal{F}$ , define the *maximum set*  $F(x) \triangleq \{f \in F : x \in X^f\}$  to be the functions in  $F$  for which  $x$  is a maximum.

Intuitively, the *haystack dimension*  $\text{HD}(\mathcal{F})$  of a function class  $\mathcal{F}$  will characterize a subset of  $\mathcal{F}$  on which no query is effective in the two senses previously discussed. For a subset  $F \subseteq \mathcal{F}$ , let

$$\rho(F, x) = \inf_{y \in \mathbb{R}} \frac{|F(x) \cup F(\langle x, y \rangle)|}{|F|} \quad \text{and} \quad \rho(F) = \sup_{x \in \mathcal{X}} \rho(F, x).$$

$\rho(F, x)$  is the fraction of functions in  $F$ , which are guaranteed to be maximized, or deemed inconsistent, by the query  $x$ , for the worst-case possibility for  $y$ . If  $\rho(F)$  is small, no query is guaranteed to be effective as either a max or information query on the subset  $F$ . Now let  $F_\theta = \arg \inf_{F \subseteq \mathcal{F}} \rho(F)$  and  $\theta = \rho(F_\theta)$ .

**Definition 1** Let  $F_\theta$  and  $\theta$  be defined as above. Then the haystack dimension  $\text{HD}(\mathcal{F})$  of  $\mathcal{F}$  is defined as  $\frac{1}{\theta}$ .

Note that the haystack dimension can be as small as 1 (all functions share a common maximum-output input) and as large as  $|\mathcal{F}|$  (every query eliminates at most one function in the senses discussed, the canonical “needle in a haystack”).

---

2. In the case of boolean functions or concepts, identifying  $f^*$  exactly is the problem of *learning from membership queries* (Angluin, 1988).  
 3. See Example 2.

## 5. Examples of the Haystack Dimension

In this section, we provide a few function classes which help illustrate how the haystack dimension characterizes the difficulty of maximizing an unknown  $f^* \in \mathcal{F}$ . Many of these examples will be useful for subsequent constructions in the paper.

The first construction considered is the “needle in a haystack”. Fix a finite  $\mathcal{X}$ . For each  $x \in \mathcal{X}$ , let  $f_x$  be the function defined to have  $f_x(x) = 1$  and  $f_x(x') = 0$  for all  $x' \neq x$ . Now let  $\mathcal{H}_{\mathcal{X}} = \{f_x \mid x \in \mathcal{X}\}$ .

**Example 1**  $HD(\mathcal{H}_{\mathcal{X}}) = |\mathcal{H}_{\mathcal{X}}|$

**Proof** For any  $x \in \mathcal{X}$ ,  $f_x$  is the only function in  $\mathcal{H}_{\mathcal{X}}$  that attains its max at  $x$ . Furthermore, all other functions output a 0 on input  $x$ . Therefore, letting  $F = \mathcal{H}_{\mathcal{X}}$ ,  $F(x) = \{f_x\}$  and  $F(\langle x, 0 \rangle) = \{f_x\}$  for any  $x \in \mathcal{X}$ . This implies that  $\rho(\mathcal{H}_{\mathcal{X}}) = |\mathcal{H}_{\mathcal{X}}|^{-1}$ . Since  $\rho(F)$  cannot be smaller than this quantity for any  $F \subseteq \mathcal{F}$ , the haystack dimension of  $\mathcal{H}_{\mathcal{X}}$  indeed equals  $|\mathcal{H}_{\mathcal{X}}|$ . ■

When  $f^* \in \mathcal{H}_{\mathcal{X}}$ , maximizing and learning  $f^*$  coincide, and both amount to guessing the  $x$  for which  $f^*(x) = 1$ .

We now describe a function class  $\mathcal{G}$  in which any algorithm is essentially forced to learn the true function  $f^*$ .

The input space  $\mathcal{X}$  will consist of two components —  $\mathcal{X}_1$  and  $\mathcal{X}_2$ , with  $\mathcal{X}$  being the union of these disjoint domains. The high-level idea is to “marry” a small shattered set (in the sense of VC dimension) to a much larger haystack construction. Subdomain  $\mathcal{X}_1$  consists of  $n$  points  $\{a_0, \dots, a_{n-1}\}$ . We construct  $\mathcal{G}$  as follows. On  $\mathcal{X}_1$ , all possible binary labelings of the  $n$  points appear in  $\mathcal{G}$ , giving a total of  $2^n$  functions. Let us think of each function in  $\mathcal{G}$  as being equated with the integer given by its binary labeling of the points in  $\mathcal{X}_1$ . So a function  $f$  is equated with the integer  $z(f) = \sum_{i=0}^{n-1} 2^i f(a_i)$ . Now let the much larger set  $\mathcal{X}_2 = \{0, \dots, 2^n - 1\}$ , and for any  $x \in \mathcal{X}_2$  define  $f(x) = 2$  if  $x = z(f)$  and  $f(x) = 0$  otherwise.

Thus, the behavior of a function on  $\mathcal{X}_1$  entirely defines the function on  $\mathcal{X}_2$  as well, and the labeling on  $\mathcal{X}_1$  gives us the index of the function’s maximum, which is always equal to 2 and occurs at exactly one point in  $\mathcal{X}_2$  (determined by the index).

Note that there is an algorithm that finds the max of  $f^*$  in  $O(n)$  queries simply by querying every input in  $\mathcal{X}_1$  and learning the identity of  $f^*$  exactly. Intuitively, no algorithm can do much better. To see why, suppose  $f^*$  were drawn uniformly at random from  $\mathcal{G}$ . Note that an action  $r \in \mathcal{X}_2$  has an exponentially small probability of being the function’s max and, in the event that  $f^*(r) = 0$ , only serves to inform the algorithm that  $f^*$  is not the single  $f \in \mathcal{G}$  with  $z(f) = r$ . Also, observe that if the “zooming” algorithm of Kleinberg et al. (2008) is applied to  $\mathcal{G}$ , it will take exponential time in the worst-case to find a maximum of  $f^*$ , essentially because it makes no attempt to exploit the special structure of  $\mathcal{G}$ .

**Example 2**  $HD(\mathcal{G}) = \Theta(n) = \Theta(\log |\mathcal{G}|)$

**Proof** We first show that there is an  $F \subseteq \mathcal{G}$  with  $\rho(F) = \frac{1}{n}$ . For each  $x \in \mathcal{X}_1$ , let  $f_x$  be the function which outputs  $f_x(x) = 1$ , and  $f_x(x') = 0$  for all  $x' \in \mathcal{X}_1, x \neq x'$ . Let  $F = \{f_x \in \mathcal{G} \mid x \in \mathcal{X}_1\} = \{f \mid z(f) \in \{2^0, 2^1, \dots, 2^{n-1}\}\}$ .  $|F| = n$ .

Consider any query  $x \in \mathcal{X}_1$ .  $F(x) = \emptyset$ , since no functions achieve their maximum on a query in  $\mathcal{X}_1$ . Furthermore,  $F(\langle x, 0 \rangle) = \{f_x\}$ , since  $f_x$  is the only function in  $F$  which doesn't output a 0 on query  $x$ . Thus  $\rho(F, x) = \frac{1}{n}$  for any  $x \in \mathcal{X}_1$ . For a query  $r \in \mathcal{X}_2$ ,  $\rho(F, r) \leq \frac{1}{n}$ , since at most one function in  $F$  (and  $\mathcal{G}$ ) achieves its maximum at  $r$ , and all other functions output a zero at  $r$ . Thus,  $\rho(F) = \frac{1}{n}$ .

We now argue that for an arbitrary  $F \subseteq \mathcal{G}$ ,  $\rho(F) \geq \frac{1}{2n}$ . Let  $r_1(x) = |f \in F : f(x) = 1|/|F|$ , be the fraction of functions that exhibit a 1 at action  $x \in \mathcal{X}_1$ . Let  $r_0(x) = 1 - r_1(x)$ . Suppose there is an action  $x \in \mathcal{X}_1$  such that  $r_1(x) \geq \frac{1}{2n}$  and  $r_0(x) \geq \frac{1}{2n}$ . Then, at least  $\frac{1}{2n}$  of the functions in  $F$  would be inconsistent with any observed output. That is, both  $\frac{|F(\langle x, 0 \rangle)|}{|F|} \geq \frac{1}{2n}$  and  $\frac{|F(\langle x, 1 \rangle)|}{|F|} \geq \frac{1}{2n}$ .

Otherwise, for every  $x$  in  $\mathcal{X}_1$ , either  $r_1(x) < \frac{1}{2n}$  or  $r_0(x) < \frac{1}{2n}$  (i.e. one outcome is quite rare). This implies that more than  $1/2$  of the functions in  $F$  exhibit the same behavior on all  $x$  in  $\mathcal{X}_1$ . However, unless  $F$  is a singleton set (in which case  $\rho(F) = 1$ ), this cannot occur since each  $f \in \mathcal{G}$  exhibits unique behavior on  $\mathcal{X}_1$ . ■

The preceding example illustrates a function class for which any algorithm querying for the max must ultimately learn the true function  $f^*$ . However, the opposite extreme is also possible. Consider a function class  $\mathcal{G}_{\max}$ . Let there be a distinguished  $x^* \in \mathcal{X}$  such that every  $f \in \mathcal{G}_{\max}$  attains its maximum at  $x^*$ . It may be arbitrarily difficult to learn the behavior of  $f^*$  on the remainder of  $\mathcal{X}$ . However, finding the maximum can be done trivially in a single query.

**Example 3**  $HD(\mathcal{G}_{\max}) = 1$

**Proof** For every  $F \subseteq \mathcal{G}_{\max}$   $F(x^*) = F$ , and the proof is immediate. ■

Finally, there are function classes which require a hybrid between learning and maximization. We construct such a class,  $\mathcal{G}^+$ . Let  $\mathcal{X}$  be the disjoint union of three sets  $\mathcal{X}_1 \cup \mathcal{X}_2 \cup \mathcal{X}_3$ . We let  $|\mathcal{X}_1| = n$ ,  $|\mathcal{X}_2| = m$  and create a function in  $\mathcal{G}^+$  for each binary labeling of  $\mathcal{X}_1$  and  $\mathcal{X}_2$ , as in the construction of  $\mathcal{G}$ . We let  $z_1(f)$  be the integer corresponding to the labeling  $f$  gives  $X_1$  and  $z_2(f)$  be the integer corresponding to the labeling  $f$  gives  $X_2$ .

Now let  $\mathcal{X}_3 = M_0 \cup M_1 \cup \dots \cup M_{2^n-1}$  where each  $|M_i| = c$ . Again, like in the construction of  $\mathcal{G}$ , for each  $f \in \mathcal{G}^+$ , there will be a single  $r \in \mathcal{X}_3$  such that  $f(r) = 2$  and  $f(r) = 0$  otherwise. However, this time, the maximizing element of  $f$  will be found in  $M_{z_1(f)}$ . And the particular element of  $M_{z_1(f)}$  will be given by  $z_2(f) \bmod c$ .

If we think of  $c < n < m$ , then there is an  $n + c$  algorithm for finding the max of  $\mathcal{F}$ . The algorithm learns the set  $M_i$  which contains the max of  $f^*$  by querying each element of  $\mathcal{X}_1$ . It then tries each of the  $c$  elements in  $M_i$ . Note that the true identity of  $f^*$  is never learned, and the “interesting” learning problem is discovering the set  $M_i$  containing the maximizing action (i.e., learning the behavior on  $\mathcal{X}_1$ ).

**Example 4** If  $c < n < m$ ,  $HD(\mathcal{G}^+) = \Theta(n)$

**Proof** We sketch the proof which proceeds almost identically to that of Example 2. There is an  $F$  with that witnesses  $\rho(F) = \frac{1}{n}$ .  $F$  is identical to that used in Example 2 on  $\mathcal{X}_1$ . The

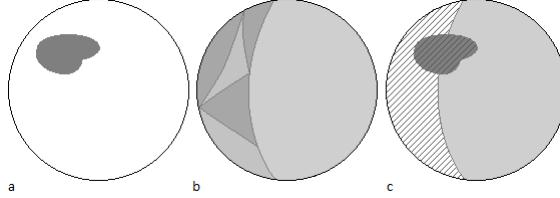


Figure 2: (a) The shaded region represents the subset of functions which attain a maximum at some input  $x$ . (b) Querying  $x$  also induces a partition of the function space, in which each piece of the partition contains functions that output the same value on  $x$ . The least informative query response is thus the largest piece of this partition. (c) Upon querying  $x$ , the greedy algorithm eliminates from its attention space a set *at least* the size of the striped region. This region is the union of the maximizing set and the *complement* of the largest partition piece.

behavior on  $\mathcal{X}_2$  is identical across all functions, and the behavior on  $\mathcal{X}_3$  is determined by these choices.

To see that for any  $F$ ,  $\rho(F) \geq \frac{1}{2n}$ , we use the same reasoning as Example 2. However, if for every  $x$  in  $\mathcal{X}_1$ , either  $r_1(x) < \frac{1}{2n}$  or  $r_0(x) < \frac{1}{2n}$ , rather than implying a contradiction, this implies that there is actually a query  $x^* \in \mathcal{X}_3$  such that  $x^*$  maximizes more than a  $\frac{1}{2c}$  fraction of the functions in  $F$ . Therefore for that particular  $F$ ,  $\rho(F) > \frac{1}{2c} > \frac{1}{2n}$ , as desired. ■

## 6. MAX Query Complexity Upper Bound

Before arguing that the haystack dimension gives lower bounds on the query complexity of any algorithm on  $\mathcal{F}$ , we observe that it motivates a simple, natural, greedy algorithm. In each round  $t$ , the *greedy algorithm*  $G$  selects  $x_t = \arg \sup_{x \in \mathcal{X}} \rho(F_t, x)$  where the *attention space*  $F_t$  is defined inductively as follows:  $F_1 = \mathcal{F}$  and  $F_{t+1} = F_t \setminus (F_t(x_t) \cup F_t(\langle x_t, y_t \rangle))$ . (Note that for fixed  $F_t$ ,  $\rho(F_t, x)$  takes values in  $\{1/|F_t|, 2/|F_t|, \dots, 1\}$ , so the supremum is achieved by an  $x \in \mathcal{X}$ .)

Essentially, the greedy algorithm always selects the query that results in the largest guaranteed reduction of the attention space<sup>4</sup>; see Figure 2. It is important to note that the attention space differs from the version space of traditional learning in that it *excludes functions that may still be consistent with the observed data*, but for which the algorithm has already played a query which would have maximized such functions. Indeed, in the extreme case the algorithm may choose to query  $x$  such that *all functions remaining* have the same output on  $x$  — in which case the query conveys no “information” in the traditional learning sense, but nevertheless functions attaining their maximum at  $x$  will be discarded.

**Theorem 2** *Let  $G$  be the greedy algorithm for the MAX problem for  $\mathcal{F}$ . Then  $Q_G \leq \text{HD}(\mathcal{F}) \log |\mathcal{F}|$ .*

---

4. Note that we assume the specified computations can in fact be implemented in finite computation time.

**Proof**

We know that the attention space  $F_t$  is nonempty for all rounds  $t \leq Q_G$  (otherwise the greedy algorithm would have selected a maximum of  $f^*$  before round  $Q_G$ ). And at least a  $\theta$  fraction of the attention space is removed in each round  $t$ , because  $\theta = \rho(F_\theta) \leq \rho(F_t) = \rho(F_t, x_t) \leq \frac{|F_t(x_t) \cup F_t(\langle x_t, y_t \rangle)|}{|F_t|}$  by the definition of  $\theta$ ,  $F_\theta$ , and  $x_t$ . Thus we have  $(1 - \theta)^{Q_G} |\mathcal{F}| \geq 1$ . Taking the logarithm of both sides of this inequality, applying  $\log(1 - x) \leq -x$  for  $x < 1$ , rearranging, and noting that  $\text{HD}(\mathcal{F}) \triangleq \frac{1}{\theta}$  by definition implies the theorem. ■

## 7. MAX Query Complexity Lower Bound

In this section we prove that that haystack dimension in fact provides a lower bound on the query complexity of *any* algorithm finding a maximum of  $f^* \in \mathcal{F}$ :

**Theorem 3** *Let  $A$  be any algorithm for the MAX problem for  $\mathcal{F}$ . Then  $Q_A \geq \frac{1}{6} \text{HD}(\mathcal{F})$ .*

The proof of Theorem 3 is somewhat involved and developed in a series of lemmas, so we shall first sketch the broad intuition. The lower bound will draw the target  $f^*$  uniformly at random from  $F_\theta$ , where  $F_\theta$  is as in the definition of haystack dimension, and is the subset of functions on which the greedy algorithm guarantees the least amount of progress (in terms of reducing its attention space, as defined above). By definition no other algorithm  $A$  can make more progress than  $\theta$  on its first query starting from  $F_\theta$  (since this is the maximum possible, and obtained by greedy). After the first step, the greedy algorithm and algorithm  $A$  may have different attention spaces, and thus on subsequent steps  $A$  may make greater progress than greedy; but  $A$  cannot make “too much” progress on (say) its second step, since otherwise its query there would have made more than  $\theta$  progress on  $F_\theta$ . This insight leads to a formal recurrence inequality governing the progress rate of  $A$ , whose solution, when combined with a Bayesian argument, leads to a lower bound that is  $\Omega(\frac{1}{\theta}) \triangleq \Omega(\text{HD}(\mathcal{F}))$ . We now proceed with the formal development.

Suppose we have an algorithm  $A$  that, given access to query-value pairs from  $f^*$ , generates a sequence of queries  $\{x_t\}$  (possibly depending on its random bits). Let  $y_t = f^*(x_t)$ . Let  $H_1 \triangleq F_\theta$  and

$$\begin{aligned} y_t^-(x) &\triangleq \arg \inf_{y \in \mathbb{R}} |H_t(x) \cup H_t(\langle x, y \rangle)| & S_t(x) &\triangleq H_t(x) \cup H_t(\langle x, y_t^-(x) \rangle) \\ H_{t+1} &\triangleq H_t \setminus S_t(x_t) & x_t^* &\triangleq \arg \sup_{x \in \mathcal{X}} |S_t(x)| & \delta_t &\triangleq \frac{|S_t(x_t^*)|}{|H_t|} \end{aligned}$$

These definitions are closely related to those for the greedy algorithm and the haystack dimension.  $y_t^-(x)$  is the “least helpful” possible output value on query  $x$ ,  $H_t$  is the attention space of algorithm  $A$  when starting from  $F_\theta$  if only least helpful outputs were returned,  $S_t(x)$  is the set of functions in  $H_t$  that either attain a maxima at  $x$  or would have behavior inconsistent with the observation  $y_t^-(x)$  on input  $x$ , and  $\delta_t$  is the progress (fractional reduction of the attention space) made by the greedy algorithm.

Our first lemmas, which codify the aforementioned Bayesian argument, show that when  $f^*$  is drawn uniformly from  $F_\theta$ , the probability a deterministic algorithm (a restriction removed shortly) finds a maximum in fewer than  $T$  steps is bounded by the sum of the  $\delta_t$ .

**Lemma 4** Fix a sequence  $\{x_t\}$ . Let  $B_t = \{f \in F_\theta : x_s \notin X^f \wedge y_s^-(x_s) = f(x_s) \forall s < t\}$ . Then  $H_t = B_t$  for all  $t$ .

**Proof** When  $t = 1$ ,  $B_1 = \{f \in F_\theta\}$  and the claim is immediate. Otherwise, assume that  $H_t = B_t$  for some  $t \geq 1$ . We have

$$\begin{aligned} B_{t+1} &= B_t \cap \{f \in F_\theta : x_t \notin X^f \wedge y_t^-(x_t) = f(x_t)\} \\ &= B_t \setminus \{f \in B_t : x_t \in X^f \vee y_t^-(x_t) \neq f(x_t)\} = H_t \setminus S_t(x_t) = H_{t+1}. \end{aligned} \quad \blacksquare$$

**Lemma 5**  $\Pr_{f^* \sim U_{F_\theta}} [T^{A,f^*} < T] \leq \sum_{t=1}^T \delta_t$  for any deterministic algorithm  $A$  and constant  $T$ .

**Proof** Since  $A$  is deterministic, the sequence  $\{x_t\}$  is determined by the choice of  $f^* \in \mathcal{F}$ . By Lemma 4 we have  $H_t = \{f \in F_\theta : x_s \notin X^f \wedge y_s^-(x_s) = f(x_s) \forall s < t\}$  for all  $t$ . Moreover, since  $y_s^-(x_s) = f^*(x_s)$  for all  $s < t$  and  $f^* \in H_t$  and algorithm  $A$  is deterministic, the sequence  $\{x_s\}_{s \leq t}$  is identical for any choice of  $f^* \in H_t$ . Let  $U_F$  denote the uniform distribution over  $F$ . We have

$$\begin{aligned} &\Pr_{f^* \sim U_{F_\theta}} [T^{A,f^*} < T] \\ &= \Pr_{f^* \sim U_{F_\theta}} [\exists t < T : x_t \in X^{f^*}] \\ &\leq \Pr_{f^* \sim U_{F_\theta}} [\exists t < T : x_t \in X^{f^*} \vee y_t^-(x_t) \neq f^*(x_t)] \\ &\leq \sum_{t=1}^T \Pr_{f^* \sim U_{F_\theta}} [x_t \in X^{f^*} \vee y_t^-(x_t) \neq f^*(x_t) \mid x_s \notin X^{f^*} \wedge y_s^-(x_s) = f^*(x_s) \forall s < t] \\ &= \sum_{t=1}^T \Pr_{f^* \sim U_{H_t}} [x_t \in X^{f^*} \vee y_t^-(x_t) \neq f^*(x_t)] \leq \sum_{t=1}^T \frac{|S_t(x_t)|}{|H_t|} \leq \sum_{t=1}^T \delta_t. \end{aligned} \quad \blacksquare$$

We thus see that one approach to lower bounding  $T^{A,f^*}$  is to bound the growth rate of the sequence  $\{\delta_t\}$ . Intuitively, we should expect that  $(1 - \delta_t)\delta_{t+1} \leq \delta_t$ . To see why, recall that  $\delta_t$  is the most progress that can be guaranteed by a single query if the function is drawn from the space  $H_t$ . This is the query that would be selected by the greedy algorithm if it were run on  $H_t$ . Suppose that it were instead that case that  $(1 - \delta_t)\delta_{t+1} > \delta_t$ . By playing  $x_{t+1}^*$  on  $H_t$  at least  $(1 - \delta_t)\delta_{t+1}$  fraction of the functions in  $H_t$  would either attain a maximum point at  $x_{t+1}^*$  or be eliminated as inconsistent with the observed value  $f^*(x_{t+1}^*)$ . Thus the query  $x_t^*$ , which only guaranteed that a  $\delta_t$  fraction of the functions in  $H_t$  have this property, was suboptimal, a contradiction. More formally:

**Lemma 6**  $(1 - \delta_t)\delta_{t+1} \leq \delta_t$  for all  $t$ .

**Proof**

$$\begin{aligned} (1 - \delta_t)\delta_{t+1} &= \left(1 - \frac{|S_t(x_t^*)|}{|H_t|}\right) \frac{|S_{t+1}(x_{t+1}^*)|}{|H_{t+1}|} = \left(\frac{|H_t| - |S_t(x_t^*)|}{|H_t|}\right) \frac{|S_{t+1}(x_{t+1}^*)|}{|H_t| - |S_t(x_t^*)|} \\ &\leq \frac{|S_{t+1}(x_{t+1}^*)|}{|H_t|} \leq \frac{|S_t(x_{t+1}^*)|}{|H_t|} \leq \frac{|S_t(x_t^*)|}{|H_t|} = \delta_t. \end{aligned}$$

Here the first inequality follows from the definition of  $x_t^*$  as a maximizer of  $|S_t(x)|$  (and thus  $(|H_t| - |S_t(x_t^*)|)/(|H_t| - |S_t(x_t)|) \leq 1$ ). The second inequality follows because  $|S_{t+1}(x)| \leq |S_t(x)|$  for all  $x \in \mathcal{X}$ , and the final inequality follows once again from the fact that  $x_t^*$  maximizes  $|S_t(x)|$ .  $\blacksquare$

We next establish that, roughly speaking,  $\delta_t$  must remain  $O(\delta_1)$  for  $\Omega(1/\delta_1)$  steps. More precisely:

**Lemma 7** *If  $t < \frac{1}{\delta_1}$  then  $\delta_t \leq \frac{\delta_1}{1-t\delta_1}$ .*

**Proof** The base case  $t = 1$  clearly holds. Now suppose for induction that  $\delta_t \leq \frac{\delta_1}{1-t\delta_1}$ . We have

$$\begin{aligned}\delta_{t+1} &\leq \frac{\delta_t}{1-\delta_t} \leq \frac{\delta_1}{1-t\delta_1} \left( \frac{1}{1-\delta_t} \right) = \frac{\delta_1}{1-t\delta_1 - \delta_t + t\delta_1\delta_t} \\ &= \frac{\delta_1}{1-(t+1)\delta_1 + \delta_1 - \delta_t + t\delta_1\delta_t} \leq \frac{\delta_1}{1-(t+1)\delta_1}\end{aligned}$$

The first inequality holds by Lemma 6, and the second inequality holds by the induction hypothesis. For the final inequality, note that  $\delta_t \leq \frac{\delta_1}{1-t\delta_1}$  implies that  $\delta_1 - \delta_t + t\delta_1\delta_t \geq 0$ , as long as  $t < \frac{1}{\delta_1}$ .  $\blacksquare$

We are now ready to prove the lower bound of Theorem 3.

**Proof** (Theorem 3) The behavior of algorithm  $A$  is partly determined by its internal randomization, which we denote as a random string  $\omega$  drawn from a distribution  $\mathcal{P}$ . Let us write  $A(\omega)$  for the deterministic algorithm corresponding to the string  $\omega$ .

Fix a positive constant  $c < 1/2$  (implying  $\frac{c}{1-c} < 1$ ), with the exact value to be specified later. For any fixed  $\omega$

$$\Pr_{f^* \sim U_{F_\theta}} \left[ T^{A(\omega), f^*} < \frac{c}{\theta} \right] \leq \sum_{t=1}^{c/\theta} \delta_t \leq \sum_{t=1}^{c/\theta} \frac{\delta_1}{1-t\delta_1} \leq \sum_{t=1}^{c/\theta} \frac{\theta}{1-c} = \frac{c}{1-c} \quad (1)$$

where we used, in order: Lemma 5; Lemma 7 and the fact that  $\delta_1 = \theta$  and  $c < 1$ ; the fact that  $\delta_1 = \theta$  again; arithmetic. Now we have

$$E_{f^* \sim U_{F_\theta}} \left[ T^{A(\omega), f^*} \right] \geq \left( 1 - \Pr_{f^* \sim U_{F_\theta}} \left[ T^{A(\omega), f^*} < \frac{c}{\theta} \right] \right) \frac{c}{\theta} \geq \left( 1 - \frac{c}{1-c} \right) \frac{c}{\theta} \quad (2)$$

where the second inequality follows from (1). Finally, we have

$$\begin{aligned}E_{f^* \sim U_{F_\theta}} \left[ T^{A, f^*} \right] &\triangleq E_{\omega \sim \mathcal{P}, f^* \sim U_{F_\theta}} \left[ T^{A(\omega), f^*} \right] \\ &= E_{\omega \sim \mathcal{P}} \left[ E_{f^* \sim U_{F_\theta}} \left[ T^{A(\omega), f^*} | \omega \right] \right] \geq E_{\omega \sim \mathcal{P}} \left[ \left( 1 - \frac{c}{1-c} \right) \frac{c}{\theta} \right] = \left( 1 - \frac{c}{1-c} \right) \frac{c}{\theta}\end{aligned}$$

where the inequality follows from (2). The choice of  $c$  implies  $\left( 1 - \frac{c}{1-c} \right) c > 0$ . Letting  $c = 2 - \sqrt{3}$  and recalling the definition  $\text{HD}(\mathcal{F}) \triangleq \frac{1}{\theta}$  implies the theorem.  $\blacksquare$

## 8. Relationship to VC Dimension and Extended Teaching Dimension

As we have demonstrated, the haystack dimension provides a lower bound on the query complexity of any algorithm for the MAX problem on a function class  $\mathcal{F}$ . This is a role analogous to the VC dimension in the PAC learning model. However, as we will demonstrate, the two are incomparable in general. We will also illustrate the haystack dimension's relationship to the *extended teaching dimension* (Hegedüs, 1995). The extended teaching dimension characterizes the number of queries required to learn  $f^* \in \mathcal{F}$  when  $\mathcal{F}$  consists of binary functions. Clearly learning  $f^*$  is sufficient for maximization and, as we will see, the haystack dimension can be much smaller than extended teaching dimension, but cannot be too much larger. Note that the VC and extended teaching dimensions are defined only for binary functions, whereas the haystack dimension and our results encompass a much more general setting.

For  $\mathcal{F}$  consisting of binary functions, we will denote the VC dimension as  $VCD(\mathcal{F})$ , where the hypothesis class is assumed to equal to the concept class. Similarly, we will denote the extended teaching dimension by  $XTD(\mathcal{F})$ . Both are defined below.

**Definition 8 (Kearns and Vazirani (1994))** *Let  $\mathcal{F}$  be a function class (concept class) where  $f : \mathcal{X} \rightarrow \{0, 1\}$  for each  $f \in \mathcal{F}$ . We say a set  $S = \{x_1, \dots, x_m\} \subseteq \mathcal{X}$  is shattered by  $\mathcal{F}$  if  $\{(f(x_1), \dots, f(x_m)) : f \in \mathcal{F}\} = \{0, 1\}^m$ .  $VCD(\mathcal{F})$  is equal to the cardinality of the largest set shattered by  $\mathcal{F}$ .*

**Definition 9 (Hegedüs (1995))** *Let  $h : \mathcal{X} \rightarrow \{0, 1\}$ . We say that  $S \subseteq \mathcal{X}$  is a specifying set for  $h$  if there is at most one  $f \in \mathcal{F}$  consistent with  $h$  on all  $x \in S$ . Let  $XTD(\mathcal{F}, h)$  be equal to the size of the smallest specifying set for  $h$ .  $XTD(\mathcal{F}) = \sup_h XTD(\mathcal{F}, h)$ .*

**Example 5** (a) For any  $d$ , there exists a function class and set of inputs  $(\mathcal{F}, \mathcal{X})$  such that  $VCD(\mathcal{F}) = d$  but  $HD(\mathcal{F}) = 1$ . (b) For any  $d$ , there exists a  $(\mathcal{F}, \mathcal{X})$  such that  $VCD(\mathcal{F}) = 2$  but  $HD(\mathcal{F}) = d$ .

**Proof** To prove (a), let  $(\mathcal{F}_d, \mathcal{X}_d)$  be any function class, and set of inputs, such that  $VCD(\mathcal{F}_d) = d$ . Let  $\mathcal{X} = \mathcal{X}_d \cup \{x^*\}$ . Construct  $(\mathcal{F}, \mathcal{X})$  by including a function  $f$  in  $\mathcal{F}$  for each  $f' \in \mathcal{F}_d$ , that is identical to  $f'$  on all inputs in  $\mathcal{X}_d$ . Also let  $f(x^*) = 1$ . For any  $F \subseteq \mathcal{F}$ ,  $x^*$  maximizes all functions in  $F$ . Therefore,  $HD(\mathcal{F}) = 1$ . However, no function in  $\mathcal{F}$  gives  $x^*$  the label 0, and so the size of the largest shattered set does not change, and  $VCD(\mathcal{F}) = d$ .

To prove (b), consider  $\mathcal{F}$  to be the “needle in the haystack” of Example 1, where  $|\mathcal{F}| = d$ . We have that  $HD(\mathcal{F}) = d$ . To compute the VC dimension, note that no function in  $\mathcal{F}$  labels more than one input with a 1, and so no set of three inputs can be shattered by  $\mathcal{F}$ . (It’s also obvious that any  $\{x_1, x_2\} \subseteq \mathcal{X}$  can be shattered). ■

**Example 6** (a) For any  $d$ , there exists a function class and set of inputs  $(\mathcal{F}, \mathcal{X})$  such that  $XTD(\mathcal{F}) = d$  but  $HD(\mathcal{F}) = 1$ . (b) For any  $\mathcal{F}$  consisting of binary functions,  $HD(\mathcal{F}) = O(HTD(\mathcal{F}) \log |\mathcal{F}|)$ .

**Proof** To prove (a), let  $(\mathcal{F}_d, \mathcal{X}_d)$  be a function class, and set of inputs, such that  $XTD(\mathcal{F}_d) = d$ . Construct  $(\mathcal{F}, \mathcal{X})$  exactly as in the proof of Example 5(a). For any  $h : \mathcal{X} \rightarrow \{0, 1\}$  with  $h(x^*) = 1$ , let  $h' : \mathcal{X} \rightarrow \{0, 1\}$  be a function identical to  $h$  on  $\mathcal{X}_d$ . It's clear that  $XTD(\mathcal{F}, h) = XTD(\mathcal{F}_d, h')$ . For any  $h : \mathcal{X} \rightarrow \{0, 1\}$  with  $h(x^*) = 0$ ,  $XTD(\mathcal{F}, h) = 1$ , since  $\{x^*\}$  is a specifying set for  $h$ . Thus,  $XTD(\mathcal{F}) = XTD(\mathcal{F}_d) = d$ .

For (b), Hegedüs (1995) gives an algorithm which learns  $f^* \in \mathcal{F}$  using  $O(XTD(\mathcal{F}) \log |\mathcal{F}|)$  membership queries. Since learning  $f^*$  is sufficient for maximization, Theorem 3 implies (b).  $\blacksquare$

## 9. Functional MAB Regret Upper Bound

In this and the next section, we return to the functional MAB problem, showing a close relationship to the functional MAX problem, and giving upper and lower bounds on regret that are order the haystack dimension  $HD(\mathcal{F})$ . We will describe a no-regret algorithm for functional MAB problems where  $\mathcal{F}$  is known, finite, but otherwise arbitrary. Our approach is to use the greedy functional MAX algorithm of Section 6 to find an action/query that maximizes  $f^*$ , and then select that action indefinitely. There is a technical complication we must overcome when implementing this approach: Each action returns a sample from a distribution, rather than a single value. We solve this by repeatedly selecting an action  $x$  to get an accurate estimate of  $f^*(x)$ , and also by restarting the algorithm after progressively longer phases (a standard “trick” in MAB algorithms).

Before giving a more detailed description of our algorithm and its analysis, we need some additional definitions. Let  $\epsilon_{\min} \triangleq \min_{f, f' \in \mathcal{F}} \inf_{x: f(x) \neq f'(x)} |f(x) - f'(x)|$  be the smallest difference between any two functions in  $\mathcal{F}$  on those points where they do differ;  $\epsilon_{\min} > 0$  allows us to determine  $f^*(x)$  by selecting  $x$  enough times.

Also, for any set  $F \subseteq \mathcal{F}$ , let us define the  $\epsilon$ -inconsistent set to be  $F(\langle x, y \rangle, \epsilon) \triangleq \{f \in F : |f(x) - y| > \epsilon\}$ . Finally, recall that  $f^*(x) \in [-b, b]$ .

The *greedy bandit algorithm*  $GB$  proceeds in phases  $i = 1, 2, \dots$ , where phase  $i$  lasts  $T_i = 2^{2^i}$  rounds, and consists of two consecutive subphases, which we will denote by  $i_{\text{explore}}$  and  $i_{\text{exploit}}$ . In subphase  $i_{\text{explore}}$ , we run a fresh instance of the greedy query algorithm  $G$  from Section 6, with some minor modifications. When the query algorithm  $G$  selects a query  $x_t^i$ , the bandit algorithm  $GB$  selects that action for  $L = \frac{32b^2}{\epsilon_{\min}^2} (\log HD(\mathcal{F}) + \log \log |\mathcal{F}|) \log T_i$  consecutive rounds, and lets  $\bar{y}_t^i$  be the average of the observations. The attention space  $F_t^i$  is then updated according to the rule  $F_{t+1}^i = F_t^i \setminus (F_t^i(x_t^i) \cup F_t^i(\langle x_t^i, \bar{y}_t^i \rangle, \epsilon_{\min}))$ . Let  $\tau_i$  be the number of queries required to empty the attention space in subphase  $i_{\text{explore}}$  (so subphase  $i_{\text{explore}}$  lasts  $\tau_i L$  rounds). In subphase  $i_{\text{exploit}}$ , in each round the bandit algorithm  $GB$  selects the action  $x_t^i$  associated with the largest  $\bar{y}_t^i$  in subphase  $i_{\text{explore}}$ .

**Theorem 10** *Let  $GB$  be the greedy bandit algorithm for the MAB problem for  $\mathcal{F}$ . Then*

$$R_{GB}(T) = O \left( b^2 \frac{HD(\mathcal{F}) \log |\mathcal{F}|}{\epsilon_{\min}^2} \left( \log HD(\mathcal{F}) + \log \log |\mathcal{F}| \right) \log T + \log \log T \right).$$

**Proof** Let  $R_i$  be the regret realized by the greedy bandit algorithm during phase  $i$ . We will first bound  $E[R_i]$  for any phase  $i$ , and then use a ‘squaring trick’ to tightly bound  $E[\sum_i R_i] \triangleq R_{GB}(T)$ .

By a similar argument as in Theorem 2, for each phase  $i$  we have  $\tau_i \leq \text{HD}(\mathcal{F}) \log |\mathcal{F}|$ . For each  $t \in [\tau_i]$ , action  $x_t^i$  is selected  $L$  times, which is a sufficient number of times to ensure (by the Chernoff and union bounds) that with probability  $1 - O(\frac{1}{T_i})$  we have  $|\bar{y}_t^i - f^*(x_t^i)| \leq \frac{\epsilon_{\min}}{2}$  for all  $t \in [\tau_i]$ . Let  $\text{good}(i)$  be this event. We have  $E[R_i] = E[R_i | \text{good}(i)] \Pr[\text{good}(i)] + E[R_i | \neg \text{good}(i)] \Pr[\neg \text{good}(i)] \leq E[R_i | \text{good}(i)] + O(1)$ , because  $R_i \leq T_i$  and  $\Pr[\neg \text{good}(i)]$  is  $O(\frac{1}{T_i})$ .

It therefore remains to bound  $E[R_i | \text{good}(i)]$ . By the definition of  $\epsilon_{\min}$ , if the event  $\text{good}(i)$  occurs then the attention space  $F_t^i$  maintained during subphase  $i_{\text{explore}}$  is updated exactly the same way as in the greedy query algorithm  $G$  from Section 6. Therefore, an action  $x^* \in X^{f^*}$  is selected in subphase  $i_{\text{explore}}$  before it ends (by Theorem 2), and thus, again by the definition of  $\epsilon_{\min}$ , the action  $x^*$  is selected in every round of subphase  $i_{\text{exploit}}$ . Since  $\tau_i \leq \text{HD}(\mathcal{F}) \log |\mathcal{F}|$  and subphase  $i_{\text{explore}}$  lasts  $\tau_i L$  rounds, we have  $E[R_i | \text{good}(i)] = O\left(\frac{b^2 \text{HD}(\mathcal{F}) \log |\mathcal{F}|}{\epsilon_{\min}^2} (\log \text{HD}(\mathcal{F}) + \log \log |\mathcal{F}|) \log T_i\right)$ .

Finally, we apply a ‘squaring trick’.<sup>5</sup> Let  $K$  be the number of phases. Since  $T_i = 2^{2^i}$ , we have  $K = O(\log \log T)$  and  $\sum_{i=1}^K \log T_i = \sum_{i=1}^{O(\log \log T)} 2^i = O(\log T)$ . Recalling that  $R_{GB}(T) \triangleq E[\sum_i R_i]$ , proves the theorem. ■

Note that the greedy bandit algorithm  $GB$  assumes that the value of the haystack dimension  $\text{HD}(\mathcal{F})$  is known. It is possible to modify the algorithm in case this information is not available: Rather than selecting each action  $x_t^i$  in subphase  $i_{\text{explore}}$  for  $L = \frac{32b^2}{\epsilon_{\min}^2} (\log \text{HD}(\mathcal{F}) + \log \log |\mathcal{F}|) \log T_i$  consecutive rounds, we instead select it for  $L = \frac{32b^2}{\epsilon_{\min}^2} (\log |\mathcal{F}| + \log \log |\mathcal{F}|) \log T_i$  consecutive rounds. Since  $\text{HD}(\mathcal{F}) \leq |\mathcal{F}|$  trivially holds, the analysis in the proof of Theorem 10 is essentially unaffected by this modification, and it adds only a  $O((\log |\mathcal{F}|)^2)$  term to the regret upper bound in Theorem 10.

## 10. Functional MAB Regret Lower Bound

In this section, we prove that the greedy bandit algorithm in Section 9 is near-optimal, at least with respect to the haystack dimension (our primary interest) and terms  $\log |\mathcal{F}|$  or smaller. With respect to the dependence on the haystack dimension, we can say something quite strong: Every MAB algorithm for every function class must suffer regret that is linear in the haystack dimension of the class. Let  $\Delta = \inf_{f \in \mathcal{F}} \inf_{\{x' \in \mathcal{X}: f(x') < \sup_x f(x)\}} \sup_x f(x) - f(x)$ , be the difference between the best action and second-best action in  $\mathcal{X}$ .

**Theorem 11** *For any function class  $\mathcal{F}$  and MAB algorithm  $A$  for  $\mathcal{F}$  we have  $R_A(T) = \Omega(\Delta \min\{T, \text{HD}(\mathcal{F})\})$ .*

The proof of Theorem 11 follows directly from Theorem 3, which implies that no MAB algorithm for function class  $\mathcal{F}$  can select the best action in fewer than  $\Omega(\text{HD}(\mathcal{F}))$  steps.

---

5. If we were to instead apply the more common ‘doubling trick’, such that  $T_i = 2^i$ , the upper bound in Theorem 10 would be  $O((\log T)^2)$ .

The next theorem proves that the terms  $\log |\mathcal{F}|$  and  $\frac{1}{\epsilon_{\min}}$  cannot be removed from the upper bound in Theorem 10.

**Theorem 12** (a) *There exists a function class  $\mathcal{F}$  such that  $\text{HD}(\mathcal{F}) = \Theta(1)$  and for any MAB algorithm  $A$  for  $\mathcal{F}$ ,  $R_A(T) = \Omega(\log |\mathcal{F}|)$ .* (b) *There exists a function class  $\mathcal{F}$  such that  $\text{HD}(\mathcal{F}) = \Theta(1)$  and for any MAB algorithm  $A$  for  $\mathcal{F}$ ,  $R_A(T) = \Omega\left(\min\left\{\frac{1}{\epsilon_{\min}}, |\mathcal{F}|\right\}\right)$ .*

### Proof

To prove (a), we will outline the construction of  $\mathcal{F}$  and omit the details of the proof, which are straightforward. The input space  $\mathcal{X}$  will have two components —  $\mathcal{X}_1$  and  $\mathcal{X}_2$ , with  $\mathcal{X}$  being the union of these disjoint domains. Subdomain  $\mathcal{X}_1$  consists of  $2^n$  points, and  $\mathcal{X}_2$  of  $n$  points, while the function class  $\mathcal{F}$  contains  $n$  deterministic functions. Each point in  $x \in \mathcal{X}_1$  corresponds to a distinct subset  $F_x \subseteq \mathcal{F}$ , and for each  $x \in \mathcal{X}_1$  let  $f(x) = \frac{1}{3}$  for half the functions in  $F_x$  and  $f(x) = \frac{2}{3}$  for the other half, and also let  $f(x) = \frac{1}{3}$  for all  $f \in \mathcal{F} \setminus F_x$ . Note that  $\mathcal{X}_1$  contains excellent information queries, because for any subset  $F \subseteq \mathcal{F}$  there is a point in  $\mathcal{X}_1$  that eliminates at least half the functions in  $F$  when that point is issued as a query, and thus  $\text{HD}(\mathcal{F}) = 2$ . Finally, map each function  $f \in \mathcal{F}$  to a distinct point  $x_f \in \mathcal{X}_2$ , and let  $f(x_f) = 1$  and  $f(x) = 0$  for all  $x \in \mathcal{X}_2 \setminus \{x_f\}$ . So each  $f \in \mathcal{F}$  has a unique maximum, contained in  $\mathcal{X}_2$ .

Suppose  $f^*$  is chosen uniformly at random from  $\mathcal{F}$ . It is clear that  $Q_A = \Omega(\log |\mathcal{F}|)$ . To see why, note that a query  $x \in \mathcal{X}_2$  has only probability  $\frac{1}{|\mathcal{F}|}$  of being the function's max, and only serves to inform the algorithm that  $f^* \neq f$  whenever  $x = x_f$  and  $f^*(x) = 0$ . Thus any algorithm is forced to learn  $f^*$  by playing actions in  $\mathcal{X}_1$ , requiring that at least  $\log_2 |\mathcal{F}|$  actions are played.

To prove (b), we simply modify the construction of  $\mathcal{F}$  to add stochasticity, as follows. For an  $x \in \mathcal{X}_1$ , if we had previously that  $f(x) = \frac{1}{3}$ , we now let  $f(x) = \epsilon$ . If we had previously that  $f(x) = \frac{2}{3}$ , we now let  $f(x) = 2\epsilon$ . Each function's behavior of  $\mathcal{X}_2$  is unchanged. Note that  $\epsilon_{\min} = \epsilon$ . When playing an  $x \in \mathcal{X}$ , rather than observe the output of the function, the algorithm now observes the outcome a Bernoulli random variable with mean  $f^*(x)$ . It can then be shown that any algorithm that wishes to make use of the information in  $\mathcal{X}_1$  must sample the actions there  $\Omega(1/\epsilon)$  times or else be forced to play actions in  $\mathcal{X}_2$ . ■

Theorem 12(b) establishes that there is at least one function class for which the haystack dimension does not provide an adequate lower bound in the MAB setting. In other words, there is a real gap between the upper and lower bounds in Theorems 10 and 11 that depends on  $\frac{1}{\epsilon_{\min}}$ . We suspect that a modified version of the haystack dimension which better accounts for the “information value” of a query under stochasticity (i.e. that uses not just the size of the largest inconsistent set, but accounts for the variance of the functions in  $\mathcal{F}$ ) would close this gap, but leave this as an open problem.

## 11. Infinite Function Classes

In the remainder of the paper, we return to studying the MAX problem. We give extensions of our basic results on query complexity, and also give examples that illustrate some aspects of our analysis. In this section, we describe how to extend our methods from Sections 6

and 7, which were restricted to finite function classes, to infinite function classes that have a *finite covering oracle*.

**Definition 13** We say  $C$  is a *finite covering oracle* for  $\mathcal{F}$  if for any  $\epsilon > 0$  the finite set  $C(\epsilon) \subseteq \mathcal{F}$  has the following property: For any  $f \in \mathcal{F}$  there exists an  $f' \in C(\epsilon)$  such that  $\sup_{x \in \mathcal{X}} |f(x) - f'(x)| \leq \epsilon$ .

Fix a possibly infinite function class  $\mathcal{F}$  with finite covering oracle  $C$ . We consider the  $\epsilon$ -MAX problem for  $\mathcal{F}$ , a relaxed version of the MAX problem. In analogy to the MAX problem, for any algorithm  $A$  let  $T^{A,f^*,\epsilon} \triangleq \min\{t : \sup_x f^*(x) - f^*(x_t) \leq \epsilon\}$  be the first round  $t$  such that the query  $x_t$  selected by  $A$  is an  $\epsilon$ -maximum of  $f^* \in \mathcal{F}$ . We are interested in bounding the *worst-case  $\epsilon$ -query complexity* of  $A$ , defined  $Q_{A,\epsilon} \triangleq \sup_{f^* \in \mathcal{F}} E[T^{A,f^*,\epsilon}]$ .

Below we give upper and lower bounds for the  $\epsilon$ -MAX problem in terms of the *lower and upper  $\epsilon$ -haystack dimensions*, which are each a different generalization of the haystack dimension. Before introducing these quantities, we need some definitions. For any set  $F \subseteq \mathcal{F}$  let  $F(x, \epsilon) \triangleq \{f \in F : \sup_{x'} f(x') - f(x) \leq \epsilon\}$  be the functions in  $F$  for which  $x$  is an  $\epsilon$ -maximum and, as in Section 9, let  $F(\langle x, y \rangle, \epsilon) \triangleq \{f \in F : |f(x) - y| > \epsilon\}$  be the functions in  $F$  that are more than  $\epsilon$ -inconsistent with the labeled example  $\langle x, y \rangle$ . Also, let

$$\begin{aligned}\theta^-(\epsilon) &\triangleq \inf_{F \subseteq C(\epsilon)} \sup_{x \in \mathcal{X}} \inf_{y \in \mathbb{R}} \frac{|F(x, \epsilon) \cup F(\langle x, y \rangle, 0)|}{|F|}, \text{ and} \\ \theta^+(\epsilon) &\triangleq \inf_{F \subseteq C(\epsilon)} \sup_{x \in \mathcal{X}} \inf_{y \in \mathbb{R}} \frac{|F(x, 0) \cup F(\langle x, y \rangle, \epsilon)|}{|F|}.\end{aligned}$$

Note that the only difference between  $\theta^-(\epsilon)$  and  $\theta^+(\epsilon)$  is the placement of  $\epsilon$  and 0. Also note that if  $\mathcal{F}$  is finite and  $C(\epsilon) = \mathcal{F}$  then  $\theta^-(0) = \theta^+(0) = \theta$ , where  $\theta$  was defined in Section 4. Now define the *lower  $\epsilon$ -haystack dimension*  $\text{HD}^-(\epsilon) \triangleq \frac{1}{\theta^-(\epsilon)}$  and the *upper  $\epsilon$ -haystack dimension*  $\text{HD}^+(\epsilon) \triangleq \frac{1}{\theta^+(\epsilon)}$ .

A simple approach to solving the  $\epsilon$ -MAX problem is just to run a slightly modified version of the greedy algorithm from Section 6 using  $C(\epsilon)$  as the initial attention space, and removing inconsistent functions only if they are inconsistent by more than  $\epsilon$ . In other words, in each round  $t$ , the  *$\epsilon$ -greedy algorithm*  $G_\epsilon$  selects  $x_t = \arg \sup_{x \in \mathcal{X}} \inf_{y \in \mathbb{R}} |F_t(x, 0) \cup F_t(\langle x, y \rangle, \epsilon)|$  where the *attention space*  $F_t$  is defined inductively as follows:  $F_1 = C(\epsilon)$  and  $F_{t+1} = F_t \setminus (F_t(x_t, 0) \cup F_t(\langle x_t, y_t \rangle, \epsilon))$ .

**Theorem 14** Let  $G_\epsilon$  be the  $\epsilon$ -greedy algorithm for the  $\epsilon$ -MAX problem. Then  $Q_{G_\epsilon, 2\epsilon} \leq \text{HD}^+(\epsilon) \log |C(\epsilon)|$  for all  $\epsilon > 0$ .

**Proof** The proof is nearly identical to the proof of Theorem 2. The algorithm  $G_\epsilon$  initializes the attention space to  $C(\epsilon)$ , and after every query at least a  $\theta^+(\epsilon)$  fraction of the attention space is eliminated. By the time the attention space is empty, the  $G_\epsilon$  algorithm has selected a maximum of some function  $\hat{f} \in C(\epsilon)$  that  $\epsilon$ -covers the true function  $f^*$ , which implies that a  $2\epsilon$ -maximum of  $f^*$  has been selected. ■

Furthermore, we can also straightforwardly lower bound the query complexity of any algorithm for the  $\epsilon$ -MAX problem.

**Theorem 15** Let  $A$  be any algorithm for the  $\epsilon$ -MAX problem. Then  $Q_{A,\epsilon} \geq \frac{1}{6} \text{HD}^-(\epsilon)$  for all  $\epsilon > 0$ .

**Proof** The proof of Theorem 3 can be repeated, with essentially no changes, to prove this theorem as well. The key is to observe that, when proving Theorem 3, we made no use of the fact that  $X^{f^*}$  contained the maxima of the true function  $f^*$ . We could have defined  $X^{f^*}$  to be any subset of  $\mathcal{X}$ , including the  $\epsilon$ -maxima of  $f^*$ . ■

Notice that the upper and lower bounds in Theorems 14 and 15 are not directly comparable, since we have not related the quantities  $\text{HD}^-(\epsilon)$  and  $\text{HD}^+(\epsilon)$ . If it happens that  $\frac{\text{HD}^+(\epsilon)}{\text{HD}^-(\epsilon)} \leq K$  for some constant  $K$ , then clearly the upper and lower bounds above are within constant and logarithmic factors of each other, just as we had for finite function classes. Indeed, a simple infinite function class for which this occurs is the class of all bounded norm hyperplanes. Let  $\mathcal{X} = \{x \in \mathbb{R}^n \mid \|x\|_\infty \leq 1\}$  and  $\mathcal{F}_{\text{hyper}} = \{\langle w, \cdot \rangle \mid w \in \mathbb{R}^n, \|w\|_\infty \leq 1\}$ , and let the finite covering oracle  $C$  be the appropriate discretization of  $\mathcal{F}_{\text{hyper}}$ , i.e.,  $C(\epsilon) = \{w \in \mathbb{R}^n \mid \forall i \ w_i \in \{0, \frac{\epsilon}{n}, \frac{2\epsilon}{n}, \dots, 1\}\}$ . Clearly  $|C(\epsilon)| = \Theta((\frac{n}{\epsilon})^n)$ , but nonetheless the ratio of the lower and upper  $\epsilon$ -haystack dimension is not large.

**Theorem 16** For function class  $\mathcal{F}_{\text{hyper}}$  we have  $\frac{\text{HD}^+(\epsilon)}{\text{HD}^-(\epsilon)} \leq n$  for all  $\epsilon > 0$ .

**Proof** By examining the definitions of  $\text{HD}^+(\epsilon)$  and  $\text{HD}^-(\epsilon)$ , we see that it suffices to show that  $\sup_{x \in \mathcal{X}} \frac{|F(x, 0)|}{|F|} \geq \frac{1}{n}$  for any finite  $F \subseteq \mathcal{F}$ . Let  $k_i = |\{\langle w, \cdot \rangle \in F \mid w_j \geq w_i, \forall j\}|$  be the number of functions in  $F$  which have their maximal component at  $i$ . Clearly there must exist a  $k_i \geq \frac{|F|}{n}$ . Let  $e_i$  be the vector with a one in its  $i$ th component and zeros everywhere else. Since  $e_i \in \mathcal{X}$  and  $e_i$  maximizes  $k_i$  functions in  $F$ , there exists an  $x \in \mathcal{X}$  with  $\frac{|F(x, 0)|}{|F|} \geq \frac{1}{n}$ . ■

Unfortunately, the bound in Theorem 16 cannot be generalized to all function classes with finite covering oracles. In the next theorem, we give an example of an infinite function class  $\mathcal{F}$  with a finite covering oracle  $C$  for which  $\frac{\text{HD}^+(\epsilon)}{\text{HD}^-(\epsilon)} = \Omega(|C(\epsilon)|)$ , which is essentially the worst possible ratio.

**Theorem 17** There exists a function class  $\mathcal{F}$  with a finite covering oracle  $C$  such that  $\text{HD}^-(\epsilon) = 1$  and  $\text{HD}^+(\epsilon) = \Omega(|C(\epsilon)|)$  for all  $\epsilon > 0$ .

**Proof** Let  $\mathcal{X} = [0, 1]$  and fix any sequence  $\{x_n\} \subset \mathcal{X}$ , all elements distinct. For visualization, it may be helpful to suppose that  $\{x_n\}$  is strictly increasing, but this is not necessary. Let  $\{\epsilon_i\}$  be the sequence defined by  $\epsilon_i = \frac{1}{2^i}$  for all  $i \in \mathbb{N}$ .

For each  $n \in \mathbb{N}$  let there be a function  $f_n \in \mathcal{F}$  whose values are zero everywhere except  $x_1, \dots, x_n$ . The nonzero values of  $f_n$  are defined as follows: Let  $f_n(x_n) = \frac{1}{n}$  and  $f_n(x_m) = \epsilon_i - \frac{1}{2^{2m}}$  for all  $m < n$ , where  $i = \lceil \log_2 n \rceil$ . Let  $C(\epsilon) = \{f_1, \dots, f_{N_\epsilon}\}$ , where  $N_\epsilon$  is the smallest integer such that  $1/N_\epsilon < \epsilon$ . We have that  $C$  is finite covering oracle because each  $C(\epsilon)$  is finite and because  $\sup_{x \in \mathcal{X}} f_n(x) \leq \epsilon$  for all  $n \geq N_\epsilon$ .

We only need the following properties of this construction, which can be verified: (1) For all  $n \in \mathbb{N}$  the elements of  $\{f_m(x_n) : m \geq n\}$  are all distinct; (2) Each  $f_n$  has a maximum

at  $x_n$  and nowhere else; (3) For all  $n \in \mathbb{N}$ , if  $i = \lceil \log_2 n \rceil$  then  $f_n(x_n) \geq \epsilon_i$  and  $f_m(x_n) < \epsilon_i$  for all  $m \neq n$ .

For the remainder of the proof fix  $\epsilon > 0$  and the smallest  $i \in \mathbb{N}$  such that  $\epsilon_i \leq \epsilon$ . We will show that  $\theta^-(\epsilon) = 1$  and  $\theta^+(\epsilon) \leq \frac{4}{N_{\epsilon_i}}$ , which suffices to prove the theorem.

First, we claim that  $\theta^-(\epsilon) = 1$ . Let  $F^+ = \arg \inf_{F \subseteq C(\epsilon)} \sup_{x \in \mathcal{X}} \inf_{y \in \mathbb{R}} \frac{|F(x, \epsilon) \cup F(\langle x, y \rangle, 0)|}{|F|}$ . Let  $n$  be the smallest integer such that  $f_n \in F^+$ . Then each  $f_m \in F^+$  has a distinct value at  $x_n$ , by property 1. So  $\inf_{y \in \mathbb{R}} |F^+(\langle x_n, y \rangle, 0)| = |F^+|$ . Next, we claim that  $\theta^+(\epsilon) \leq \frac{4}{N_{\epsilon_i}}$ . Let  $F_i = \{f_n \in \mathcal{F} : i = \lceil \log_2 n \rceil\}$ , and note that  $F_i \subseteq C(\epsilon)$  and  $|F_i| = \frac{N_{\epsilon_i}}{2}$ . For any  $x \in \mathcal{X}$  we have  $|F_i(x, 0)| \leq 1$ , by property 2, and  $\inf_{y \in \mathbb{R}} |F_i(\langle x, y \rangle, \epsilon)| \leq 1$ , by property 3. ■

## 12. Computational Complexity

Our results have so far ignored computational complexity. In general a function class  $\mathcal{F}$ , for which finding the optimal query is computationally intractable, might nevertheless have small haystack dimension, admitting algorithms with low query complexity. Consider the following simple example. Let  $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \mathcal{X}_3$  where  $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$  are disjoint,  $|\mathcal{X}_1| = |\mathcal{X}_2| = n$  and  $|\mathcal{X}_3| = 2^n$ . Each function in  $\mathcal{F}$  will attain its maximum on some action in  $\mathcal{X}_3$ . The location of that maximum, as in Example 2, will be encoded by  $\mathcal{X}_1$  and  $\mathcal{X}_2$ . However, we will now encode the location cryptographically, with a function's behavior on  $\mathcal{X}_1$  representing a public key, and a function's behavior on  $\mathcal{X}_2$  representing the encrypted location of its maximum.

More precisely, let  $z$  be an  $n$ -bit number. For a pair of  $\frac{n}{2}$ -bit primes  $p$  and  $q$ , let  $N_{pq} = pq$ . Also let  $e(z, N_{pq}) = z^2 \pmod{N_{pq}}$  be  $z$  encrypted with public key  $N_{pq}$ .

Now let  $f_{z,p,q}$  be the function that gives the  $i$ th bit of  $N_{pq}$  as output to the  $i$ th input of  $\mathcal{X}_1$  and the  $i$ th bit of  $e(z, N_{pq})$  as output to the  $i$ th input of  $\mathcal{X}_2$ . On the  $z$ th input of  $\mathcal{X}_3$ ,  $f_{z,p,q}$  outputs a 2. On all other inputs of  $\mathcal{X}_3$ ,  $f_{z,p,q}$  outputs 0. Let  $\mathcal{F}$  be the function class consisting of all functions  $f_{z,p,q}$  for every pair of  $\frac{n}{2}$ -bit primes  $p, q$  and  $n$ -bit integer  $z$ .

There exists an algorithm with query complexity  $O(n)$  for any  $f^* \in \mathcal{F}$ . That algorithm queries each action in  $\mathcal{X}_1 \cup \mathcal{X}_2$ , retrieving the public key  $N_{pq}$  and the cypher  $e(z, N_{pq})$ . Information-theoretically, the maximum of  $f^*$  can be found in a single additional query. The algorithm may simply test every  $n$ -bit  $z'$ , checking if  $e(z, N_{pq}) = e(z', N_{pq})$ . However, computing  $z$  is as hard as factorization (Kearns and Vazirani, 1994).

## Acknowledgments

We thank Alex Slivkins and the anonymous reviewers for their helpful comments.

## References

- Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21th Annual Conference on Computational Learning Theory*, 2008.

- Kareem Amin, Michael Kearns, and Umar Syed. Graphical models for bandit problems. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 2011.
- Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1988.
- Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. Online optimization in X-armed bandits. In *Advances in Neural Information Processing Systems 21*, 2008.
- Varsha Dani and Thomas P. Hayes. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2006.
- Abraham Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceeding of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2005.
- Tibor Hegedüs. Generalized teaching dimensions and the query complexity of learning. In *Proceedings of the 8th Annual Conference on Computational Learning Theory*, 1995.
- Michael Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.
- Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, 2008.
- John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems 20*, 2007.
- Tyler Lu, Dávid Pál, and Martin Pál. Showing relevant ads via Lipschitz context multi-armed bandits. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.
- Adam J. Mersereau, Paat Rusmevichientong, and John N. Tsitsiklis. A structured multi-armed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12):2787–2802, 2009.
- Rob Nowak. Noisy generalized binary search. In *Advances in Neural Information Processing Systems 22*, 2009.
- Aleksandrs Slivkins. Contextual bandits with similarity information. In *Proceedings of the 24th Annual Conference on Computational Learning Theory*, 2011.
- Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, 2008.
- William Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3–4):285–294, 1933.

A MIN KEARNS SYED

# Minimax Policies for Combinatorial Prediction Games

**Jean-Yves Audibert**

*Imagine, Univ. Paris Est, and Sierra,  
CNRS/ENS/INRIA, Paris, France*

AUDIBERT@IMAGINE.ENPC.FR

**Sébastien Bubeck**

*Operations Research & Financial Engineering  
Princeton University, USA*

SBUBECK@PRINCETON.EDU

**Gábor Lugosi**

*ICREA and Pompeu Fabra University  
Barcelona, Spain*

LUGOSI@UPF.ES

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We address the online linear optimization problem when the actions of the forecaster are represented by binary vectors. Our goal is to understand the magnitude of the minimax regret for the worst possible set of actions. We study the problem under three different assumptions for the feedback: full information, and the partial information models of the so-called “semi-bandit”, and “bandit” problems. We consider both  $L_\infty$ -, and  $L_2$ -type of restrictions for the losses assigned by the adversary.

We formulate a general strategy using Bregman projections on top of a potential-based gradient descent, which generalizes the ones studied in the series of papers György et al. (2007); Dani et al. (2008); Abernethy et al. (2008); Cesa-Bianchi and Lugosi (2009); Helmbold and Warmuth (2009); Koolen et al. (2010); Uchiya et al. (2010); Kale et al. (2010) and Audibert and Bubeck (2010). We provide simple proofs that recover most of the previous results. We propose new upper bounds for the semi-bandit game. Moreover we derive lower bounds for all three feedback assumptions. With the only exception of the bandit game, the upper and lower bounds are tight, up to a constant factor. Finally, we answer a question asked by Koolen et al. (2010) by showing that the exponentially weighted average forecaster is suboptimal against  $L_\infty$  adversaries.

## 1. Introduction

In the sequential decision making problems considered in this paper, at each time instance  $t = 1, \dots, n$ , the forecaster chooses, possibly in a randomized way, an action from a given set  $\mathcal{S}$  where  $\mathcal{S}$  is a subset of the  $d$ -dimensional hypercube  $\{0, 1\}^d$ . The action chosen by the forecaster at time  $t$  is denoted by  $V_t = (V_{1,t}, \dots, V_{d,t}) \in \mathcal{S}$ . Simultaneously to the forecaster, the adversary chooses a loss vector  $\ell_t = (\ell_{1,t}, \dots, \ell_{d,t}) \in [0, +\infty)^d$  and the loss incurred by the forecaster is  $\ell_t^T V_t$ . The goal of the forecaster is to minimize the expected cumulative loss  $\mathbb{E} \sum_{t=1}^n \ell_t^T V_t$  where the expectation is taken with respect to the forecaster’s internal randomization. This problem is an instance of an “online linear optimization” problem<sup>1</sup>, see, e.g., Awerbuch and Kleinberg (2004); McMahan and

---

1. In online linear optimization problems, the action set is often not restricted to be a subset of  $\{0, 1\}^d$  but can be an arbitrary subset of  $\mathbb{R}^d$ . However, in the most interesting cases, actions are naturally represented by Boolean vectors and we restrict our attention to this case.

Parameters: set of actions  $\mathcal{S} \subset \{0, 1\}^d$ ; number of rounds  $n \in \mathbb{N}$ .

For each round  $t = 1, 2, \dots, n$ :

- (1) the forecaster chooses  $V_t \in \mathcal{S}$  with the help of an external randomization;
- (2) simultaneously the adversary selects a loss vector  $\ell_t \in [0, +\infty)^d$  (without revealing it);
- (3) the forecaster incurs the loss  $\ell_t^T V_t$ . He observes
  - the loss vector  $\ell_t$  in the full information game,
  - the coordinates  $\ell_{i,t} \mathbb{1}_{V_{i,t}=1}$  in the semi-bandit game,
  - the instantaneous loss  $\ell_t^T V_t$  in the bandit game.

*Goal:* The forecaster tries to minimize his cumulative loss  $\sum_{t=1}^n \ell_t^T V_t$ .

Figure 1: Combinatorial prediction games.

Blum (2004); Kalai and Vempala (2005); György et al. (2007); Dani et al. (2008); Abernethy et al. (2008); Cesa-Bianchi and Lugosi (2009); Helmbold and Warmuth (2009); Koolen et al. (2010); Uchiya et al. (2010) and Kale et al. (2010)

We consider three variants of the problem, distinguished by the type of information that becomes available to the forecaster at each time instance, after taking an action. (1) In the *full information game* the forecaster observes the entire loss vector  $\ell_t$ ; (2) in the *semi-bandit game* only those components  $\ell_{i,t}$  of  $\ell_t$  are observable for which  $V_{i,t} = 1$ ; (3) in the *bandit game* only the total loss  $\ell_t^T V_t$  becomes available to the forecaster.

We refer to these problems as *combinatorial prediction games*. All three prediction games are sketched in Figure 1. For all three games, we define the regret<sup>2</sup> of the forecaster as

$$\bar{R}_n = \mathbb{E} \sum_{t=1}^n \ell_t^T V_t - \min_{v \in \mathcal{S}} \mathbb{E} \sum_{t=1}^n \ell_t^T v.$$

In order to make meaningful statements about the regret, one needs to restrict the possible loss vectors the adversary may assign. We work with two different natural assumptions that have been considered in the literature:

**$L_\infty$  assumption:** here we assume that  $\|\ell_t\|_\infty \leq 1$  for all  $t = 1, \dots, n$

**$L_2$  assumption:** assume that  $\ell_t^T v \leq 1$  for all  $t = 1, \dots, n$  and  $v \in \mathcal{S}$ .

Note that, without loss of generality, we may assume that for all  $i \in \{1, \dots, d\}$ , there exists  $v \in \mathcal{S}$  with  $v_i = 1$ , and then the  $L_2$  assumption implies the  $L_\infty$  assumption.

The goal of this paper is to study the *minimax regret*, that is, the performance of the forecaster that minimizes the regret for the worst possible sequence of loss assignments. This, of course, depends on the set  $\mathcal{S}$  of actions. Our aim is to determine the order of magnitude of the minimax regret for the most difficult set to learn. More precisely, for a given game, if we write  $\text{sup}$  for the

2. For the full information game, one can directly upper bound the stronger notion of regret  $\mathbb{E} \sum_{t=1}^n \ell_t^T V_t - \mathbb{E} \min_{v \in \mathcal{S}} \sum_{t=1}^n \ell_t^T v$  which is always larger than  $\bar{R}_n$ . However, for partial information games, this requires more work.

	$L_\infty$			$L_2$		
	Full Info	Semi-Bandit	Bandit	Full Info	Semi-Bandit	Bandit
Lower Bound	$d\sqrt{n}$	$d\sqrt{n}$	$\mathbf{d}^{3/2}\sqrt{n}$	$\sqrt{dn}$	$\sqrt{dn}$	$d\sqrt{n}$
Upper Bound	$d\sqrt{n}$	$\mathbf{d}\sqrt{n}$	$d^{5/2}\sqrt{n}$	$\sqrt{dn}$	$\sqrt{dn \log d}$	$d^{3/2}\sqrt{n}$

 Table 1: Bounds on  $R_n$  proved in this paper (up to constant factor). The new results are set in bold.

	$L_\infty$			$L_2$		
	Full Info	Semi-Bandit	Bandit	Full Info	Semi-Bandit	Bandit
EXP2	$d^{3/2}\sqrt{n}$	$d^{3/2}\sqrt{n}$	$d^{5/2}\sqrt{n}$	$\sqrt{dn}$	$\mathbf{d}\sqrt{n}^*$	$d^{3/2}\sqrt{n}$
LINEXP	$d\sqrt{n}$	$\mathbf{d}\sqrt{n}$	$\mathbf{d}^2n^{2/3}$	$\sqrt{dn}$	$\mathbf{d}\sqrt{n}^*$	$\mathbf{d}^2n^{2/3}$
LINPOLY	$\mathbf{d}\sqrt{n}$	$\mathbf{d}\sqrt{n}$	-	$\sqrt{dn}$	$\sqrt{dn \log d}$	-

 Table 2: Upper bounds on  $\bar{R}_n$  for specific forecasters. The new results are in bold. We also show that the bound for EXP2 in the full information game is unimprovable. Note that the bound for (Bandit, LINEXP) is very weak. The bounds with \* become  $\sqrt{dn \log d}$  if we restrict our attention to sets  $\mathcal{S}$  that are “almost symmetric” in the sense that for some  $k$ ,  $\mathcal{S} \subset \{v \in \{0, 1\}^d : \sum_{i=1}^d v_i \leq k\}$  and  $\text{Conv}(\mathcal{S}) \cap [\frac{k}{2d}; 1]^d \neq \emptyset$ .

supremum over all allowed adversaries (that is, either  $L_\infty$  or  $L_2$  adversaries) and inf for the infimum over all forecaster strategies for this game, we are interested in the maximal minimax regret

$$R_n = \max_{\mathcal{S} \subset \{0, 1\}^d} \inf \sup \bar{R}_n.$$

Note that in this paper we do not restrict our attention to computationally efficient algorithms. The following example illustrates the different games that we introduced above.

**Example 1** Consider the well studied example of path planning in which, at every time instance, the forecaster chooses a path from one fixed vertex to another in a graph. At each time, a loss is assigned to every edge of the graph and, depending on the model of the feedback, the forecaster observes either the losses of all edges, the losses of each edge on the chosen path, or only the total loss of the chosen path. The goal is to minimize the total loss for any sequence of loss assignments. This problem can be cast as a combinatorial prediction game in dimension  $d$  for  $d$  the number of edges in the graph.

Our contribution is threefold. First, we propose a variant of the algorithm used to track the best linear predictor (Herbster and Warmuth, 1998) that is well-suited to our combinatorial prediction games. This leads to an algorithm called CLEB that generalizes various approaches that have been proposed. This new point of view on algorithms that were defined for specific games (only the full information game, or only the standard multi-armed bandit game) allows us to generalize them easily to all combinatorial prediction games, leading to new algorithms such as LINPOLY. This algorithmic contribution leads to our second main result, the improvement of the known upper bounds for the semi-bandit game. This point of view also leads to a different proof of the minimax

$\sqrt{nd}$  regret bound in the standard  $d$ -armed bandit game that is much simpler than the one provided in Audibert and Bubeck (2010). A summary of the bounds proved in this paper can be found in Table 1 and Table 2. In addition we prove several lower bounds. First, we establish lower bounds on the minimax regret in all three games and under both types of adversaries, whereas only the cases ( $L_2/L_\infty$ , Full Information) and ( $L_2$ , Bandit) were previously treated in the literature. Moreover we also answer a question of Koolen et al. (2010) by showing that the traditional exponentially weighted average forecaster is suboptimal against  $L_\infty$  adversaries.

In particular, this paper leads to the following (perhaps unexpected) conclusions:

- The full information game is as hard as the semi-bandit game. More precisely, in terms of  $R_n$ , the price that one pays for the limited feedback of the semi-bandit game compared to the full information game is only a constant factor (or a  $\sqrt{\log d}$  factor for the  $L_2$  setting).
- In the full information and semi-bandit game, the traditional exponentially weighted average forecaster is provably suboptimal for  $L_\infty$  adversaries while it is optimal for  $L_2$  adversaries in the full information game.
- Denote by  $\mathcal{A}_2$  (respectively  $\mathcal{A}_\infty$ ) the set of adversaries that satisfy the  $L_2$  assumption (respectively the  $L_\infty$  assumption). We clearly have  $\mathcal{A}_2 \subset \mathcal{A}_\infty \subset d\mathcal{A}_2$ . We prove that, in the full information game,  $R_n$  gains an additional factor of  $\sqrt{d}$  at each inclusion. In the semi-bandit game, we show that the same statement remains true up to a logarithmic factor.

**Notation.** The convex hull of  $\mathcal{S}$  is denoted  $\text{Conv}(\mathcal{S})$ .

## 2. Combinatorial learning with Bregman projections

In this section we introduce a general forecaster that we call CLEB (Combinatorial LEarning with Bregman projections). Every forecaster investigated in this paper is a special case of CLEB.

Let  $\mathcal{D}$  be a convex subset of  $\mathbb{R}^d$  with nonempty interior  $\text{Int}(\mathcal{D})$  and boundary  $\partial\mathcal{D}$ .

**Definition 1** We call Legendre any function  $F : \mathcal{D} \rightarrow \mathbb{R}$  such that

- (i)  $F$  is strictly convex and admits continuous first partial derivatives on  $\text{Int}(\mathcal{D})$
- (ii) For any  $u \in \partial\mathcal{D}$ , for any  $v \in \text{Int}(\mathcal{D})$ , we have

$$\lim_{s \rightarrow 0, s > 0} (u - v)^T \nabla F((1 - s)u + sv) = +\infty.$$

The Bregman divergence  $D_F : \mathcal{D} \times \text{Int}(\mathcal{D})$  associated to a Legendre function  $F$  is defined by

$$D_F(u, v) = F(u) - F(v) - (u - v)^T \nabla F(v).$$

We consider the algorithm CLEB described in Figure 2. The basic idea is to use a potential-based gradient descent (1) followed by a projection (2) with respect to the Bregman divergence of the potential onto the convex hull of  $\mathcal{S}$  to ensure that the resulting weight vector  $w_{t+1}$  can be viewed as  $w_{t+1} = \mathbb{E}_{V \sim p_{t+1}} V$  for some distribution  $p_{t+1}$  on  $\mathcal{S}$ . The combination of Bregman projections with potential-based gradient descent was first used in Herbster and Warmuth (1998). Online learning

Parameters:

- a Legendre function  $F$  defined on  $\mathcal{D}$  with  $\text{Conv}(\mathcal{S}) \cap \text{Int}(\mathcal{D}) \neq \emptyset$
- $w_1 \in \text{Conv}(\mathcal{S}) \cap \text{Int}(\mathcal{D})$

For each round  $t = 1, 2, \dots, n$ :

- (a) Let  $p_t$  be a distribution on the set  $\mathcal{S}$  such that  $w_t = \mathbb{E}_{V \sim p_t} V$ .
- (b) Draw a random action  $V_t$  according to the distribution  $p_t$  and observe
  - the loss vector  $\ell_t$  in the full information game,
  - the coordinates  $\ell_{i,t} \mathbb{1}_{V_{i,t}=1}$  in the semi-bandit game,
  - the instantaneous loss  $\ell_t^T V_t$  in the bandit game.
- (c) Estimate the loss  $\ell_t$  by  $\tilde{\ell}_t$ . For instance, one may take
  - $\tilde{\ell}_t = \ell_t$  in the full information game,
  - $\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{\sum_{v \in \mathcal{S}: v_i=1} p_t(v)} V_{i,t}$  in the semi-bandit game,
  - $\tilde{\ell}_t = P_t^+ V_t V_t^T \ell_t$ , with  $P_t = \mathbb{E}_{v \sim p_t} (vv^T)$  in the bandit game.
- (d) Let  $w'_{t+1} \in \text{Int}(\mathcal{D})$  satisfying

$$\nabla F(w'_{t+1}) = \nabla F(w_t) - \tilde{\ell}_t. \quad (1)$$

- (e) Project the weight vector  $w'_{t+1}$  defined by (1) on the convex hull of  $\mathcal{S}$ :

$$w_{t+1} \in \underset{w \in \text{Conv}(\mathcal{S}) \cap \text{Int}(\mathcal{D})}{\operatorname{argmin}} D_F(w, w'_{t+1}). \quad (2)$$

Figure 2: Combinatorial learning with Bregman projections (CLEB).

with Bregman divergences without the projection step has a long history (see Section 11.11 of Cesa-Bianchi and Lugosi (2006)). As discussed below, CLEB may be viewed as a generalization of the forecasters LINEXP and INF.

The Legendre conjugate  $F^*$  of  $F$  is defined by  $F^*(u) = \sup_{v \in \mathcal{D}} \{u^T v - F(v)\}$ . The following theorem establishes the first step of all upper bounds for the regret of CLEB.

**Theorem 2** CLEB satisfies for any  $u \in \text{Conv}(\mathcal{S}) \cap \mathcal{D}$ ,

$$\sum_{t=1}^n \tilde{\ell}_t^T w_t - \sum_{t=1}^n \tilde{\ell}_t^T u \leq D_F(u, w_1) + \sum_{t=1}^n D_{F^*}(\nabla F(w_t) - \tilde{\ell}_t, \nabla F(w_t)). \quad (3)$$

**Proof** By applying the definition of the Bregman divergences (or equivalently using Lemma 11.1 of Cesa-Bianchi and Lugosi (2006)), we obtain

$$\begin{aligned} \tilde{\ell}_t^T w_t - \tilde{\ell}_t^T u &= (u - w_t)^T (\nabla F(w'_{t+1}) - \nabla F(w_t)) \\ &= D_F(u, w_t) + D_F(w_t, w'_{t+1}) - D_F(u, w'_{t+1}). \end{aligned}$$

By the Pythagorean theorem (Lemma 11.3 of Cesa-Bianchi and Lugosi (2006)), we have  $D_F(u, w'_{t+1}) \geq D_F(u, w_{t+1}) + D_F(w_{t+1}, w'_{t+1})$ , hence

$$\tilde{\ell}_t^T w_t - \tilde{\ell}_t^T u \leq D_F(u, w_t) + D_F(w_t, w'_{t+1}) - D_F(u, w_{t+1}) - D_F(w_{t+1}, w'_{t+1}).$$

Summing over  $t$  then gives

$$\sum_{t=1}^n \tilde{\ell}_t^T w_t - \sum_{t=1}^n \tilde{\ell}_t^T u \leq D_F(u, w_1) - D_F(u, w_{n+1}) + \sum_{t=1}^n (D_F(w_t, w'_{t+1}) - D_F(w_{t+1}, w'_{t+1})). \quad (4)$$

By the nonnegativity of the Bregman divergences, we get

$$\sum_{t=1}^n \tilde{\ell}_t^T w_t - \sum_{t=1}^n \tilde{\ell}_t^T u \leq D_F(u, w_1) + \sum_{t=1}^n D_F(w_t, w'_{t+1}).$$

From Proposition 11.1 of Cesa-Bianchi and Lugosi (2006), we have  $D_F(w_t, w'_{t+1}) = D_{F^*}(\nabla F(w_t) - \tilde{\ell}_t, \nabla F(w_t))$ , which concludes the proof.  $\blacksquare$

As we will see below, by the equality  $\mathbb{E} \sum_{t=1}^n \tilde{\ell}_t^T V_t = \mathbb{E} \sum_{t=1}^n \tilde{\ell}_t^T w_t$ , and provided that  $\tilde{\ell}_t^T V_t$  and  $\tilde{\ell}_t^T u$  are unbiased estimates of  $\mathbb{E} \ell_t^T V_t$  and  $\mathbb{E} \ell_t^T u$ , Theorem 2 leads to an upper bound on the regret  $\bar{R}_n$  of CLEB, which allows us to obtain the bounds of Table 2 by using appropriate choices of  $F$ . Moreover, if  $F$  admits an Hessian, denoted  $\nabla^2 F$ , that is always invertible, then one can prove that up to a third-order term (in  $\tilde{\ell}_t$ ), the regret bound can be written as:

$$\sum_{t=1}^n \tilde{\ell}_t^T w_t - \sum_{t=1}^n \tilde{\ell}_t^T u \lesssim D_F(u, w_1) + \sum_{t=1}^n \tilde{\ell}_t^T (\nabla^2 F(w_t))^{-1} \tilde{\ell}_t. \quad (5)$$

In this paper, we restrict our attention to the combinatorial learning setting in which  $\mathcal{S}$  is a subset of  $\{0, 1\}^d$ . However, one should note that this specific form of  $\mathcal{S}$  plays no role in the definition of CLEB, meaning that the algorithm on Figure 2 can be used to handle general online linear optimization problems, where  $\mathcal{S}$  is any subset of  $\mathbb{R}^d$ .

### 3. Different instances of CLEB

In this section we describe several instances of CLEB and relate them to existing algorithms. Figure 3 summarizes the relationship between the various algorithms introduced below.

#### 3.1. EXP2 (Expanded Exponentially weighted average forecaster)

The simplest approach to combinatorial prediction games is to consider each vertex of  $\mathcal{S}$  as an independent expert, and then apply a strategy designed for the expert problem. We call EXP2 the resulting strategy when one uses the traditional exponentially weighted average forecaster (also called Hedge, Freund and Schapire (1997)), see Figure 4. In the full information game, EXP2 corresponds to Expanded Hedge defined in Koolen et al. (2010), where it was studied under the  $L_\infty$  assumption. It was also studied in the full information game under the  $L_2$  assumption in Dani et al. (2008). In

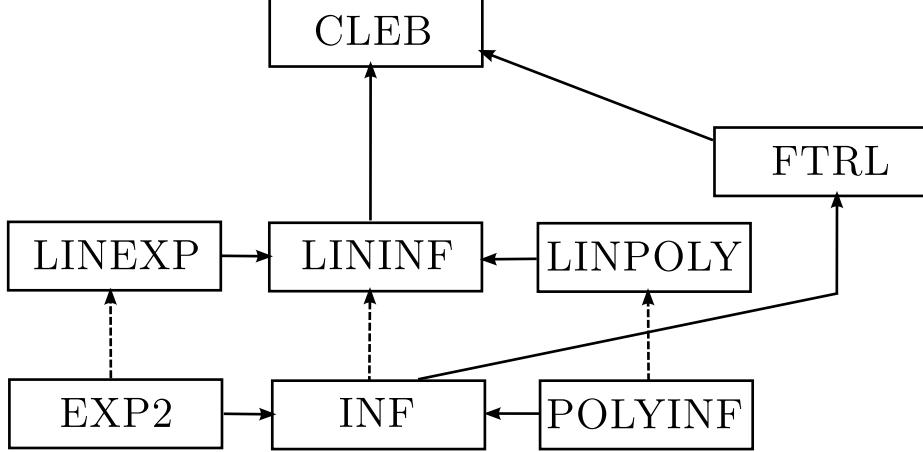


Figure 3: The figure sketches the relationship of the algorithms studied in this paper with arrows representing “is a special case of”. Dotted arrows indicate that the link is obtained by “expanding”  $\mathcal{S}$ , that is seeing  $\mathcal{S}$  as the set of basis vector in  $\mathbb{R}^{|\mathcal{S}|}$  rather than seeing it as a (structured) subset of  $\{0, 1\}^d$  (see Section 3.1). The six algorithms on the bottom use a Legendre function with a diagonal Hessian. On the contrary, the FTRL algorithm (see Section 3.3) may consider Legendre functions more adapted to the geometry of the convex hull of  $\mathcal{S}$ . POLYINF is the algorithm considered in Theorem 22.

the semi-bandit game, EXP2 was studied in György et al. (2007) under the  $L_\infty$  assumption. Finally in the bandit game, EXP2 corresponds to the strategy proposed by Dani et al. (2008) and also to the ComBand strategy, studied under the  $L_\infty$  assumption in Cesa-Bianchi and Lugosi (2009) and under the  $L_2$  assumption in Cesa-Bianchi and Lugosi (2010). (These last strategies differ in how the losses are estimated.)

EXP2 is a CLEB strategy in dimension  $|\mathcal{S}|$  that uses  $\mathcal{D} = [0, +\infty)^{|\mathcal{S}|}$  and the function  $F : u \mapsto \frac{1}{\eta} \sum_{i=1}^{|\mathcal{S}|} u_i \log(u_i)$ , for some  $\eta > 0$  (this can be proved by using the fact that the Kullback-Leibler projection on the simplex is equivalent to a  $L_1$ -normalization). The following theorem shows the regret bound that one can obtain for EXP2 (for instance with Theorem 5 applied to the case where  $\mathcal{S}$  is replaced by  $\mathcal{S}' = \{u \in \{0, 1\}^{|\mathcal{S}|} : \sum_{v \in \mathcal{S}} u_v = 1\}$ ).

**Theorem 3** *For the EXP2 forecaster, provided that  $\mathbb{E}\tilde{\ell}_t = \ell_t$ , we have*

$$\bar{R}_n \leq \frac{\log(|\mathcal{S}|)}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{v \in \mathcal{S}} \mathbb{E}[p_t(v)(\tilde{\ell}_t^T v)^2 \max(1, \exp(-\eta \tilde{\ell}_t^T v))].$$

### 3.2. LINEXP (Linear Exponentially weighted average forecaster)

We call LINEXP the CLEB strategy that uses  $\mathcal{D} = [0, +\infty)^d$  and the function  $F : u \mapsto \frac{1}{\eta} \sum_{i=1}^d u_i \log(u_i)$  associated to the Kullback-Leibler divergence, for some  $\eta > 0$ . In the full information game, LINEXP corresponds to Component Hedge defined in Koolen et al. (2010), where it was studied under the  $L_\infty$  assumption. In the semi-bandit game, LINEXP was studied in Uchiya et al. (2010); Kale

EXP2:

Parameter: Learning rate  $\eta$ .

Let  $w_1 = (\frac{1}{|\mathcal{S}|}, \dots, \frac{1}{|\mathcal{S}|}) \in \mathbb{R}^{|\mathcal{S}|}$ .

For each round  $t = 1, 2, \dots, n$ ;

(a) Let  $p_t$  the distribution on  $\mathcal{S}$  such that  $p_t(v) = w_{v,t}$  for any  $v \in \mathcal{S}$ .

(b) Play  $V_t \sim p_t$  and observe

- the loss vector  $\ell_t$  in the full information game,
- the coordinates  $\ell_{i,t} \mathbb{1}_{V_{i,t}=1}$  in the semi-bandit game,
- the instantaneous loss  $\ell_t^T V_t$  in the bandit game.

(c) Estimate the loss vector  $\ell_t$  by  $\tilde{\ell}_t$ . For instance, one may take

- $\tilde{\ell}_t = \ell_t$  in the full information game,
- $\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{\sum_{v \in \mathcal{S}: v_i=1} p_{v,t}}$  in the semi-bandit game,
- $\tilde{\ell}_t = P_t^+ V_t V_t^T \ell_t$ , with  $P_t = \mathbb{E}_{v \sim p_t}(vv^T)$  in the bandit game.

(d) Update the weights, for all  $v \in \mathcal{S}$ ,

$$w_{v,t+1} = \frac{\exp(-\eta \tilde{\ell}_t^T v) w_{v,t}}{\sum_{u \in \mathcal{S}} \exp(-\eta \tilde{\ell}_t^T u) w_{u,t}}.$$

Figure 4: EXP2 forecaster.

et al. (2010) under the  $L_\infty$  assumption, and for the particular set  $\mathcal{S}$  with all vertices of  $L_1$  norm equal to some value  $k$ .

### 3.3. FTRL (Follow the Regularized Leader)

If  $\text{Conv}(\mathcal{S}) \subset \mathcal{D}$  and  $w_1 \in \arg\min_{w \in \mathcal{D}} F(w)$ , steps (d) and (e) are equivalent to

$$w_{t+1} \in \arg\min_{w \in \text{Conv}(\mathcal{S})} \left( \sum_{s=1}^t \tilde{\ell}_s^T w + F(w) \right),$$

showing that in this case CLEB can be interpreted as a regularized follow-the-leader algorithm. This type of algorithm was studied in Abernethy and Rakhlin (2009) in the full information and bandit setting (see also the lecture notes Rakhlin and Tewari (2008)). A survey of FTRL strategies for the full information game can be found in Hazan (2010). In the bandit game, FTRL with  $F$  being a self-concordant barrier function and a different estimate than the one proposed in Figure 2 was studied in Abernethy et al. (2008).

### 3.4. LININF (Linear Implicitly Normalized Forecaster)

Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ . The function  $f$  has a diagonal Hessian if and only if it can be written as  $f(u) = \sum_{i=1}^d f_i(u_i)$ , for some twice differentiable functions  $f_i : \mathbb{R} \rightarrow \mathbb{R}$ ,  $i = 1, \dots, d$ . The Hessian is

called exchangeable when the functions  $f_1'', \dots, f_d''$  are identical. In this case, up to adding an affine function of  $u$  (note that this does not alter neither the Bregman divergence nor CLEB), we have  $f(u) = \sum_{i=1}^d g(u_i)$  for some twice differentiable function  $g$ . In this section, we consider this type of Legendre functions. To underline the surprising link<sup>3</sup> with the Implicitly Normalized Forecaster proposed in Audibert and Bubeck (2010), we consider  $g$  of the form  $x \mapsto \int_x^\infty \psi^{-1}(s)ds$ , and will refer to the algorithm presented hereafter as LININF.

**Definition 4** Let  $\omega \geq 0$ . A function  $\psi : (-\infty, a) \rightarrow \mathbb{R}_+^*$  for some  $a \in \mathbb{R} \cup \{+\infty\}$  is called an  $\omega$ -potential if and only if it is convex, continuously differentiable, and satisfies

$$\begin{aligned} \lim_{x \rightarrow -\infty} \psi(x) &= \omega & \lim_{x \rightarrow a} \psi(x) &= +\infty \\ \psi' > 0 & & \int_\omega^{\omega+1} |\psi^{-1}(s)|ds &< +\infty. \end{aligned}$$

**Theorem 5** Let  $\omega \geq 0$  and let  $\psi$  be an  $\omega$ -potential function. The function  $F$  defined on  $\mathcal{D} = [\omega, +\infty)^d$  by  $F(u) = \sum_{i=1}^d \int_\omega^{u_i} \psi^{-1}(s)ds$  is Legendre. The associated CLEB satisfies, for any  $u \in \text{Conv}(\mathcal{S}) \cap \mathcal{D}$ ,

$$\sum_{t=1}^n \tilde{\ell}_t^T w_t - \sum_{t=1}^n \tilde{\ell}_t^T u \leq D_F(u, w_1) + \frac{1}{2} \sum_{t=1}^n \sum_{i=1}^d \tilde{\ell}_{i,t}^2 \max \left( \psi'(\psi^{-1}(w_{i,t})), \psi'(\psi^{-1}(w_{i,t}) - \tilde{\ell}_{i,t}) \right), \quad (6)$$

where for any  $(u, v) \in \mathcal{D} \times \text{Int}(\mathcal{D})$ ,

$$D_F(u, v) = \sum_{i=1}^d \left( \int_{v_i}^{u_i} \psi^{-1}(s)ds - (u_i - v_i)\psi^{-1}(v_i) \right). \quad (7)$$

In particular, when the estimates  $\tilde{\ell}_{i,t}$  are nonnegative, we have

$$\sum_{t=1}^n \tilde{\ell}_t^T w_t - \sum_{t=1}^n \tilde{\ell}_t^T u \leq D_F(u, w_1) + \sum_{t=1}^n \sum_{i=1}^d \frac{\tilde{\ell}_{i,t}^2}{2(\psi^{-1})'(w_{i,t})}. \quad (8)$$

**Proof** It is easy to check that  $F$  is a Legendre function and that (7) holds. We also have  $\nabla F^*(u) = (\nabla F)^{-1}(u) = (\psi(u_1), \dots, \psi(u_d))$ , hence

$$D_{F^*}(u, v) = \sum_{i=1}^d \left( \int_{v_i}^{u_i} \psi(s)ds - (u_i - v_i)\psi(v_i) \right).$$

From the Taylor-Lagrange expansion, we have  $D_{F^*}(u, v) \leq \sum_{i=1}^d \max_{s \in [u_i, v_i]} \frac{1}{2}\psi'(s)(u_i - v_i)^2$ . Since the function  $\psi$  is convex, we have  $\max_{s \in [u_i, v_i]} \psi'(s) \leq \psi'(\max(u_i, v_i))$ , which gives the desired results.  $\blacksquare$

---

3. detailed in Appendix A.

Note that LINEXP is an instance of LININF with  $\psi : x \mapsto \exp(\eta x)$ . On the other hand, Audibert and Bubeck (2010) recommend the choice  $\psi(x) = (-\eta x)^{-q}$  with  $\eta > 0$  and  $q > 1$  since it leads to the minimax optimal rate  $\sqrt{nd}$  for the standard  $d$ -armed bandit game (while the best bound for Exp3 is of the order of  $\sqrt{nd \log d}$ ). This corresponds to a function  $F$  of the form  $F(u) = -\frac{q}{(q-1)\eta} \sum_{i=1}^d u_i^{(q-1)/q}$ . We refer to the corresponding CLEB as LINPOLY. In Appendix A we show that a simple application of Theorem 5 proves that LINPOLY with  $q = 2$  satisfies  $\bar{R}_n \leq 2\sqrt{2nd}$ . This improves on the bound  $\bar{R}_n \leq 8\sqrt{nd}$  obtained in Theorem 11 of Audibert and Bubeck (2010).

#### 4. Full Information Game

This section details the upper bounds of the forecasters EXP2, LINEXP and LINPOLY under the  $L_2$  and  $L_\infty$  assumptions for the full information game. All results are gathered in Table 2 (page 109). The proofs can be found in Appendix B. Up to numerical constants, the results concerning (EXP2,  $L_2$  and  $L_\infty$ ) and (LINEXP,  $L_\infty$ ) appeared or can be easily derived from respectively Dani et al. (2008) and Koolen et al. (2010).

**Theorem 6 (LINEXP,  $L_\infty$ )** *Under the  $L_\infty$  assumption, for LINEXP with  $\tilde{\ell}_t = \ell_t$ ,  $\eta = \sqrt{2/n}$  and  $w_1 = \operatorname{argmin}_{w \in \operatorname{Conv}(\mathcal{S})} D_F(w, (1, \dots, 1)^T)$ , we have*

$$\bar{R}_n \leq d\sqrt{2n}.$$

**Theorem 7 (LINEXP,  $L_2$ )** *Under the  $L_2$  assumption, for LINEXP with  $\tilde{\ell}_t = \ell_t$ ,  $\eta = \sqrt{2d/n}$  and  $w_1 = \operatorname{argmin}_{w \in \operatorname{Conv}(\mathcal{S})} D_F(w, (1, \dots, 1)^T)$ , we have*

$$\bar{R}_n \leq \sqrt{2nd}.$$

**Theorem 8 (LINPOLY,  $L_\infty$ )** *Under the  $L_\infty$  assumption, for LINPOLY with  $\tilde{\ell}_t = \ell_t$ ,  $\eta = \sqrt{\frac{2}{q(q-1)n}}$  and  $w_1 = \operatorname{argmin}_{w \in \operatorname{Conv}(\mathcal{S})} D_F(w, (1, \dots, 1)^T)$ , we have*

$$\bar{R}_n \leq d\sqrt{\frac{2qn}{q-1}}.$$

**Theorem 9 (LINPOLY,  $L_2$ )** *Under the  $L_2$  assumption, for LINPOLY with  $\tilde{\ell}_t = \ell_t$ ,  $\eta = \sqrt{\frac{2d}{q(q-1)n}}$  and  $w_1 = \operatorname{argmin}_{w \in \operatorname{Conv}(\mathcal{S})} D_F(w, (1, \dots, 1)^T)$ , we have*

$$\bar{R}_n \leq \sqrt{\frac{2qdn}{q-1}}.$$

**Theorem 10 (EXP2,  $L_\infty$ )** *Under the  $L_\infty$  assumption, for EXP2 with  $\tilde{\ell}_t = \ell_t$ , we have*

$$\bar{R}_n \leq \frac{d \log 2}{\eta} + \frac{\eta nd^2}{2}.$$

In particular for  $\eta = \sqrt{\frac{2 \log 2}{nd}}$ , we have  $\bar{R}_n \leq \sqrt{2d^3 n \log 2}$ .

From Theorem 19, the above upper bound is tight, and consequently there exists  $\mathcal{S}$  for which the algorithm EXP2 is not minimax optimal in the full information game under the  $L_\infty$  assumption.

**Theorem 11 (EXP2,  $L_2$ )** *Under the  $L_2$  assumption, for EXP2 with  $\tilde{\ell}_t = \ell_t$ , we have*

$$\bar{R}_n \leq \frac{d \log 2}{\eta} + \frac{\eta n}{2}.$$

In particular for  $\eta = \sqrt{\frac{2d \log 2}{n}}$ , we have  $\bar{R}_n \leq \sqrt{2dn \log 2}$ .

## 5. Semi-Bandit Game

This section details the upper bounds of the forecasters EXP2, LINEXP and LINPOLY under the  $L_2$  and  $L_\infty$  assumptions for the semi-bandit game. These bounds are gathered in Table 2 (page 109). The proofs can be found in Appendix C. Up to the numerical constant, the result concerning (EXP2,  $L_\infty$ ) appeared in György et al. (2007) in the context of the online shortest path problem. Uchiya et al. (2010) and Kale et al. (2010) studied the semi-bandit problem under the  $L_\infty$  assumption for action sets of the form  $\mathcal{S} = \{v \in \{0, 1\}^d : \sum_{i=1}^d v_i = k\}$  for some value  $k$ . Their common algorithm corresponds to LINEXP and the bounds are of order  $\sqrt{knd \log(d/k)}$ . Our upper bounds for the regret of LINEXP extend these results to more general sets of arms and to the  $L_2$  assumption.

**Theorem 12 (LINEXP,  $L_\infty$ )** *Under the  $L_\infty$  assumption, for LINEXP with  $\tilde{\ell}_{i,t} = \ell_{i,t} \frac{V_{i,t}}{w_{i,t}}$ ,  $\eta = \sqrt{2/n}$  and  $w_1 = \operatorname{argmin}_{w \in \operatorname{Conv}(\mathcal{S})} D_F(w, (1, \dots, 1)^T)$ , we have*

$$\bar{R}_n \leq d\sqrt{2n}.$$

Since the  $L_2$  assumption implies the  $L_\infty$  assumption, we also have  $\bar{R}_n \leq d\sqrt{2n}$  under the  $L_2$  assumption.

Let us now detail how LINEXP behaves for almost symmetric action sets as defined below.

**Definition 13** *The set  $\mathcal{S} \subset \{0, 1\}^d$  is called almost symmetric if for some  $k \in \mathbb{N}$ ,  $\mathcal{S} \subset \{v \in \{0, 1\}^d : \sum_{i=1}^d v_i \leq k\}$  and  $\operatorname{Conv}(\mathcal{S}) \cap [\frac{k}{2d}; 1]^d \neq \emptyset$ . The integer  $k$  is called the order of the symmetry.*

The set  $\mathcal{S} = \{v \in \{0, 1\}^d : \sum_{i=1}^d v_i = k\}$  considered in Uchiya et al. (2010) and Kale et al. (2010) is a particular almost symmetric set.

**Theorem 14 (LINEXP, almost symmetric  $\mathcal{S}$ )** *Let  $\mathcal{S}$  be an almost symmetric set of order  $k \in \mathbb{N}$ . Consider LINEXP with  $\tilde{\ell}_{i,t} = \ell_{i,t} \frac{V_{i,t}}{w_{i,t}}$  and  $w_1 = \operatorname{argmin}_{w \in \operatorname{Conv}(\mathcal{S})} D_F(w, (\frac{k}{d}, \dots, \frac{k}{d})^T)$ . Let  $\mathcal{L} = \max(\log(\frac{d}{k}), 1)$ .*

- Under the  $L_\infty$  assumption, taking  $\eta = \sqrt{\frac{2k\mathcal{L}}{nd}}$ , we have  $\bar{R}_n \leq \sqrt{2knd\mathcal{L}}$ .
- Under the  $L_2$  assumption, taking  $\eta = k\sqrt{\frac{\mathcal{L}}{nd}}$ , we have  $\bar{R}_n \leq 2\sqrt{nd\mathcal{L}}$ .

In particular, it means that under the  $L_2$  assumption, there is a gain in the regret bound of a factor  $\sqrt{d/\mathcal{L}}$  when the set of actions is an almost symmetric set of order  $k$ .

**Theorem 15 (LINPOLY,  $L_\infty$ )** Under the  $L_\infty$  assumption, for LINPOLY with  $\tilde{\ell}_{i,t} = \ell_{i,t} \frac{V_{i,t}}{w_{i,t}}$ ,  $\eta = \sqrt{\frac{2}{q(q-1)n}}$  and  $w_1 = \operatorname{argmin}_{w \in \operatorname{Conv}(\mathcal{S})} D_F(w, (1, \dots, 1)^T)$ , we have

$$\bar{R}_n \leq d \sqrt{\frac{2qn}{q-1}}.$$

**Theorem 16 (LINPOLY,  $L_2$ )** Under the  $L_2$  assumption, for LINPOLY with  $\tilde{\ell}_{i,t} = \ell_{i,t} \frac{V_{i,t}}{w_{i,t}}$ ,  $\eta = \sqrt{\frac{2d^{\frac{1}{q}}}{q(q-1)n}}$  and  $w_1 = \operatorname{argmin}_{w \in \operatorname{Conv}(\mathcal{S})} D_F(w, (1, \dots, 1)^T)$ , we have

$$\bar{R}_n \leq \sqrt{\frac{2qnd}{q-1} d^{1-\frac{1}{q}}}.$$

In particular, for  $q = 1 + (\log d)^{-1}$ , we have  $\bar{R}_n \leq \sqrt{2nde \log(ed)}$ .

**Theorem 17 (EXP2,  $L_\infty$ )** Under the  $L_\infty$  assumption, for the EXP2 forecaster described in Figure 4 using  $\tilde{\ell}_{i,t} = \ell_{i,t} \frac{V_{i,t}}{w_{i,t}}$ , we have

$$\bar{R}_n \leq \frac{d \log 2}{\eta} + \frac{\eta nd^2}{2}.$$

In particular for  $\eta = \sqrt{\frac{2 \log 2}{nd}}$ , we have  $\bar{R}_n \leq \sqrt{2d^3 n \log 2}$ .

The corresponding lower bound is given in Theorem 19.

**Theorem 18 (EXP2,  $L_2$ )** Under the  $L_2$  assumption, for EXP2 with  $\tilde{\ell}_{i,t} = \ell_{i,t} \frac{V_{i,t}}{w_{i,t}}$ , we have

$$\bar{R}_n \leq \frac{d \log 2}{\eta} + \frac{\eta nd}{2}.$$

In particular for  $\eta = \sqrt{\frac{2 \log 2}{n}}$ , we have  $\bar{R}_n \leq d \sqrt{2n \log 2}$ .

Note that as for LINEXP, we end up upper bounding  $\sum_{i=1}^d \ell_{i,t}$  by  $d$ . In the case of almost symmetric set  $\mathcal{S}$  of order  $k$ , this sum can be bounded by  $2d/k$ , while  $\log(|\mathcal{S}|)$  is upper bounded by  $k \log(d+1)$ . So as for LINEXP, this leads to a regret bound of order  $\sqrt{nd \log d}$  when the set of actions is an almost symmetric set.

## 6. Bandit Game

The upper bounds for EXP2 in the bandit case proposed in Table 2 (page 109) are extracted from Dani et al. (2008). The approach proposed by the authors is to use EXP2 in the space described by a barycentric spanner. More precisely, let  $m = \dim(\operatorname{Span}(\mathcal{S}))$  and  $e_1, \dots, e_m$  be a barycentric spanner of  $\mathcal{S}$ ; for instance, take  $(e_1, \dots, e_m) \in \operatorname{argmax}_{(x_1, \dots, x_m) \in \mathcal{S}^m} |\det_{\operatorname{Span}(\mathcal{S})}(x_1, \dots, x_m)|$  (see Awerbuch and Kleinberg, 2004). We introduce the transformations  $T_1 : \mathbb{R}^d \rightarrow \mathbb{R}^m$  such that for  $x \in \mathbb{R}^d$ ,  $T_1(x) = (x^T e_1, \dots, x^T e_m)^T$ , and  $T_2 : \mathcal{S} \rightarrow [-1, 1]^m$  such that for  $v \in \mathcal{S}$ ,  $v =$

$\sum_{i=1}^m (T_2(v))_i e_i$ . Note that for any  $v \in \mathcal{S}$ , we have  $\ell_t^T v = T_1(\ell_t)^T T_2(v)$ . Then the loss estimate for  $v \in \mathcal{S}$  is

$$\tilde{\ell}_t^T v = (Q_t^+ T_2(V_t) T_2(V_t)^T T_1(\ell_t))^T T_2(v), \text{ where } Q_t = \mathbb{E}_{V \sim p_t} T_2(V) T_2(V)^T.$$

Moreover the authors also add a forced exploration which is uniform over the barycentric spanner.

A concurrent approach is the one proposed in Cesa-Bianchi and Lugosi (2009, 2010). There the authors study EXP2 directly in the original space, with the estimate described in Figure 4, and with an additional forced exploration which is uniform over  $\mathcal{S}$ . They work out several examples of sets  $\mathcal{S}$  for which they improve the regret bound by a factor  $\sqrt{d}$  with respect to Dani et al. (2008). Unfortunately there exists sets  $\mathcal{S}$  for which this approach fails to provide a bound polynomial in  $d$ . In general one needs to replace the uniform exploration over  $\mathcal{S}$  by an exploration that is tailored to this set. How to do this in general is still an open question.

The upper bounds for LINEXP in the bandit case proposed in Table 2 (page 109) are derived by using the trick of Dani et al. (2008) (that is, by working with a barycentric spanner). The proof of this result is omitted, since it does not yield the optimal dependency in  $n$ . Moreover we can not analyze LINPOLY since (1) is not well defined in this case, because  $\tilde{\ell}_t$  can be non-positive. In general we believe that the LININF approach is not sound for the bandit case, and that one needs to work with a Legendre function with non-diagonal Hessian.

The only known CLEB with non-diagonal Hessian is the one proposed in Abernethy et al. (2008), where the authors use a self-concordant barrier function. In this case, they are able to propose a loss estimate related to the structure of the Hessian. This approach is powerful, and under the  $L_2$  assumption leads to a regret upper bound of order  $d\sqrt{\theta n \log n}$  for  $\theta > 0$  such that  $\text{Conv}(\mathcal{S})$  admits a  $\theta$ -self-concordant barrier function (see Abernethy et al., 2008, section 5). When  $\text{Conv}(\mathcal{S})$  admits a  $O(1)$ -self-concordant barrier function, the upper bound matches the lower bound  $O(d\sqrt{n})$ . The open question is to determine for which sets  $\mathcal{S}$ , this occurs.

## 7. Lower Bounds

We start this Section with a result that shows that EXP2 is suboptimal against  $L_\infty$  adversaries. This answers a question of Koolen et al. (2010).

**Theorem 19** *Let  $n \geq d$ . There exists a subset  $\mathcal{S} \subset \{0,1\}^d$  such that in the full information game, for the EXP2 strategy (for any learning rate  $\eta$ ), we have*

$$\sup \bar{R}_n \geq 0.02 d^{3/2} \sqrt{n},$$

where the supremum is taken over all  $L_\infty$  adversaries.

**Proof** For sake of simplicity we assume here that  $d$  is a multiple of 4 and that  $n$  is even. We consider the following subset of the hypercube:

$$\begin{aligned} \mathcal{S} = \left\{ v \in \{0,1\}^d : \sum_{i=1}^{d/2} v_i = d/4 \text{ and} \right. \\ \left. \left( v_i = 1, \forall i \in \{d/2 + 1, \dots, d/2 + d/4\} \right) \text{ or } \left( v_i = 1, \forall i \in \{d/2 + d/4 + 1, \dots, d\} \right) \right\}. \end{aligned}$$

That is, choosing a point in  $\mathcal{S}$  corresponds to choosing a subset of  $d/4$  elements in the first half of the coordinates, and choosing one of the two first disjoint intervals of size  $d/4$  in the second half of the coordinates.

We will prove that for any parameter  $\eta$ , there exists an adversary such that Exp (with parameter  $\eta$ ) has a regret of at least  $\frac{nd}{16} \tanh\left(\frac{\eta d}{8}\right)$ , and that there exists another adversary such that its regret is at least  $\min\left(\frac{d \log 2}{12\eta}, \frac{nd}{12}\right)$ . As a consequence, we have

$$\begin{aligned} \sup \bar{R}_n &\geq \max\left(\frac{nd}{16} \tanh\left(\frac{\eta d}{8}\right), \min\left(\frac{d \log 2}{12\eta}, \frac{nd}{12}\right)\right) \\ &\geq \min\left(\max\left(\frac{nd}{16} \tanh\left(\frac{\eta d}{8}\right), \frac{d \log 2}{12\eta}\right), \frac{nd}{12}\right) \geq \min\left(A, \frac{nd}{12}\right), \end{aligned}$$

with

$$\begin{aligned} A &= \min_{\eta \in [0, +\infty)} \max\left(\frac{nd}{16} \tanh\left(\frac{\eta d}{8}\right), \frac{d \log 2}{12\eta}\right) \\ &\geq \min\left\{\min_{\eta d \geq 8} \frac{nd}{16} \tanh\left(\frac{\eta d}{8}\right), \min_{\eta d < 8} \max\left(\frac{nd}{16} \tanh\left(\frac{\eta d}{8}\right), \frac{d \log 2}{12\eta}\right)\right\} \\ &\geq \min\left\{\frac{nd}{16} \tanh(1), \min_{\eta d < 8} \max\left(\frac{nd}{16} \frac{\eta d}{8} \frac{1}{\tanh(1)}, \frac{d \log 2}{12\eta}\right)\right\} \\ &\geq \min\left\{\frac{nd}{16} \tanh(1), \sqrt{\frac{nd^3 \log 2}{128 \times 12 \times \tanh(1)}}\right\} \geq \min(0.04 nd, 0.02 d^{3/2} \sqrt{n}). \end{aligned}$$

Let us first prove the lower bound  $\frac{nd}{16} \tanh\left(\frac{\eta d}{8}\right)$ . We define the following adversary:

$$\ell_{i,t} = \begin{cases} 1 & \text{if } i \in \{d/2 + 1; \dots, d/2 + d/4\} \text{ and } t \text{ odd,} \\ 1 & \text{if } i \in \{d/2 + d/4 + 1, \dots, d\} \text{ and } t \text{ even,} \\ 0 & \text{otherwise.} \end{cases}$$

This adversary always put a zero loss on the first half of the coordinates, and alternates between a loss of  $d/4$  for choosing the first interval (in the second half of the coordinates) and the second interval. At the beginning of odd rounds, any vertex  $v \in \mathcal{S}$  has the same cumulative loss and thus Exp picks its expert uniformly at random, which yields an expected cumulative loss equal to  $nd/16$ . On the other hand at even rounds the probability distribution to select the vertex  $v \in \mathcal{S}$  is always the same. More precisely the probability of selecting a vertex which contains the interval  $\{d/2 + d/4 + 1, \dots, d\}$  (i.e, the interval with a  $d/4$  loss at this round) is exactly  $\frac{1}{1 + \exp(-\eta d/4)}$ . This adds an expected cumulative loss equal to  $\frac{nd}{8} \frac{1}{1 + \exp(-\eta d/4)}$ . Finally note that the loss of any fixed vertex is  $nd/8$ . Thus we obtain

$$\bar{R}_n = \frac{nd}{16} + \frac{nd}{8} \frac{1}{1 + \exp(-\eta d/4)} - \frac{nd}{8} = \frac{nd}{16} \tanh\left(\frac{\eta d}{8}\right).$$

We move now to the dependency in  $1/\eta$ . Here we consider the adversary defined by:

$$\ell_{i,t} = \begin{cases} 1 - \varepsilon & \text{if } i \leq d/4, \\ 1 & \text{if } i \in \{d/4 + 1, \dots, d/2\}, \\ 0 & \text{otherwise.} \end{cases}$$

Note that against this adversary the choice of the interval (in the second half of the components) does not matter. Moreover by symmetry the weight of any coordinate in  $\{d/4 + 1, \dots, d/2\}$  is the same (at any round). Finally remark that this weight is decreasing with  $t$ . Thus we have the following identities (in the big sums  $i$  represents the number of components selected in the first  $d/4$  components):

$$\begin{aligned}\bar{R}_n &= \mathbb{E} \left( \varepsilon \sum_{t=1}^n \sum_{i=d/4+1}^{d/2} V_{i,t} \right) = \varepsilon \frac{d}{4} \sum_{t=1}^n \mathbb{E} V_{d/2,t} \geq \frac{n\varepsilon d}{4} \mathbb{P}(V_{d/2,n} = 1) \\ &= \frac{n\varepsilon d}{4} \frac{\sum_{v \in \mathcal{S}: v_{d/2}=1} \exp(-\eta n \ell_2^T v)}{\sum_{v \in \mathcal{S}} \exp(-\eta n \ell_2^T v)} \\ &= \frac{n\varepsilon d}{4} \frac{\sum_{i=0}^{d/4-1} \binom{d/4}{i} \binom{d/4-1}{d/4-i-1} \exp(-\eta(nd/4 - in\varepsilon))}{\sum_{i=0}^{d/4} \binom{d/4}{i} \binom{d/4}{d/4-i} \exp(-\eta(nd/4 - in\varepsilon))} \\ &= \frac{n\varepsilon d}{4} \frac{\sum_{i=0}^{d/4-1} \binom{d/4}{i} \binom{d/4-1}{d/4-i-1} \exp(\eta in\varepsilon)}{\sum_{i=0}^{d/4} \binom{d/4}{i} \binom{d/4}{d/4-i} \exp(\eta in\varepsilon)} \\ &= \frac{n\varepsilon d}{4} \frac{\sum_{i=0}^{d/4-1} \left(1 - \frac{4i}{d}\right) \binom{d/4}{i} \binom{d/4}{d/4-i} \exp(\eta in\varepsilon)}{\sum_{i=0}^{d/4} \binom{d/4}{i} \binom{d/4}{d/4-i} \exp(\eta in\varepsilon)}\end{aligned}$$

where we used  $\binom{d/4-1}{d/4-i-1} = \left(1 - \frac{4i}{d}\right) \binom{d/4}{d/4-i}$  in the last equality. Thus taking  $\varepsilon = \min\left(\frac{\log 2}{\eta n}, 1\right)$  yields

$$\bar{R}_n \geq \min\left(\frac{d \log 2}{4\eta}, \frac{nd}{4}\right) \frac{\sum_{i=0}^{d/4-1} \left(1 - \frac{4i}{d}\right) \binom{d/4}{i}^2 \min(2, \exp(\eta n))^i}{\sum_{i=0}^{d/4} \binom{d/4}{i}^2 \min(2, \exp(\eta n))^i} \geq \min\left(\frac{d \log 2}{12\eta}, \frac{nd}{12}\right),$$

where the last inequality follows from Lemma 23 (see Appendix E). This concludes the proof of the lower bound.  $\blacksquare$

The next two theorems give lower bounds under the three feedback assumptions and the two types of adversaries. The cases  $(L_2, \text{Full Information})$  and  $(L_2, \text{Bandit})$  already appeared in Dani et al. (2008), while the case  $(L_\infty, \text{Full Information})$  was treated in Koolen et al. (2010) (with more precise lower bounds for subsets  $\mathcal{S}$  of particular interest). Note that the lower bounds for the semi-bandit case trivially follow from the ones for the full information game. Thus our main contribution here is the lower bound for  $(L_\infty, \text{Bandit})$ , which is technically quite different from the other cases. We also give explicit constants in all cases.

**Theorem 20** *Let  $n \geq d$ . Against  $L_\infty$  adversaries in the cases of full information and semi-bandit games, we have*

$$R_n \geq 0.008 d\sqrt{n},$$

*and in the bandit game*

$$R_n \geq 0.01 d^{3/2} \sqrt{n}.$$

**Proof** In this proof we consider the following subset of  $\{0, 1\}^d$ :

$$\mathcal{S} = \{v \in \{0, 1\}^d : \forall i \in \{1, \dots, \lfloor d/2 \rfloor\}, v_{2i-1} + v_{2i} = 1\}.$$

Under full information, playing in  $\mathcal{S}$  corresponds to playing  $\lfloor d/2 \rfloor$  independent standard full information games with 2 experts. Thus we can apply [Theorem 30, Audibert and Bubeck (2010)] to obtain:

$$R_n \geq \lfloor d/2 \rfloor \times 0.03\sqrt{n \log 2} \geq 0.008 d\sqrt{n}.$$

We now move to the bandit game, for which the proof is more challenging. For the sake of simplicity, we assume in the following that  $d$  is even. Moreover, we restrict our attention to deterministic forecasters, the extension to general forecaster can be done by a routine application of Fubini's theorem.

*First step: definitions.*

We denote by  $I_{i,t} \in \{1, 2\}$  the random variable such that  $V_{2i,t} = 1$  if and only if  $I_{i,t} = 2$ . That is,  $I_{i,t}$  is the expert chosen at time  $t$  in the  $i^{th}$  game. We also define the empirical distribution of plays  $q_n^i = (q_{1,n}^i, q_{2,n}^i)$  in game  $i$  as  $q_{j,n}^i = \frac{\sum_{t=1}^n \mathbb{1}_{I_{i,t}=j}}{n}$ . Let  $J_{i,n}$  be drawn according to  $q_n^i$ .

In this proof we consider a set of  $2^{d/2}$  adversaries. For  $\alpha = (\alpha_1, \dots, \alpha_{d/2}) \in \{1, 2\}^{d/2}$  we define the  $\alpha$ -adversary as follows: For any  $t \in \{1, \dots, n\}$ , the loss of expert  $\alpha_i$  in game  $i$  is drawn from a Bernoulli of parameter  $1/2$  while the loss of the other expert in game  $i$  is drawn from a Bernoulli of parameter  $1/2 + \varepsilon$ . We note  $\mathbb{E}_\alpha$  when we integrate with respect to the reward generation process of the  $\alpha$ -adversary. We note  $\mathbb{P}_{i,\alpha}$  the law of  $J_{i,n}$  when the forecaster plays against the  $\alpha$ -adversary. Remark that we have  $\mathbb{P}_{i,\alpha}(J_{i,n} = j) = \mathbb{E}_\alpha \frac{1}{n} \sum_{t=1}^n \mathbb{1}_{I_{i,t}=j}$ , hence, against the  $\alpha$ -adversary we have:

$$\bar{R}_n = \mathbb{E}_\alpha \sum_{t=1}^n \sum_{i=1}^{d/2} \varepsilon \mathbb{1}_{I_{i,t} \neq \alpha_i} = n\varepsilon \sum_{i=1}^{d/2} (1 - \mathbb{P}_{i,\alpha}(J_{i,t} = \alpha_i)),$$

which implies (since the maximum is larger than the mean)

$$\sup_{\alpha \in \{1, 2\}^{d/2}} \bar{R}_n \geq n\varepsilon \sum_{i=1}^{d/2} \left( 1 - \frac{1}{2^{d/2}} \sum_{\alpha \in \{1, 2\}^{d/2}} \mathbb{P}_{i,\alpha}(J_{i,n} = \alpha_i) \right). \quad (9)$$

*Second step: information inequality.*

Let  $\mathbb{P}_{-i,\alpha}$  be the law of  $J_{i,n}$  against the adversary which plays like the  $\alpha$ -adversary except that in the  $i^{th}$  game, the losses of both coordinates are drawn from a Bernoulli of parameter  $1/2 + \varepsilon$  (we call it the  $(-i, \alpha)$ -adversary). Now we use Pinsker's inequality which gives:

$$\mathbb{P}_{i,\alpha}(J_{i,n} = \alpha_i) \leq \mathbb{P}_{-i,\alpha}(J_{i,n} = \alpha_i) + \sqrt{\frac{1}{2} \text{KL}(\mathbb{P}_{-i,\alpha}, \mathbb{P}_{i,\alpha})},$$

and thus, (thanks to the concavity of the square root)

$$\frac{1}{2^{d/2}} \sum_{\alpha \in \{1, 2\}^{d/2}} \mathbb{P}_{i,\alpha}(J_{i,n} = \alpha_i) \leq \frac{1}{2} + \sqrt{\frac{1}{2^{d/2+1}} \sum_{\alpha \in \{1, 2\}^{d/2}} \text{KL}(\mathbb{P}_{-i,\alpha}, \mathbb{P}_{i,\alpha})}. \quad (10)$$

Third step: computation of  $\text{KL}(\mathbb{P}_{-i,\alpha}, \mathbb{P}_{i,\alpha})$  with the chain rule for Kullback-Leibler divergence.

Note that since the forecaster is deterministic, the sequence of observed losses (up to time  $n$ )  $W_n \in \{0, 1, \dots, d\}^n$  uniquely determines the empirical distribution of plays  $q_n^i$ , and in particular the law of  $J_{i,n}$  conditionally to  $W_n$  is the same for any adversary. Thus, if we note  $\mathbb{P}_\alpha^n$  (respectively  $\mathbb{P}_{-i,\alpha}^n$ ) the law of  $W_n$  when the forecaster plays against the  $\alpha$ -adversary (respectively the  $(-i, \alpha)$ -adversary), then one can easily prove that  $\text{KL}(\mathbb{P}_{-i,\alpha}, \mathbb{P}_{i,\alpha}) \leq \text{KL}(\mathbb{P}_{-i,\alpha}^n, \mathbb{P}_\alpha^n)$ . Now we use the chain rule for Kullback-Leibler divergence iteratively to introduce the laws  $\mathbb{P}_\alpha^t$  of the observed losses  $W_t$  up to time  $t$ . More precisely, we have,

$$\begin{aligned}\text{KL}(\mathbb{P}_{-i,\alpha}^n, \mathbb{P}_\alpha^n) &= \text{KL}(\mathbb{P}_{-i,\alpha}^1, \mathbb{P}_\alpha^1) + \sum_{t=2}^n \sum_{w_{t-1} \in \{0,1,\dots,d\}^{t-1}} \mathbb{P}_{-i,\alpha}^{t-1}(w_{t-1}) \text{KL}(\mathbb{P}_{-i,\alpha}^t(\cdot|w_{t-1}), \mathbb{P}_\alpha^t(\cdot|w_{t-1})) \\ &= \text{KL}(\mathcal{B}_\emptyset, \mathcal{B}'_\emptyset) \mathbb{1}_{I_{i,1}=\alpha_i} + \sum_{t=2}^n \sum_{w_{t-1}: I_{i,t}=\alpha_i} \mathbb{P}_{-i,\alpha}^{t-1}(w_{t-1}) \text{KL}(\mathcal{B}_{w_{t-1}}, \mathcal{B}'_{w_{t-1}}),\end{aligned}$$

where  $\mathcal{B}_{w_{t-1}}$  and  $\mathcal{B}'_{w_{t-1}}$  are sums of  $d/2$  Bernoulli distributions with parameters in  $\{1/2, 1/2 + \varepsilon\}$  and such that the number of Bernoullis with parameter  $1/2 + \varepsilon$  in  $\mathcal{B}_{w_{t-1}}$  is equal to the number of Bernoullis with parameter  $1/2 + \varepsilon$  in  $\mathcal{B}'_{w_{t-1}}$  plus one. Now using Lemma 24 (see Appendix E) we obtain  $\text{KL}(\mathbb{P}_{-i,\alpha}^n, \mathbb{P}_\alpha^n) \leq \frac{16\varepsilon^2}{d} \mathbb{E}_{-i,\alpha} \sum_{t=1}^n \mathbb{1}_{I_{i,t}=\alpha_i}$ . Summing and plugging this into (10) we obtain  $\frac{1}{2^{d/2}} \sum_{\alpha \in \{1,2\}^{d/2}} \mathbb{P}_{i,\alpha}(J_{i,n} = \alpha_i) \leq \frac{1}{2} + 2\varepsilon\sqrt{\frac{n}{d}}$ . To conclude the proof one needs to plug in this last equation in (9) along with straightforward computations. ■

**Theorem 21** Let  $n \geq d$ . Against  $L_2$  adversaries in the cases of full information and semi-bandit games, we have

$$R_n \geq 0.05\sqrt{dn},$$

and in the bandit game

$$R_n \geq 0.05 \min(n, d\sqrt{n}).$$

## References

- J. Abernethy and A. Rakhlin. Beating the adaptive bandit with high probability. In *22nd annual conference on learning theory*, 2009.
- J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In Rocco A. Servedio and Tong Zhang, editors, *COLT*, pages 263–274. Omnipress, 2008.
- J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *JMLR*, 11:2635–2686, 2010.

- B. Awerbuch and R.D. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *STOC '04: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53. ACM, 2004.
- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. ISBN 0521841089.
- N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. In *22nd annual conference on learning theory*, 2009.
- N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Submitted*, 2010.
- V. Dani, T. Hayes, and S.M. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, volume 20, pages 345–352, 2008.
- Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.
- A. György, T. Linder, G. Lugosi, and G. Ottucsák. The on-line shortest path problem under partial monitoring. *J. Mach. Learn. Res.*, 8:2369–2403, 2007.
- E. Hazan. A survey: The convex optimization approach to regret minimization. Working draft, 2010.
- D. P. Helmbold and M. K. Warmuth. Learning permutations with exponential weights. *JMLR*, 10: 1705–1736, 2009.
- M. Herbster and M. K. Warmuth. Tracking the best expert. *Mach. Learn.*, 32:151–178, August 1998. ISSN 0885-6125.
- A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71:291307, 2005.
- S. Kale, L. Reyzin, and R. Schapire. Non-stochastic bandit slate problems. *Advances in Neural Information Processing Systems*, pages 1054–1062, 2010.
- W. M. Koolen, M. K. Warmuth, and J. Kivinen. Hedging structured concepts. In *23rd annual conference on learning theory*, 2010.
- H. B. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *In Proceedings of the 17th Annual Conference on Learning Theory*, pages 109–123, 2004.
- A. Rakhlin and A. Tewari. Lecture notes on online learning. 2008.
- T. Uchiya, A. Nakamura, and M. Kudo. Algorithms for adversarial bandit problems with multiple plays. In *Proc. of the 21st International Conference on Algorithmic Learning Theory*, 2010.

Parameters: set of actions  $\mathcal{A} = \{1, \dots, d\}$ ; number of rounds  $n \in \mathbb{N}$ .

For each round  $t = 1, 2, \dots, n$ :

- (1) the forecaster chooses  $A_t \in \mathcal{A}$ , with the help of an external randomization;
- (2) simultaneously the adversary selects a loss vector  $\ell_t = (\ell_{1,t}, \dots, \ell_{d,t})^T \in \mathbb{R}^d$  (without revealing it);
- (3) the forecaster incurs the loss  $\ell_{A_t,t}$ . He observes
  - the loss vector  $\ell_t$  in the full information game,
  - the coordinate  $\ell_{A_t,t}$  in the bandit game.

*Goal:* The forecaster tries to minimize his cumulative loss  $\sum_{t=1}^n \ell_{A_t,t}$ .

Figure 5: Standard prediction games.

## Appendix A. Standard prediction games

It is well-known that the standard prediction games described in Figure 5 are specific cases of the combinatorial prediction games described in Figure 1. Indeed, consider  $\mathcal{S} = \{\mathbf{a}_1, \dots, \mathbf{a}_d\}$ , where  $\mathbf{a}_i \in \{0, 1\}^d$  is the vector whose only nonzero component is the  $i$ -th one. The standard and combinatorial prediction games are then equivalent by using  $V_t = \mathbf{a}_{A_t}$  and noticing that  $\ell_t^T \mathbf{a}_i = \ell_{i,t}$ . In particular, the semi-bandit and bandit combinatorial prediction games are then both equivalent to the traditional multi-armed bandit game.

We now show that INF (defined in Figure 2 of Audibert and Bubeck (2010)) is a special case of LININF.

**Proof** Indeed, suppose that the estimates  $\tilde{\ell}_1, \dots, \tilde{\ell}_n$  are nonnegative (coordinate-wise), and take  $w_1 = (\frac{1}{d}, \dots, \frac{1}{d})$ . Then the vector  $w'_{t+1}$  satisfying (1) exists, and is defined coordinate-wise by  $\psi^{-1}(w'_{i,t+1}) = \psi^{-1}(w_{i,t}) - \tilde{\ell}_{i,t}$ . Besides, the optimality condition of (2) implies the existence of  $c_t \in \mathbb{R}$  (independent of  $i$ ) such that  $\psi^{-1}(w_{i,t+1}) = \psi^{-1}(w'_{i,t+1}) + c_t$ . It implies  $\psi^{-1}(w_{i,t}) = \psi^{-1}(w_{i,1}) - \sum_{s=1}^{t-1} (\tilde{\ell}_{i,s} - c_s)$  for any  $t \geq 1$ . So there exists  $C_t \in \mathbb{R}$  such that  $w_{i,t} = \psi(\sum_{s=1}^{t-1} (1 - \tilde{\ell}_{i,s}) - C_t)$ . Since  $w_t \in \text{Conv}(\mathcal{S})$ , the constant  $C_t$  should satisfy  $\sum_{i=1}^n w_{i,t} = 1$ . We thus recover INF with the estimate  $1 - \tilde{\ell}_{i,s}$  of the reward  $1 - \ell_{i,t}$ . So the Bregman projection has here a simple solution depending on a unique constant  $C_t$  obtained by solving the equality  $\sum_{i=1}^n \psi(\sum_{s=1}^{t-1} (1 - \tilde{\ell}_{i,s}) - C_t) = 1$ . ■

Next we show how to obtain the minimax  $\sqrt{nd}$  regret bound, with a much simpler proof than the one proposed in Audibert and Bubeck (2010), as well as a better constant.

**Theorem 22** Let  $q > 1$ . For the INF forecaster (that is for CLEB with  $w_1 = (\frac{1}{d}, \dots, \frac{1}{d})^T$  and  $\mathcal{S} = \{\mathbf{a}_1, \dots, \mathbf{a}_d\}$ ) using  $\psi(x) = (-\eta x)^{-q}$  and  $\tilde{\ell}_{i,t} = \ell_{i,t} \frac{V_{i,t}}{w_{i,t}}$ , we have

$$\bar{R}_n \leq \frac{qd^{\frac{1}{q}}}{\eta(q-1)} + \frac{q\eta nd^{1-\frac{1}{q}}}{2}.$$

In particular for  $\eta = \sqrt{2}d^{\frac{1}{q}-\frac{1}{2}}[(q-1)n]^{-\frac{1}{2}}$ , we have  $\bar{R}_n \leq q\sqrt{\frac{2nd}{q-1}}$ .

In view of this last bound, the optimal  $q$  is  $q = 2$ , which leads to  $\bar{R}_n \leq 2\sqrt{2nd}$ . This improves on the bound  $\bar{R}_n \leq 8\sqrt{nd}$  obtained in Theorem 11 of Audibert and Bubeck (2010). The INF forecaster with the above polynomial  $\psi$  is referred to as POLYINF in Figure 3.

**Proof** We apply (8). First we bound the divergence term. We have  $\psi^{-1}(s) = -\frac{1}{\eta}s^{-1/q}$ , and

$$D_F(u, v) = \frac{1}{\eta} \sum_{i=1}^d \left( \frac{1}{q-1} v_i^{1-\frac{1}{q}} - \frac{q}{q-1} u_i^{1-\frac{1}{q}} + u_i v_i^{-\frac{1}{q}} \right), \text{ hence}$$

$$\max_{u \in \text{Conv}(\mathcal{S})} D_F(u, w_1) = D_F\left((1, 0, \dots, 0)^T, \left(\frac{1}{d}, \dots, \frac{1}{d}\right)^T\right) = \frac{q}{\eta(q-1)}(d^{\frac{1}{q}} - 1).$$

Combining this with  $(\psi^{-1})'(w_{i,t}) = \frac{1}{q\eta}w_{i,t}^{-1-\frac{1}{q}}$  and (8), we obtain

$$\begin{aligned} \bar{R}_n &\leq \frac{qd^{\frac{1}{q}}}{\eta(q-1)} + \frac{q\eta}{2} \mathbb{E} \sum_{t=1}^n \sum_{i=1}^d w_{i,t}^{1+\frac{1}{q}} \ell_{i,t}^2 \\ &= \frac{qd^{\frac{1}{q}}}{\eta(q-1)} + \frac{q\eta}{2} \sum_{t=1}^n \sum_{i=1}^d \mathbb{E}(w_{i,t}^{\frac{1}{q}-1} V_{i,t} \ell_{i,t}^2) \leq \frac{qd^{\frac{1}{q}}}{\eta(q-1)} + \frac{q\eta}{2} \sum_{t=1}^n \sum_{i=1}^d \mathbb{E}(w_{i,t}^{\frac{1}{q}}) \leq \frac{qd^{\frac{1}{q}}}{\eta(q-1)} + \frac{q\eta}{2} nd^{1-\frac{1}{q}}, \end{aligned}$$

where in the last step we use that by Hölder's inequality,  $\sum_{i=1}^d (w_{i,t}^{\frac{1}{q}} \times 1) \leq (\sum_{i=1}^d w_{i,t})^{\frac{1}{q}} \times d^{1-\frac{1}{q}}$ . ■

## Appendix B. Proofs of Theorems in Section 4

### Proof of Theorem 6

We have  $D_F(u, v) = \frac{1}{\eta} \sum_{i=1}^d \left( u_i \log\left(\frac{u_i}{v_i}\right) - u_i + v_i \right)$ , hence from the Pythagorean theorem,

$$D_F(u, w_1) \leq D_F(u, (1, \dots, 1)^T) \leq \frac{d}{\eta}.$$

Since we have  $\sum_{i=1}^d w_{i,t} \leq d$ , Theorem 5 implies  $\bar{R}_n \leq \frac{d}{\eta} + \frac{\eta}{2} \mathbb{E} \sum_{t=1}^n \sum_{i=1}^d w_{i,t} \ell_{i,t}^2 \leq \frac{d}{\eta} + \frac{nd\eta}{2}$ .

### Proof of Theorem 7

As in the previous proof, we have  $D_F(u, w_1) \leq \frac{d}{\eta}$ , but under the  $L_2$  constraint, we can improve the bound on  $\sum_{i=1}^d w_{i,t} \ell_{i,t}^2$  by using  $\sum_{i=1}^d w_{i,t} \ell_{i,t} \leq 1$  (since  $w_t \in \text{Conv}(\mathcal{S})$ ). This gives  $\bar{R}_n \leq \frac{d}{\eta} + \frac{n\eta}{2}$ .

### Proof of Theorem 8

We have  $D_F(u, v) = \frac{1}{\eta} \sum_{i=1}^d \left( \frac{1}{q-1} v_i^{1-\frac{1}{q}} - \frac{q}{q-1} u_i^{1-\frac{1}{q}} + u_i v_i^{-\frac{1}{q}} \right)$ , hence  $D_F(u, w_1) \leq \frac{d}{\eta(q-1)}$ .

Since we have  $w_{i,t}^{1+\frac{1}{q}} \ell_{i,t}^2 \leq 1$ , Theorem 5 implies  $\bar{R}_n \leq \frac{d}{\eta(q-1)} + \frac{\eta q}{2} \mathbb{E} \sum_{t=1}^n \sum_{i=1}^d w_{i,t}^{1+\frac{1}{q}} \ell_{i,t}^2 \leq \frac{d}{\eta(q-1)} + \frac{ndq\eta}{2}$ .

**Proof of Theorem 9**

As in the previous proof, we have  $D_F(u, w_1) \leq \frac{d}{\eta(q-1)}$ . Under the  $L_2$  constraint, we can improve the bound on  $\sum_{i=1}^d w_{i,t}^{1+\frac{1}{q}} \ell_{i,t}^2$  by using  $\sum_{i=1}^d w_{i,t} \ell_{i,t} \leq 1$  (since  $w_t \in \text{Conv}(\mathcal{S})$ ). This gives

$$\bar{R}_n \leq \frac{d}{\eta(q-1)} + \frac{\eta q}{2} \mathbb{E} \sum_{t=1}^n \sum_{i=1}^d w_{i,t}^{1+\frac{1}{q}} \ell_{i,t}^2 \leq \frac{d}{\eta(q-1)} + \frac{nq\eta}{2}.$$

**Proof of Theorem 10**

Using  $\log(|\mathcal{S}|) \leq d \log 2$ ,  $0 \leq \tilde{\ell}_t^T v \leq d$  and  $\sum_{v \in \mathcal{S}} w_{v,t} = 1$  in Theorem 3, we get the result.

**Proof of Theorem 11**

Using  $\log(|\mathcal{S}|) \leq d \log 2$ ,  $0 \leq \tilde{\ell}_t^T v \leq 1$  and  $\sum_{v \in \mathcal{S}} w_{v,t} = 1$  in Theorem 3, we get the result.

**Appendix C. Proofs of Theorems in Section 5**
**Proof of Theorem 12**

We have again  $D_F(u, w_1) \leq \frac{d}{\eta}$ . Since we have  $\sum_{i=1}^d \ell_{i,t} \leq d$ , Theorem 5 implies

$$\bar{R}_n \leq \frac{d}{\eta} + \frac{\eta}{2} \mathbb{E} \sum_{t=1}^n \sum_{i=1}^d \ell_{i,t}^2 \frac{V_{i,t}}{w_{i,t}} = \frac{d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{i=1}^d \mathbb{E}(\ell_{i,t}^2) \leq \frac{d}{\eta} + \frac{nd\eta}{2}.$$

**Proof of Theorem 14**

The starting point is Theorem 5, which, by using  $\mathbb{E}(\ell_{i,t}^2 \frac{V_{i,t}}{w_{i,t}}) = \mathbb{E}\ell_{i,t}^2 \leq \mathbb{E}\ell_{i,t}$ , implies

$$\bar{R}_n \leq D_F(u, w_1) + \frac{\eta}{2} \mathbb{E} \sum_{t=1}^n \sum_{i=1}^d \ell_{i,t}^2 \frac{V_{i,t}}{w_{i,t}} \leq D_F(u, w_1) + \frac{\eta}{2} \mathbb{E} \sum_{t=1}^n \sum_{i=1}^d \ell_{i,t}. \quad (11)$$

For any  $u \in [0, 1]^d$  such that  $\sum_{i=1}^d u_i \leq k$ , we have

$$D_F(u, w_1) \leq D_F\left(u, \left(\frac{k}{d}, \dots, \frac{k}{d}\right)^T\right) \leq \frac{1}{\eta} \left(k + \sum_{i=1}^d u_i \log\left(\frac{du_i}{ke}\right)\right) \leq \frac{k\mathcal{L}}{\eta}, \quad (12)$$

where the last inequality can be obtained by writing the optimality conditions. More precisely, two cases are considered depending on whether holds  $\sum_{i=1}^d u_i = k$  at the optimum: when it is the case, the maximum is achieved for  $u$  of the form  $u = (1, \dots, 1, 0, \dots, 0)^T$ ; otherwise,  $u = (0, \dots, 0)^T$  achieves the maximum. The desired results are then obtained by combining (11), (12) and an upper bound on  $\sum_{i=1}^d \ell_{i,t}$ : indeed, under the  $L_\infty$  assumption, we have  $\sum_{i=1}^d \ell_{i,t} \leq d$ . Under the  $L_2$  assumption, since  $\mathcal{S}$  is an almost symmetric set of order  $k$ , there exists  $z \in \text{Conv}(\mathcal{S}) \cap [\frac{k}{2d}; 1]^d$ , and consequently  $\sum_{i=1}^d \ell_{i,t} \leq \sum_{i=1}^d (\frac{2d}{k} z_i) \ell_{i,t} \leq \frac{2d}{k}$ .

**Proof of Theorem 15**

We have again  $D_F(u, w_1) \leq \frac{d}{\eta(q-1)}$ . Since we have  $w_{i,t}^{\frac{1}{q}} \ell_{i,t}^2 \leq 1$ , Theorem 5 implies

$$\bar{R}_n \leq \frac{d}{\eta(q-1)} + \frac{\eta q}{2} \sum_{t=1}^n \sum_{i=1}^d \mathbb{E}(w_{i,t}^{\frac{1}{q}} \ell_{i,t}^2) \leq \frac{d}{\eta(q-1)} + \frac{ndq\eta}{2}.$$

**Proof of Theorem 16**

We have again  $D_F(u, w_1) \leq \frac{d}{\eta(q-1)}$ . From  $\mathbb{E}(w_{i,t}^{1+\frac{1}{q}} \tilde{\ell}_{i,t}^2) = \mathbb{E}(w_{i,t}^{\frac{1}{q}} \ell_{i,t}^2) \leq \mathbb{E}[(w_{i,t} \ell_{i,t})^{\frac{1}{q}}]$  and Theorem 5, we get

$$\bar{R}_n \leq \frac{d}{\eta(q-1)} + \frac{\eta q}{2} \sum_{t=1}^n \sum_{i=1}^d \mathbb{E}[(w_{i,t} \ell_{i,t})^{\frac{1}{q}}] \leq \frac{d}{\eta(q-1)} + \frac{nd^{1-\frac{1}{q}}q\eta}{2}.$$

where we use  $\sum_{i=1}^d (w_{i,t} \ell_{i,t})^{\frac{1}{q}} \leq (\sum_{i=1}^d w_{i,t} \ell_{i,t})^{\frac{1}{q}} \times d^{1-\frac{1}{q}}$  in the last step.

**Proof of Theorem 17**

Let  $q_{i,t} = \sum_{v \in \mathcal{S}: v_i=1} p_{v,t} = \mathbb{E}_{V_t \sim p_t} V_{i,t}$  for  $i \in \{1, \dots, d\}$ . We have

$$\begin{aligned} \mathbb{E}_{V_t \sim p_t} \sum_{v \in \mathcal{S}} p_{v,t} (\tilde{\ell}_t^T v)^2 &= \mathbb{E}_{V_t \sim p_t, V'_t \sim p_t} (\tilde{\ell}_t^T V'_t)^2 \\ &= \mathbb{E}_{V_t \sim p_t, V'_t \sim p_t} \sum_{i,j} \frac{\ell_{i,t} V_{i,t} \ell_{j,t} V_{j,t}}{q_{i,t} q_{j,t}} V'_{i,t} V'_{j,t} \\ &\leq \mathbb{E}_{V_t \sim p_t, V'_t \sim p_t} \sum_{i,j} \ell_{i,t} \ell_{j,t} \frac{V_{i,t}}{q_{i,t}} \frac{V'_{j,t}}{q_{j,t}} = \left( \sum_{i=1}^d \ell_{i,t} \right)^2 \leq d^2. \end{aligned}$$

Using  $\log(|\mathcal{S}|) \leq d \log 2$ , the result then follows from Theorem 3.

**Proof of Theorem 18**

Let  $q_{i,t} = \sum_{v \in \mathcal{S}: v_i=1} p_{v,t} = \mathbb{E}_{V_t \sim p_t} V_{i,t}$  for  $i \in \{1, \dots, d\}$ . We have

$$\begin{aligned} \mathbb{E}_{V_t \sim p_t} \sum_{v \in \mathcal{S}} p_{v,t} (\tilde{\ell}_t^T v)^2 &= \mathbb{E}_{V_t \sim p_t, V'_t \sim p_t} \sum_{i,j} \frac{\ell_{i,t} V_{i,t} \ell_{j,t} V_{j,t}}{q_{i,t} q_{j,t}} V'_{i,t} V'_{j,t} \\ &\leq \mathbb{E}_{V_t, V'_t} \sum_{i,j} \frac{\ell_{i,t} V_{i,t}}{q_{i,t}} \frac{V'_{j,t}}{q_{j,t}} \ell_{j,t} V_{j,t} \\ &= \mathbb{E}_{V_t} \sum_{i=1}^d \frac{\ell_{i,t} V_{i,t}}{q_{i,t}} \sum_{j=1}^d \ell_{j,t} V_{j,t} \leq \mathbb{E}_{V_t} \sum_{i=1}^d \frac{\ell_{i,t} V_{i,t}}{q_{i,t}} = \sum_{i=1}^d \ell_{i,t} \leq d. \end{aligned}$$

Using  $\log(|\mathcal{S}|) \leq d \log 2$ , the result then follows from Theorem 3.

## Appendix D. Proof of Theorem 21

We consider the bandit game first. We use the notation and adversaries defined in the proof of Theorem 20. We modify these adversaries as follows: at each turn one selects uniformly at random  $E_t \in \{1, \dots, d\}$ . Then, at time  $t$ , the losses of all coordinates but  $E_t$  are set to 0. This new adversary is clearly in  $L_2$ . For this new set of adversaries, one has to do only two modifications in the proof of Theorem 20. First (9) is replaced by:

$$\sup_{\alpha \in \{1,2\}^{d/2}} \bar{R}_n \geq \frac{n\varepsilon}{d} \sum_{i=1}^{d/2} \left( 1 - \frac{1}{2^{d/2}} \sum_{\alpha \in \{1,2\}^{d/2}} \mathbb{P}_{i,\alpha}(J_{i,n} = \alpha_i) \right).$$

Second  $\mathcal{B}_{w_{t-1}}$  is now a Bernoulli with mean  $\mu_t \in [\frac{1}{2} + \frac{\varepsilon}{d}, \frac{1}{2} + \frac{\varepsilon}{2}]$  and  $\mathcal{B}'_{w_{t-1}}$  is a Bernoulli with mean  $\mu_t - \frac{\varepsilon}{d}$ , and thus we have

$$\text{KL}(\mathcal{B}_{w_{t-1}}, \mathcal{B}'_{w_{t-1}}) \leq \frac{4\varepsilon^2}{(1-\varepsilon^2)d^2}.$$

The proof is then concluded again with straightforward computations.

The proof for the full information game is exactly the same as the one for bandit information, except that the definition of  $W_t$  is slightly different and implies that  $\mathcal{B}_{w_{t-1}}$  is now a Bernoulli with mean  $\frac{1}{d}(\frac{1}{2} + \varepsilon)$  and  $\mathcal{B}'_{w_{t-1}}$  is a Bernoulli with mean  $\frac{1}{2d}$ , which gives

$$\text{KL}(\mathcal{B}_{w_{t-1}}, \mathcal{B}'_{w_{t-1}}) \leq \frac{4\varepsilon^2}{2d-1}.$$

## Appendix E. Technical lemmas

We prove here two technical lemmas that were used in the proofs above.

**Lemma 23** *For any  $k \in \mathbb{N}^*$ , for any  $1 \leq c \leq 2$ , we have*

$$\frac{\sum_{i=0}^k (1-i/k) \binom{k}{i}^2 c^i}{\sum_{i=0}^k \binom{k}{i}^2 c^i} \geq 1/3.$$

**Proof** Let  $f(c)$  denote the left-hand side term of the inequality. Introduce the random variable  $X$ , which is equal to  $i \in \{0, \dots, k\}$  with probability  $\binom{k}{i}^2 c^i / \sum_{j=0}^k \binom{k}{j}^2 c^j$ . We have  $f'(c) = \frac{1}{c} \mathbb{E}[X(1-X/k)] - \frac{1}{c} \mathbb{E}(X) \mathbb{E}(1-X/k) = -\frac{1}{c} \text{Var } X \leq 0$ . So the function  $f$  is decreasing on  $[1, 2]$ , and, from now on, we consider  $c = 2$ . Numerator and denominator of the left-hand side (l.h.s.) differ only by the  $1 - i/k$  factor. A lower bound for the left-hand side can thus be obtained by showing that the terms for  $i$  close to  $k$  are not essential to the value of the denominator. To prove this, we may use the Stirling formula: for any  $n \geq 1$

$$\left(\frac{n}{e}\right)^n \sqrt{2\pi n} < n! < \left(\frac{n}{e}\right)^n \sqrt{2\pi n} e^{1/(12n)} \quad (13)$$

Indeed, this inequality implies that for any  $k \geq 2$  and  $i \in [1, k - 1]$

$$\left(\frac{k}{i}\right)^i \left(\frac{k}{k-i}\right)^{k-i} \frac{\sqrt{k}}{\sqrt{2\pi i(k-i)}} e^{-1/6} < \binom{k}{i} < \left(\frac{k}{i}\right)^i \left(\frac{k}{k-i}\right)^{k-i} \frac{\sqrt{k}}{\sqrt{2\pi i(k-i)}} e^{1/12},$$

hence

$$\left(\frac{k}{i}\right)^{2i} \left(\frac{k}{k-i}\right)^{2(k-i)} \frac{ke^{-1/3}}{2\pi i(k-i)} < \binom{k}{i}^2 < \left(\frac{k}{i}\right)^{2i} \left(\frac{k}{k-i}\right)^{2(k-i)} \frac{ke^{1/6}}{2\pi i}$$

Introduce  $\lambda = i/k$  and  $\chi(\lambda) = \frac{2^\lambda}{\lambda^{2\lambda}(1-\lambda)^{2(1-\lambda)}}$ . We have

$$[\chi(\lambda)]^k \frac{2e^{-1/3}}{\pi k} < \binom{k}{i}^2 2^i < [\chi(\lambda)]^k \frac{e^{1/6}}{2\pi\lambda}. \quad (14)$$

Lemma 23 can be numerically verified for  $k \leq 10^6$ . We now consider  $k > 10^6$ . For  $\lambda \geq 0.666$ , since the function  $\chi$  can be shown to be decreasing on  $[0.666, 1]$ , the inequality  $\binom{k}{i}^2 2^i < [\chi(0.666)]^k \frac{e^{1/6}}{2 \times 0.666 \times \pi}$  holds. We have  $\chi(0.657)/\chi(0.666) > 1.0002$ . Consequently, for  $k > 10^6$ , we have  $[\chi(0.666)]^k < 0.001 \times [\chi(0.657)]^k/k^2$ . So for  $\lambda \geq 0.666$  and  $k > 10^6$ , we have

$$\begin{aligned} \binom{k}{i}^2 2^i &< 0.001 \times [\chi(0.657)]^k \frac{e^{1/6}}{2\pi \times 0.666 \times k^2} < [\chi(0.657)]^k \frac{2e^{-1/3}}{1000\pi k^2} \\ &= \min_{\lambda \in [0.656, 0.657]} [\chi(\lambda)]^k \frac{2e^{-1/3}}{1000\pi k^2} \\ &< \frac{1}{1000k} \max_{i \in \{1, \dots, k-1\} \cap [0, 0.666k]} \binom{k}{i}^2 2^i. \end{aligned} \quad (15)$$

where the last inequality comes from (14) and the fact that there exists  $i \in \{1, \dots, k-1\}$  such that  $i/k \in [0.656, 0.657]$ . Inequality (15) implies that for any  $i \in \{1, \dots, k\}$ , we have

$$\sum_{\frac{5}{6}k \leq i \leq k} \binom{k}{i}^2 2^i < \frac{1}{1000} \max_{i \in \{1, \dots, k-1\} \cap [0, 0.666k]} \binom{k}{i}^2 2^i < \frac{1}{1000} \sum_{0 \leq i < 0.666k} \binom{k}{i}^2 2^i.$$

To conclude, introducing  $A = \sum_{0 \leq i < 0.666k} \binom{k}{i}^2 2^i$ , we have

$$\frac{\sum_{i=0}^k (1 - i/k) \binom{k}{i} \binom{k}{k-i} 2^i}{\sum_{i=0}^k \binom{k}{i} \binom{k}{k-i} 2^i} > \frac{(1 - 0.666)A}{A + 0.001A} \geq \frac{1}{3}.$$

■

**Lemma 24** Let  $\ell$  and  $n$  be integers with  $\frac{1}{2} \leq \frac{n}{2} \leq \ell \leq n$ . Let  $p, p', q, p_1, \dots, p_n$  be real numbers in  $(0, 1)$  with  $q \in \{p, p'\}$ ,  $p_1 = \dots = p_\ell = q$  and  $p_{\ell+1} = \dots = p_n$ . Let  $\mathcal{B}$  (resp.  $\mathcal{B}'$ ) be the sum of  $n+1$  independent Bernoulli distributions with parameters  $p, p_1, \dots, p_n$  (resp.  $p', p_1, \dots, p_n$ ). We have

$$\text{KL}(\mathcal{B}, \mathcal{B}') \leq \frac{2(p' - p)^2}{(1 - p')(n + 2)q}.$$

**Proof** Let  $Z, Z', Z_1, \dots, Z_n$  be independent Bernoulli distributions with parameters  $p, p', p_1, \dots, p_n$ . Define  $S = \sum_{i=1}^{\ell} Z_i$ ,  $T = \sum_{i=\ell+1}^n Z_i$  and  $V = Z + S$ . By slight abuse of notation, merging in the same notation the distribution and the random variable, we have

$$\begin{aligned}\text{KL}(\mathcal{B}, \mathcal{B}') &= \text{KL}((Z + S) + T, (Z' + S) + T) \\ &\leq \text{KL}((Z + S, T), (Z' + S, T)) \\ &= \text{KL}(Z + S, Z' + S).\end{aligned}$$

Let  $s_k = \mathbb{P}(S = k)$  for  $k = -1, 0, \dots, \ell + 1$ . Using the equalities

$$s_k = \binom{\ell}{k} q^k (1-q)^{\ell-k} = \frac{q}{1-q} \frac{\ell-k+1}{k} \binom{\ell}{k-1} q^{k-1} (1-q)^{\ell-k+1} = \frac{q}{1-q} \frac{\ell-k+1}{k} s_{k-1},$$

which holds for  $1 \leq k \leq \ell + 1$ , we obtain

$$\begin{aligned}\text{KL}(Z + S, Z' + S) &= \sum_{k=0}^{\ell+1} \mathbb{P}(V = k) \log \left( \frac{\mathbb{P}(Z + S = k)}{\mathbb{P}(Z' + S = k)} \right) \\ &= \sum_{k=0}^{\ell+1} \mathbb{P}(V = k) \log \left( \frac{ps_{k-1} + (1-p)s_k}{p's_{k-1} + (1-p')s_k} \right) \\ &= \sum_{k=0}^{\ell+1} \mathbb{P}(V = k) \log \left( \frac{p \frac{1-q}{q} k + (1-p)(\ell-k+1)}{p' \frac{1-q}{q} k + (1-p')(\ell-k+1)} \right) \\ &= \mathbb{E} \log \left( \frac{(p-q)V + (1-p)q(\ell+1)}{(p'-q)V + (1-p')q(\ell+1)} \right).\end{aligned}\tag{16}$$

*First case:*  $q = p'$ .

By Jensen's inequality, using that  $\mathbb{E}V = p'(\ell + 1) + p - p'$  in this case, we then get

$$\begin{aligned}\text{KL}(Z + S, Z' + S) &\leq \log \left( \frac{(p-p')\mathbb{E}(V) + (1-p)p'(\ell+1)}{(1-p')p'(\ell+1)} \right) \\ &= \log \left( \frac{(p-p')^2 + (1-p')p'(\ell+1)}{(1-p')p'(\ell+1)} \right) \\ &= \log \left( 1 + \frac{(p-p')^2}{(1-p')p'(\ell+1)} \right) \leq \frac{(p-p')^2}{(1-p')p'(\ell+1)}.\end{aligned}$$

*Second case:*  $q = p$ .

In this case,  $V$  is a binomial distribution with parameters  $\ell + 1$  and  $p$ . From (16), we have

$$\begin{aligned}\text{KL}(Z + S, Z' + S) &\leq -\mathbb{E} \log \left( \frac{(p'-p)V + (1-p')p(\ell+1)}{(1-p)p(\ell+1)} \right) \\ &\leq -\mathbb{E} \log \left( 1 + \frac{(p'-p)(V - \mathbb{E}V)}{(1-p)p(\ell+1)} \right).\end{aligned}\tag{17}$$

To conclude, we will use the following lemma.

**Lemma 25** *The following inequality holds for any  $x \geq x_0$  with  $x_0 \in (0, 1)$ :*

$$-\log(x) \leq -(x - 1) + \frac{(x - 1)^2}{2x_0}.$$

**Proof** Introduce  $f(x) = -(x - 1) + \frac{(x - 1)^2}{2x_0} + \log(x)$ . We have  $f'(x) = -1 + \frac{x-1}{x_0} + \frac{1}{x}$ , and  $f''(x) = \frac{1}{x_0} - \frac{1}{x^2}$ . From  $f'(x_0) = 0$ , we get that  $f'$  is negative on  $(x_0, 1)$  and positive on  $(1, +\infty)$ . This leads to  $f$  nonnegative on  $[x_0, +\infty)$ .  $\blacksquare$

Finally, from Lemma 25 and (17), using  $x_0 = \frac{1-p'}{1-p}$ , we obtain

$$\begin{aligned} \text{KL}(Z + S, Z' + S) &\leq \left( \frac{p' - p}{(1-p)p(\ell+1)} \right)^2 \frac{\mathbb{E}[(V - \mathbb{E}V)^2]}{2x_0} \\ &= \left( \frac{p' - p}{(1-p)p(\ell+1)} \right)^2 \frac{(\ell+1)p(1-p)^2}{2(1-p')} \\ &= \frac{(p' - p)^2}{2(1-p')(\ell+1)p}. \end{aligned}$$

$\blacksquare$

# Minimax Regret of Finite Partial-Monitoring Games in Stochastic Environments\*

Gábor Bartók

BARTOK@CS.UALBERTA.CA

Dávid Pál

DPAL@CS.UALBERTA.CA

Csaba Szepesvári

SZEPESVA@CS.UALBERTA.CA

*Department of Computing Science, University of Alberta, Edmonton, T6G 2E8, AB, Canada*

**Editors:** Sham Kakade, Ulrike von Luxburg

## Abstract

In a partial monitoring game, the learner repeatedly chooses an action, the environment responds with an outcome, and then the learner suffers a loss and receives a feedback signal, both of which are fixed functions of the action and the outcome. The goal of the learner is to minimize his regret, which is the difference between his total cumulative loss and the total loss of the best fixed action in hindsight. Assuming that the outcomes are generated in an i.i.d. fashion from an arbitrary and unknown probability distribution, we characterize the minimax regret of any partial monitoring game with finitely many actions and outcomes. It turns out that the minimax regret of any such game is either zero,  $\tilde{\Theta}(\sqrt{T})$ ,  $\Theta(T^{2/3})$ , or  $\Theta(T)$ . We provide a computationally efficient learning algorithm that achieves the minimax regret within logarithmic factor for any game.

**Keywords:** Online learning, Imperfect feedback, Regret analysis

## 1. Introduction

Partial monitoring provides a mathematical framework for sequential decision making problems with imperfect feedback. Various problems of interest can be modeled as partial monitoring instances, such as learning with expert advice (Littlestone and Warmuth, 1994), the multi-armed bandit problem (Auer et al., 2002), dynamic pricing (Kleinberg and Leighton, 2003), the dark pool problem (Agarwal et al., 2010), label efficient prediction (Cesa-Bianchi et al., 2005), and linear and convex optimization with full or bandit feedback (Zinkevich, 2003; Abernethy et al., 2008; Flaxman et al., 2005).

In this paper we restrict ourselves to finite games, *i.e.*, games where both the set of actions available to the learner and the set of possible outcomes generated by the environment are finite. A finite partial monitoring game  $\mathbf{G}$  is described by a pair of  $N \times M$  matrices: the *loss matrix*  $\mathbf{L}$  and the *feedback matrix*  $\mathbf{H}$ . The entries  $\ell_{i,j}$  of  $\mathbf{L}$  are real numbers lying in, say, the interval  $[0, 1]$ . The entries  $h_{i,j}$  of  $\mathbf{H}$  belong to an alphabet  $\Sigma$  on which we do not impose any structure and we only assume that learner is able to distinguish distinct elements of the alphabet.

The game proceeds in  $T$  rounds according to the following protocol. First,  $\mathbf{G} = (\mathbf{L}, \mathbf{H})$  is announced for both players. In each round  $t = 1, 2, \dots, T$ , the learner chooses an action  $I_t \in$

---

\* This work was supported in part by AICML, AITF (formerly iCore and AIF), NSERC and the PASCAL2 Network of Excellence under EC grant no. 216886.

$\{1, 2, \dots, N\}$  and simultaneously, the environment chooses an outcome  $J_t \in \{1, 2, \dots, M\}$ . Then, the learner receives as a feedback the entry  $h_{I_t, J_t}$ . The learner incurs *instantaneous loss*  $\ell_{I_t, J_t}$ , which is *not revealed* to him. The feedback can be thought of as a masked information about the outcome  $J_t$ . In some cases  $h_{I_t, J_t}$  might uniquely determine the outcome, in other cases the feedback might give only partial or no information about the outcome. In this paper, we shall assume that  $J_t$  is chosen randomly from a fixed multinomial distribution.

The learner is scored according to the loss matrix  $\mathbf{L}$ . In round  $t$  the learner incurs an *instantaneous loss* of  $\ell_{I_t, J_t}$ . The goal of the learner is to keep low his *total loss*  $\sum_{t=1}^T \ell_{I_t, J_t}$ . Equivalently, the learner's performance can also be measured in terms of his regret, *i.e.*, the total loss of the learner is compared with the loss of best fixed action in hindsight. The regret is defined as the difference of these two losses.

In general, the regret grows with the number of rounds  $T$ . If the regret is sublinear in  $T$ , the learner is said to be Hannan consistent, and this means that the learner's average per-round loss approaches the average per-round loss of the best action in hindsight.

Piccolboni and Schindelhauer (2001) were one of the first to study the regret of these games. In fact, they have studied the problem without making any probabilistic assumptions about the outcome sequence  $J_t$ . They proved that for any finite game  $(\mathbf{L}, \mathbf{H})$ , either for any algorithm the regret can be  $\Omega(T)$  in the worst case, or there exists an algorithm which has regret  $\tilde{O}(T^{3/4})$  on any outcome sequence<sup>1</sup>. This result was later improved by Cesa-Bianchi et al. (2006) who showed that the algorithm of Piccolboni and Schindelhauer has regret  $O(T^{2/3})$ . Furthermore, they provided an example of a finite game, a variant of label-efficient prediction, for which any algorithm has regret  $\Theta(T^{2/3})$  in the worst case.

However, for many games  $O(T^{2/3})$  is not optimal. For example, games with full feedback (*i.e.*, when the feedback uniquely determines the outcome) can be viewed as a special instance of the problem of learning with expert advice and in this case it is known that the “EWA forecaster” has regret  $O(\sqrt{T})$ ; see *e.g.*, Lugosi and Cesa-Bianchi (2006, Chapter 3). Similarly, for games with “bandit feedback” (*i.e.*, when the feedback determines the instantaneous loss) the INF algorithm (Audibert and Bubeck, 2009) and the Exp3 algorithm (Auer et al., 2002) achieve  $O(\sqrt{T})$  regret as well.<sup>2</sup>

This leaves open the problem of determining the minimax regret (*i.e.*, optimal worst-case regret) of any given game  $(\mathbf{L}, \mathbf{H})$ . A partial progress was made in this direction by Bartók et al. (2010) who characterized (almost) all finite games with  $M = 2$  outcomes. They showed that the minimax regret of any “non-degenerate” finite game with two outcomes falls into one of four categories: zero,  $\tilde{\Theta}(\sqrt{T})$ ,  $\Theta(T^{2/3})$  or  $\Theta(T)$ . They gave a combinatoric-geometric condition on the matrices  $\mathbf{L}, \mathbf{H}$  which determines the category a game belongs to. Additionally, they constructed an efficient algorithm which, for any game, achieves the minimax regret rate associated to the game within poly-logarithmic factor.

In this paper, we consider the same problem, with two exceptions. In pursuing a general result, we will consider *all* finite games. However, at the same time, we will only deal with *stochastic* environments, *i.e.*, when the outcome sequences are generated from a fixed probability distribution in an i.i.d. manner.

---

1. The notations  $\tilde{O}(\cdot)$  and  $\tilde{\Theta}(\cdot)$  hide polylogarithmic factors.

2. We ignore the dependence of regret on the number of actions or any other parameters.

The regret against stochastic environments is defined as the difference between the cumulative loss suffered by the algorithm and that of the action with the lowest expected loss. That is, given an algorithm  $\mathcal{A}$  and a time horizon  $T$ , if the outcomes are generated from a probability distribution  $p$ , the regret is

$$R_T(\mathcal{A}, p) = \sum_{t=1}^T \ell_{I_t, J_t} - \min_{1 \leq i \leq N} \mathbb{E}_p \left[ \sum_{t=1}^T \ell_{i, J_t} \right].$$

In this paper we analyze the *minimax* expected regret (in what follows, minimax regret) of games, defined as

$$R_T(\mathbf{G}) = \inf_{\mathcal{A}} \sup_{p \in \Delta_M} \mathbb{E}_p [R_T(\mathcal{A}, p)].$$

We show that the minimax regret of any finite game falls into four categories: zero,  $\tilde{\Theta}(\sqrt{T})$ ,  $\Theta(T^{2/3})$ , or  $\Theta(T)$ . Accordingly, we call the games *trivial*, *easy*, *hard*, and *hopeless*. We give a simple and efficiently computable characterization of these classes using a geometric condition on  $(\mathbf{L}, \mathbf{H})$ . We provide lower-bounds and algorithms that achieve them within poly-logarithmic factor. Our result is an extension of the result of Bartók et al. (2010) for stochastic environments.

It is clear that any lower bound which holds for stochastic environments must hold for adversarial environments too. On the other hand, algorithms and regret upper bounds for stochastic environments, of course, do not transfer to algorithms and regret upper bounds for the adversarial case. Our characterization is a stepping stone towards understanding the minimax regret of partial monitoring games. In particular, we conjecture that our characterization holds without any change for unrestricted environments.

## 2. Preliminaries

In this section, we introduce our conventions, along with some definitions. By default, all vectors are column vectors. We denote by  $\|v\| = \sqrt{v^\top v}$  the Euclidean norm of a vector  $v$ . For a vector  $v$ , the notation  $v \geq 0$  means that all entries of  $v$  are non-negative, and the notation  $v > 0$  means that all entries are positive. For a matrix  $A$ ,  $\text{Im } A$  denotes its *image space*, *i.e.*, the vector space generated by its columns, and the notation  $\text{Ker } A$  denotes its *kernel*, *i.e.*, the set  $\{x : Ax = 0\}$ .

Consider a game  $\mathbf{G} = (\mathbf{L}, \mathbf{H})$  with  $N$  actions and  $M$  outcomes. That is,  $\mathbf{L} \in \mathbb{R}^{N \times M}$  and  $\mathbf{H} \in \Sigma^{N \times M}$ . For the sake of simplicity and, without loss of generality, we assume that no symbol  $\sigma \in \Sigma$  can be present in two different rows of  $\mathbf{H}$ . The *signal matrix* of an action is defined as follows:

**Definition 1 (Signal matrix)** *Let  $\{\sigma_1, \dots, \sigma_{s_i}\}$  be the set of symbols listed in the  $i^{\text{th}}$  row of  $\mathbf{H}$ . (Thus,  $s_i$  denotes the number of different symbols in row  $i$  of  $\mathbf{H}$ ). The signal matrix  $S_i$  of action  $i$  is defined as an  $s_i \times M$  matrix with entries  $a_{k,j} = \mathbb{I}(h_{i,j} = \sigma_k)$  for  $1 \leq k \leq s_i$  and  $1 \leq j \leq M$ . The signal matrix for a set of actions is defined as the signal matrices of the actions in the set, stacked on top of one another, in the ordering of the actions.*

For an example of a signal matrix, see Section 3.1. We identify the strategy of a stochastic opponent with an element of the probability simplex  $\Delta_M = \{p \in \mathbb{R}^M : p \geq 0, \sum_{j=1}^M p_j = 1\}$ . Note that for any opponent strategy  $p$ , if the learner chooses action  $i$  then the vector  $S_i p \in \mathbb{R}^{s_i}$  is the probability distribution of the observed feedback:  $(S_i p)_k$  is the probability of observing the  $k^{\text{th}}$  symbol.

We denote by  $\ell_i^\top$  the  $i^{\text{th}}$  row of the loss matrix  $\mathbf{L}$  and we call  $\ell_i$  the *loss vector of action  $i$* . We say that action  $i$  is *optimal* under opponent strategy  $p \in \Delta_M$  if for any  $1 \leq j \leq N$ ,  $\ell_i^\top p \leq \ell_j^\top p$ . Action  $i$  is said to be *Pareto-optimal* if there exists an opponent strategy  $p$  such that action  $i$  is optimal under  $p$ . We now define the *cell decomposition* of  $\Delta_M$  induced by  $\mathbf{L}$  (for an example, see Figure 2):

**Definition 2 (Cell decomposition)** *For an action  $i$ , the cell  $C_i$  associated with  $i$  is defined as  $C_i = \{p \in \Delta_M : \text{action } i \text{ is optimal under } p\}$ . The cell decomposition of  $\Delta_M$  is defined as the multiset  $\mathcal{C} = \{C_i : 1 \leq i \leq N, C_i \text{ has positive } (M-1)\text{-dimensional volume}\}$ .*

Actions whose cell is of positive  $(M-1)$ -dimensional volume are called *strongly Pareto-optimal*. Actions that are Pareto-optimal but not strongly Pareto-optimal are called *degenerate*. Note that the cells of the actions are defined with linear inequalities and thus they are convex polytopes. It follows that strongly Pareto-optimal actions are the actions whose cells are  $(M-1)$ -dimensional polytopes. It is also important to note that the cell decomposition is a multiset, since some actions can share the same cell. Nevertheless, if two actions have the same cell of dimension  $(M-1)$ , their loss vectors will necessarily be identical.<sup>3</sup>

We call two cells of  $\mathcal{C}$  *neighbors* if their intersection is an  $(M-2)$ -dimensional polytope. The actions corresponding to these cells will also be called neighbors. Neighborship is not defined for cells outside of  $\mathcal{C}$ . For two neighboring cells  $C_i, C_j \in \mathcal{C}$ , we define the *neighborhood action set*  $A_{i,j} = \{1 \leq k \leq N : C_i \cap C_j \subseteq C_k\}$ . It follows from the definition that actions  $i$  and  $j$  are in  $A_{i,j}$  and thus  $A_{i,j}$  is nonempty. However, one can have more than two actions in the neighborhood action set.

When discussing lower bounds we will need the definition of algorithms. For us, an algorithm  $\mathcal{A}$  is a mapping  $\mathcal{A} : \Sigma^* \rightarrow \{1, 2, \dots, N\}$  which maps past feedback sequences to actions. That the algorithms are deterministic is assumed for convenience. In particular, the lower bounds we prove can be extended to randomized algorithms by conditioning on the internal randomization of the algorithm. Note that the algorithms we design are themselves deterministic.

### 3. Classification of finite partial-monitoring games

In this section we present our main result: we state the theorem that classifies all finite stochastic partial-monitoring games based on how their minimax regret scales with the time horizon. Thanks to the previous section, we are now equipped to define a notion which will play a key role in the classification theorem:

---

3. One could think that actions with identical loss vectors are redundant and that all but one of such actions could be removed without loss of generality. However, since different actions can lead to different observations and thus yield different information, removing the duplicates can be harmful.

**Definition 3 (Observability)** Let  $S$  be the signal matrix for the set of all actions in the game. For actions  $i$  and  $j$ , we say that  $\ell_i - \ell_j$  is globally observable if  $\ell_i - \ell_j \in \text{Im } S^\top$ . Furthermore, if  $i$  and  $j$  are two neighboring actions, then  $\ell_i - \ell_j$  is called locally observable if  $\ell_i - \ell_j \in \text{Im } S_{(i,j)}^\top$ , where  $S_{(i,j)}$  is the signal matrix for the neighborhood action set  $A_{i,j}$ .

As we will see, global observability implies that we can estimate the difference of the expected losses after choosing each action once. Local observability means we only need actions from the neighborhood action set to estimate the difference.

The classification theorem, which is our main result, is the following:

**Theorem 4 (Classification)** Let  $\mathbf{G} = (\mathbf{L}, \mathbf{H})$  be a partial-monitoring game with  $N$  actions and  $M$  outcomes. Let  $\mathcal{C} = \{C_1, \dots, C_k\}$  be its cell decomposition, with corresponding loss vectors  $\ell_1, \dots, \ell_k$ . The game  $\mathbf{G}$  falls into one of the following four categories:

- (a)  $R_T(\mathbf{G}) = 0$  if there exists an action  $i$  with  $C_i = \Delta_M$ . This case is called trivial.
- (b)  $R_T(\mathbf{G}) = \Theta(T)$  if there exist two strongly Pareto-optimal actions  $i$  and  $j$  such that  $\ell_i - \ell_j$  is not globally observable. This case is called hopeless.
- (c)  $R_T(\mathbf{G}) = \tilde{\Theta}(\sqrt{T})$  if it is not trivial and for all pairs of (strongly Pareto-optimal) neighboring actions  $i$  and  $j$ ,  $\ell_i - \ell_j$  is locally observable. These games are called easy.
- (d)  $R_T(\mathbf{G}) = \Theta(T^{2/3})$  if  $\mathbf{G}$  is not hopeless and there exists a pair of neighboring actions  $i$  and  $j$  such that  $\ell_i - \ell_j$  is not locally observable. These games are called hard.

Note that the conditions listed under (a)–(d) are mutually exclusive and cover all finite partial-monitoring games. The only non-obvious implication is that if a game is easy then it cannot be hopeless. The reason this holds is because for any pair of cells  $C_i, C_j$  in  $\mathcal{C}$ , the vector  $\ell_i - \ell_j$  can be expressed as a telescoping sum of the differences of loss vectors of neighboring cells.

The remainder of the paper is dedicated to proving Theorem 4. We start with the simple cases. If there exists an action whose cell covers the whole probability simplex then choosing that action in every round will yield zero regret, proving case (a). The condition in Case (b) is due to Piccolboni and Schindelhauer (2001), who showed that under the condition mentioned there, there is no algorithm that achieves sublinear regret<sup>4</sup>. The upper bound for case (d) is achieved by the FeedExp3 algorithm due to Piccolboni and Schindelhauer (2001), for which a regret bound of  $O(T^{2/3})$  was shown by Cesa-Bianchi et al. (2006). The lower bound for case (c) was proved by Antos et al. (2011). For a visualization of previous results, see Figure 1.

The above assertions help characterize trivial and hopeless games, and show that if a game is not trivial and not hopeless then its minimax regret falls between  $\Omega(\sqrt{T})$  and  $O(T^{2/3})$ . Our contribution in this paper is that we give exact minimax rates (up to logarithmic factors) for these games. To prove the upper bound for case (c), we introduce a new algorithm, which we call BALATON, for “Bandit Algorithm for Loss Annihilation”<sup>5</sup>. This algorithm is presented in Section 4, while its analysis is given in Section 5. The lower bound for case (d) is presented in Section 6.

---

4. Although Piccolboni and Schindelhauer state their theorem for adversarial environments, their proof applies to stochastic environments without any change (which is important for the lower bound part).

5. Balaton is a lake in Hungary. We thank Gergely Neu for suggesting the name.

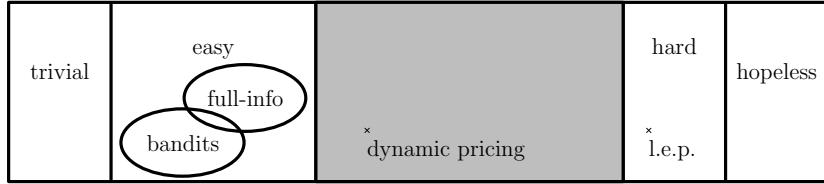


Figure 1: Partial monitoring games and their minimax regret as it was known previously.

The big rectangle denotes the set of all games. Inside the big rectangle, the games are ordered from left to right based on their minimax regret. In the “hard” area, l.e.p. denotes label-efficient prediction. The grey area contains games whose minimax regret is between  $\Omega(\sqrt{T})$  and  $O(T^{2/3})$  but their exact regret rate was unknown. This area is now eliminated, and the dynamic pricing problem is proven to be hard.

### 3.1. Example

In this section, as a corollary of Theorem 4 we show that the discretized dynamic pricing game (see, *e.g.*, Cesa-Bianchi et al. (2006)) is *hard*. Dynamic pricing is a game between a vendor (learner) and a customer (environment). In each round, the vendor sets a price he wants to sell his product at (action), and the costumer sets a maximum price he is willing to buy the product (outcome). If the product is not sold, the vendor suffers some constant loss, otherwise his loss is the difference between the customer’s maximum and his price. The customer never reveals the maximum price and thus the vendor’s only feedback is whether he sold the product or not.

The discretized version of the game with  $N$  actions (and outcomes) is defined by the matrices

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 2 & \cdots & N-1 \\ c & 0 & 1 & \cdots & N-2 \\ \vdots & & \ddots & & \vdots \\ c & \cdots & c & 0 & 1 \\ c & \cdots & \cdots & c & 0 \end{pmatrix} \quad \mathbf{H} = \begin{pmatrix} 1 & \cdots & \cdots & 1 \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix},$$

where  $c$  is a positive constant (see Figure 2 for the cell-decomposition for  $N = 3$ ). It is easy to see that all the actions are strongly Pareto-optimal. Also, after some linear algebra it turns out that the cells underlying the actions have a single common vertex in the interior of the probability simplex. It follows that any two actions are neighbors. On the other hand, if we take two non-consecutive actions  $i$  and  $i'$ ,  $\ell_i - \ell_{i'}$  is not locally observable. For example, the signal matrix for action 1 and action  $N$  is

$$S_{(1,N)} = \begin{pmatrix} 1 & \cdots & 1 & 1 \\ 1 & \cdots & 1 & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix},$$

whereas  $\ell_N - \ell_1 = (c, c-1, \dots, c-N+2, -N+1)^\top$ . It is obvious that  $\ell_N - \ell_1$  is not in the row space of  $S_{(1,N)}$ .

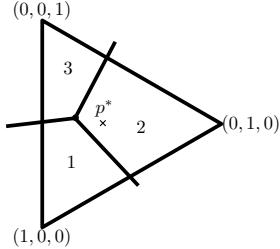


Figure 2: The cell decomposition of the discretized dynamic pricing game with 3 actions.  
If the opponent strategy is  $p^*$ , then action 2 is the optimal action.

#### 4. Balaton: An algorithm for easy games

In this section we present our algorithm that achieves  $\tilde{O}(\sqrt{T})$  expected regret for easy games (case (c) of Theorem 4). The input of the algorithm is the loss matrix  $\mathbf{L}$ , the feedback matrix  $\mathbf{H}$ , the time horizon  $T$  and an error probability  $\delta$ , to be chosen later. Before describing the algorithm, we introduce some notation. We define a graph  $\mathcal{G}$  associated with game  $\mathbf{G}$  the following way. Let the vertex set be the set of cells of the cell decomposition  $\mathcal{C}$  of the probability simplex such that cells  $C_i, C_j \in \mathcal{C}$  share the same vertex when  $C_i = C_j$ . The graph has an edge between vertices whose corresponding cells are neighbors. This graph is connected, since the probability simplex is convex and the cell decomposition covers the simplex.

Recall that for neighboring cells  $C_i, C_j$ , the signal matrix  $S_{(i,j)}$  is defined as the signal matrix for the neighborhood action set  $A_{i,j}$  of cells  $i, j$ . Assuming that the game satisfies the condition of case (c) of Theorem 4, we have that for all neighboring cells  $C_i$  and  $C_j$ ,  $\ell_i - \ell_j \in \text{Im } S_{(i,j)}^\top$ . This means that there exists a *coefficient vector*  $v_{(i,j)}$  such that  $\ell_i - \ell_j = S_{(i,j)}^\top v_{(i,j)}$ . We define the  $k^{\text{th}}$  segment of  $v_{(i,j)}$ , denoted by  $v_{(i,j),k}$ , as the vector of components of  $v_{(i,j)}$  that correspond to the  $k^{\text{th}}$  action in the neighborhood action set. That is, if  $S_{(i,j)}^\top = (S_1^\top \cdots S_r^\top)$ , then  $\ell_i - \ell_j = S_{(i,j)}^\top v_{(i,j)} = \sum_{s=1}^r S_s^\top v_{(i,j),s}$ , where  $S_1, \dots, S_r$  are the signal matrices of the individual actions in  $A_{i,j}$ .

Let  $J_t \in \{1, \dots, M\}$  denote the outcome at time step  $t$ . For  $1 \leq k \leq M$ , let  $e_k \in \mathbb{R}^M$  be the  $k^{\text{th}}$  unit vector. For an action  $i$ , let  $O_i(t) = S_i e_{J_t}$  be the *observation vector* of action  $i$  at time step  $t$ . If the rows of the signal matrix  $S_i$  correspond to symbols  $\sigma_1, \dots, \sigma_{s_i}$  and action  $i$  is chosen at time step  $t$  then the unit vector  $O_i(t)$  indicates which symbol was observed in that time step. Thus,  $O_{I_t}(t)$  holds the same information as the feedback at time  $t$  (recall that  $I_t$  is the action chosen by the learner at time step  $t$ ). From now on, for simplicity, we will assume that the feedback at time step  $t$  is the observation vector  $O_{I_t}(t)$  itself.

The main idea of the algorithm is to successively eliminate actions in an efficient, yet safe manner. When all remaining strongly Pareto optimal actions share the same cell, the elimination phase finishes and from this point, one of the remaining actions is played. During the elimination phase, the algorithm works in rounds. In each round each ‘alive’ Pareto optimal action is played once. The resulting observations are used to estimate the loss-difference between the alive actions. If some estimate becomes sufficiently precise, the action of the pair deemed to be suboptimal is eliminated (possibly together with other

---

**Algorithm 1** BALATON

---

**Input:**  $\mathbf{L}, \mathbf{H}, T, \delta$

**Initialization:**

$$[\mathcal{G}, \mathcal{C}, \{v_{(i,j),k}\}, \{path_{(i,j)}\}, \{(LB_{(i,j)}, UB_{(i,j)}, \sigma_{(i,j)}, R_{(i,j)})\}] \leftarrow \text{INITIALIZE}(\mathbf{L}, \mathbf{H})$$

$$t \leftarrow 0, n \leftarrow 0$$

$$\text{aliveActions} \leftarrow \{1 \leq i \leq N : C_i \cap \text{interior}(\Delta_M) \neq \emptyset\}$$

**main loop**

**while**  $|V_{\mathcal{G}}| > 1$  and  $t < T$  **do**

$$n \leftarrow n + 1$$

**for each**  $i \in \text{aliveActions}$  **do**

$$O_i \leftarrow \text{EXECUTEACTION}(i)$$

$$t \leftarrow t + 1$$

**end for**

**for each** edge  $(i, j)$  in  $\mathcal{G}$ :  $\mu_{(i,j)} \leftarrow \sum_{k \in A_{i,j}} O_k^\top v_{(i,j),k}$  **end for**

**for each** non-adjacent vertex pair  $(i, j)$  in  $\mathcal{G}$ :  $\mu_{(i,j)} \leftarrow \sum_{(k,l) \in path_{(i,j)}} \mu_{(k,l)}$  **end for**

$\text{haveEliminated} \leftarrow \text{false}$

**for each** vertex pair  $(i, j)$  in  $\mathcal{G}$  **do**

$$\hat{\mu}_{(i,j)} \leftarrow \left(1 - \frac{1}{n}\right) \hat{\mu}_{(i,j)} + \frac{1}{n} \mu_{(i,j)}$$

**if**  $\text{BSTOPSTEP}(\hat{\mu}_{(i,j)}, LB_{(i,j)}, UB_{(i,j)}, \sigma_{(i,j)}, R_{(i,j)}, n, 1/2, \delta)$  **then**

$$[\text{aliveActions}, \mathcal{C}, \mathcal{G}] \leftarrow \text{ELIMINATE}(i, j, \text{sgn}(\hat{\mu}_{(i,j)}))$$

$\text{haveEliminated} \leftarrow \text{true}$

**end if**

**end for**

**if**  $\text{haveEliminated}$  **then**

$$\{path_{(i,j)}\} \leftarrow \text{REGENERATEPATHS}(\mathcal{G})$$

**end if**

**end while**

Let  $i$  be a strongly Pareto-optimal action in  $\text{aliveActions}$

**while**  $t < T$  **do**

EXECUTEACTION(i)

$$t \leftarrow t + 1$$

**end while**

---

actions). To determine if an estimate is sufficiently precise, we will use an appropriate stopping rule. A small regret will be achieved by tuning the error probability of the stopping rule appropriately.

The details of the algorithm are as follows: In the preprocessing phase, the algorithm constructs the neighbourhood graph, the signal matrices  $S_{(i,j)}$  assigned to the edges of the graph, the coefficient vectors  $v_{(i,j)}$  and their segment vectors  $v_{(i,j),k}$ . In addition, it constructs a path in the graph connecting any pairs of nodes, and initializes some variables used by the stopping rule.

In the elimination phase, the algorithm runs a loop. In each round of the loop, the algorithm chooses each of the alive actions once and, based on the observations, the estimates  $\hat{\mu}_{(i,j)}$  of the loss-differences  $(\ell_i - \ell_j)^\top p^*$  are updated, where  $p^*$  is the actual opponent

strategy. The algorithm maintains the set  $\mathcal{C}$  of cells of alive actions and their neighborhood graph  $\mathcal{G}$ .

The estimates are calculated as follows. First we calculate estimates for neighboring actions  $(i, j)$ . In round<sup>6</sup>  $n$ , for every action  $k$  in  $A_{i,j}$  let  $O_k$  be the observation vector for action  $k$ . Let  $\mu_{(i,j)} = \sum_{k \in A_{i,j}} O_k^\top v_{(i,j),k}$ . From the local observability condition and the construction of  $v_{(i,j),k}$ , with simple algebra it follows that  $\mu_{(i,j)}$  are unbiased estimates of  $(\ell_i - \ell_j)^\top p^*$  (see Lemma 5). For non-neighboring action pairs, we use telescoping sums: since the graph  $\mathcal{G}$  (induced by the alive actions) stays connected, we can take a path  $i = i_0, i_1, \dots, i_r = j$  in the graph, and the estimate  $\mu_{(i,j)}(n)$  will be the sum of the estimates along the path:  $\sum_{l=1}^r \mu_{(i_{l-1}, i_l)}$ . The estimate of the difference of the expected losses after round  $n$  will be the average  $\hat{\mu}_{(i,j)} = (1/n) \sum_{l=1}^n \mu_{(i,j)}(s)$ , where  $\mu_{(i,j)}(s)$  denotes the estimate for pair  $(i, j)$  computed in round  $s$ .

After updating the estimates, the algorithm decides which actions to eliminate. For each pair of vertices  $i, j$  of the graph, the expected difference of their loss is tested for its sign by the BSTOPSTEP subroutine, based on the estimate  $\hat{\mu}_{(i,j)}$  and its relative error. This subroutine uses a stopping rule based on Bernstein's inequality.

The subroutine's pseudocode is shown as Algorithm 2 and is essentially based on the work by Mnih et al. (2008). The algorithm maintains two values, LB, UB, computed from the supplied sequence of sample means ( $\hat{\mu}$ ) and the deviation bounds

$$c(\sigma, R, n, \delta) = \sigma \sqrt{\frac{2L(\delta, n)}{n}} + \frac{RL(\delta, n)}{3n}, \quad \text{where } L(\delta, n) = \log \left( 3 \frac{p}{p-1} \frac{n^p}{\delta} \right). \quad (1)$$

Here  $p > 1$  is an arbitrarily chosen parameter of the algorithm,  $\sigma$  is a (deterministic) upper bound on the (conditional) variance of the random variables whose common mean  $\mu$  we wish to estimate, while  $R$  is a (deterministic) upper bound on their range. This is a general stopping rule method, which stops when it produced an  $\epsilon$ -relative accurate estimate of the unknown mean. The algorithm is guaranteed to be correct outside of a failure event whose probability is bounded by  $\delta$ .

Algorithm BALATON calls this method with  $\epsilon = 1/2$ . As a result, when BSTOPSTEP returns true, outside of the failure event the sign of the estimate  $\hat{\mu}$  supplied to BALATON will match the sign of the mean to be estimated. The conditions under which the algorithm indeed produces  $\epsilon$ -accurate estimates (with high probability) are given in Lemma 11 (see Appendix), which also states that also with high probability, the time when the algorithm stops is bounded by

$$C \cdot \max \left( \frac{\sigma^2}{\epsilon^2 \mu^2}, \frac{R}{\epsilon |\mu|} \right) \left( \log \frac{1}{\delta} + \log \frac{R}{\epsilon |\mu|} \right),$$

where  $\mu \neq 0$  is the true mean. Note that the choice of  $p$  in (1) influences only  $C$ .

If BSTOPSTEP returns true for an estimate  $\mu_{(i,j)}$ , function ELIMINATE is called. If, say,  $\mu_{(i,j)} > 0$ , this function takes the closed half space  $\{q \in \Delta_M : (\ell_i - \ell_j)^\top q \leq 0\}$  and eliminates *all* actions whose cell lies completely in the half space. The function also drops the vertices from the graph that correspond to eliminated cells. The elimination necessarily

---

6. Note that a round of the algorithm is not the same as the time step  $t$ . In a round, the algorithm chooses each of the alive actions once.

---

**Algorithm 2** Algorithm BSTOPSTEP. Note that, somewhat unusually at least in pseudocodes, the arguments LB, UB are passed by reference, i.e., the algorithm rewrites the values of these arguments (which are thus returned back to the caller).

---

**Input:**  $\hat{\mu}, \text{LB}, \text{UB}, \sigma, R, n, \varepsilon, \delta$   
 $\text{LB} \leftarrow \max(\text{LB}, |\hat{\mu}| - c(\delta, \sigma, R, n))$   
 $\text{UB} \leftarrow \min(\text{UB}, |\hat{\mu}| + c(\delta, \sigma, R, n))$   
**return**  $(1 + \epsilon)\text{LB} < (1 - \epsilon)\text{UB}$

---

concerns all actions with corresponding cell  $C_i$ , and possibly other actions as well. The remaining cells are redefined by taking their intersection with the complement half space  $\{q \in \Delta_M : (\ell_i - \ell_j)^\top q \geq 0\}$ .

By construction, after the elimination phase, the remaining graph is still connected, but some paths used in the round may have lost vertices or edges. For this reason, in the last phase of the round, new paths are constructed for vertex pairs with broken paths.

The main loop of the algorithm continues until either one vertex remains in the graph or the time horizon  $T$  is reached. In the former case, one of the actions corresponding to that vertex is chosen until the time horizon is reached.

## 5. Analysis of the algorithm

In this section we prove that the algorithm described in the previous section achieves  $\tilde{O}(\sqrt{T})$  expected regret.

Let us assume that the outcomes are generated following the probability vector  $p^* \in \Delta_M$ . Let  $j^*$  denote an optimal action, that is, for every  $1 \leq i \leq N$ ,  $\ell_{j^*}^\top p^* \leq \ell_i^\top p^*$ . For every pair of actions  $i, j$ , let  $\alpha_{i,j} = (\ell_i - \ell_j)^\top p^*$  be the expected difference of their instantaneous loss. The expected regret of the algorithm can be rewritten as

$$\mathbb{E} \left[ \sum_{t=1}^T \ell_{I_t, J_t} - \min_{1 \leq i \leq N} \mathbb{E} \left[ \sum_{t=1}^T \ell_{i, J_t} \right] \right] = \sum_{i=1}^N \mathbb{E} [\tau_i] \alpha_{i, j^*}, \quad (2)$$

where  $\tau_i$  is the number of times action  $i$  is chosen by the algorithm.

Throughout the proof, the value that BALATON assigns to a variable  $x$  in round  $n$  will be denoted by  $x(n)$ . Further, for  $1 \leq k \leq N$ , we introduce the i.i.d. random sequence  $(J_k(n))_{n \geq 1}$ , taking values on  $\{1, \dots, M\}$ , with common multinomial distribution satisfying,  $\mathbb{P}[J_k(n) = j] = p_j^*$ . Clearly, a statistically equivalent model to the one where  $(J_t)$  is an i.i.d. sequence with multinomial  $p^*$  is when  $(J_t)$  is defined through

$$J_t = J_{I_t} \left( \sum_{s=1}^t \mathbb{I}(I_s = I_t) \right). \quad (3)$$

Note that this claim holds, independently of the algorithm generating the actions,  $I_t$ . Therefore, in what follows, we assume that the outcome sequence is generated through (3). As we will see, this construction significantly simplifies subsequent steps of the proof. In particular, the construction will be very convenient since if action  $k$  is selected by our algorithm in the  $n^{\text{th}}$  elimination round then the outcome obtained in response is going to be

$O_k(n) = S_k u_k(n)$ , where  $u_k(n) = e_{J_k(n)}$ . (This holds because in the elimination rounds all alive actions are tried exactly once by BALATON.)

Let  $(\mathcal{F}_n)_n$  be the filtration defined as  $\mathcal{F}_n = \sigma(u_k(m); 1 \leq k \leq N, 1 \leq m \leq n)$ . We also introduce the notations  $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_n]$  and  $\text{Var}_n(\cdot) = \text{Var}(\cdot | \mathcal{F}_n)$ , the conditional expectation and conditional variance operators corresponding to  $\mathcal{F}_n$ . Note that  $\mathcal{F}_n$  contains the information known to BALATON (and more) at the end of the elimination round  $n$ . Our first (trivial) observation is that  $\mu_{(i,j)}(n)$ , the estimate of  $\alpha_{i,j}$  obtained in round  $n$  is  $\mathcal{F}_n$ -measurable. The next lemma establishes that, furthermore,  $\mu_{(i,j)}(n)$  is an unbiased estimate of  $\alpha_{i,j}$ :

**Lemma 5** *For any  $n \geq 1$  and  $i, j$  such that  $C_i, C_j \in \mathcal{C}$ ,  $\mathbb{E}_{n-1}[\mu_{(i,j)}(n)] = \alpha_{i,j}$ .*

**Proof** Consider first the case when actions  $i$  and  $j$  are neighbors. In this case,

$$\mu_{(i,j)}(n) = \sum_{k \in A_{i,j}} O_k(n)^\top v_{(i,j),k} = \sum_{k \in A_{i,j}} (S_k u_k(n))^\top v_{(i,j),k} = \sum_{k \in A_{i,j}} u_k(n)^\top S_k^\top v_{(i,j),k},$$

and thus

$$\begin{aligned} \mathbb{E}_{n-1}[\mu_{(i,j)}(n)] &= \sum_{k \in A_{i,j}} \mathbb{E}_{n-1}[u_k(n)^\top] S_k^\top v_{(i,j),k} = p^{*\top} \sum_{k \in A_{i,j}} S_k^\top v_{(i,j),k} = p^{*\top} S_{(i,j)}^\top v_{(i,j)} \\ &= p^{*\top} (\ell_i - \ell_j) = \alpha_{i,j}. \end{aligned}$$

For non-adjacent  $i$  and  $j$ , we have a telescoping sum:

$$\begin{aligned} \mathbb{E}_{n-1}[\mu_{(i,j)}(n)] &= \sum_{k=1}^r \mathbb{E}_{n-1}[\mu_{(i_{k-1}, i_k)}(n)] \\ &= p^{*\top} (\ell_{i_0} - \ell_{i_1} + \ell_{i_1} - \ell_{i_2} + \cdots + \ell_{i_{r-1}} - \ell_{i_r}) = \alpha_{i,j}, \end{aligned}$$

where  $i = i_0, i_1, \dots, i_r = j$  is the path the algorithm uses in round  $n$ , known at the end of round  $n-1$ .  $\blacksquare$

**Lemma 6** *The conditional variance of  $\mu_{(i,j)}(n)$ ,  $\text{Var}_{n-1}(\mu_{(i,j)}(n))$ , is upper bounded by  $V = 2 \sum_{\{i,j \text{ neighbors}\}} \|v_{(i,j)}\|_2^2$ .*

**Proof** For neighboring cells  $i, j$ , we write

$$\begin{aligned} \mu_{(i,j)}(n) &= \sum_{k \in A_{i,j}} O_k(n)^\top v_{(i,j),k} \quad \text{and thus} \\ \text{Var}_{n-1}(\mu_{(i,j)}(n)) &= \text{Var}_{n-1} \left( \sum_{k \in A_{i,j}} O_k(n)^\top v_{(i,j),k} \right) \\ &= \sum_{k \in A_{i,j}} \mathbb{E}_{n-1} \left[ v_{(i,j),k}^\top (O_k(n) - \mathbb{E}_{n-1}[O_k(n)]) (O_k(n) - \mathbb{E}_{n-1}[O_k(n)])^\top v_{(i,j),k} \right] \\ &\leq \sum_{k \in A_{i,j}} \|v_{(i,j),k}\|_2^2 \mathbb{E}_{n-1} [\|O_k(n) - \mathbb{E}_{n-1}[O_k(n)]\|_2^2] \\ &\leq \sum_{k \in A_{i,j}} \|v_{(i,j),k}\|_2^2 = \|v_{(i,j)}\|_2^2, \end{aligned} \tag{4}$$

where in (4) we used that  $O_k(n)$  is a unit vector and  $\mathbb{E}_{n-1}[O_k(n)]$  is a probability vector.

For  $i, j$  non-neighboring cells, let  $i = i_0, i_1, \dots, i_r = j$  the path used for the estimate in round  $n$ . Then  $\mu_{(i,j)}(n)$  can be written as

$$\mu_{(i,j)}(n) = \sum_{s=1}^r \mu_{(i_{s-1}, i_s)}(n) = \sum_{s=1}^r \sum_{k \in A_{i_{s-1}, i_s}} O_k(n)^\top v_{(i_{s-1}, i_s), k}.$$

It is not hard to see that an action can only be in at most two neighborhood action sets in the path and so the double sum can be rearranged as

$$\sum_{k \in \bigcup A_{i_{s-1}, i_s}} O_k(n)^\top (v_{(i_{s-1}, i_s), k} + v_{(i_s, i_{s+1}), k}),$$

and thus  $\text{Var}_{n-1}(\mu_{(i,j)}(n)) \leq 2 \sum_{s=1}^r \|v_{(i_{s-1}, i_s)}\|_2^2 \leq 2 \sum_{\{i,j \text{ neighbors}\}} \|v_{(i,j)}\|_2^2$ . ■

**Lemma 7** *The range of the estimates  $\mu_{(i,j)}(n)$  is upper bounded by  $R = \sum_{\{i,j \text{ neighbors}\}} \|v_{(i,j)}\|_1$ .*

**Proof** The bound trivially follows from the definition of the estimates. ■

Let  $\delta$  be the confidence parameter used in BSTOPSTEP. Since, according to Lemmas 5, 6 and 7,  $(\mu_{(i,j)})$  is a “shifted” martingale difference sequence with conditional mean  $\alpha_{i,j}$ , bounded conditional variance and range, we can apply Lemma 11 stated in the Appendix. By the union bound, the probability that any of the confidence bounds fails during the game is at most  $N^2\delta$ . Thus, with probability at least  $1 - N^2\delta$ , if BSTOPSTEP returns true for a pair  $(i, j)$  then  $\text{sgn}(\alpha_{i,j}) = \text{sgn}(\mu_{(i,j)})$  and the algorithm eliminates all the actions whose cell is contained in the closed half space defined by  $\mathcal{H} = \{p : \text{sgn}(\alpha_{i,j})p^\top (\ell_i - \ell_j) \leq 0\}$ . By definition  $\alpha_{i,j} = (\ell_i - \ell_j)^\top p^*$ . Thus  $p^* \notin \mathcal{H}$  and none of the eliminated actions can be optimal under  $p^*$ .

From Lemma 11 we also see that, with probability at least  $1 - N^2\delta$ , the number of times  $\tau_i^*$  the algorithm experiments with a suboptimal action  $i$  during the elimination phase is bounded by

$$\tau_i^* \leq \frac{c(\mathbf{G})}{\alpha_{i,j^*}^2} \log \frac{R}{\delta \alpha_{i,j^*}} = T_i, \quad (5)$$

where  $c(\mathbf{G}) = C(V + R)$  is a problem dependent constant.

The following lemma, the proof of which can be found in the Appendix, shows that degenerate actions will be eliminated in time.

**Lemma 8** *Let action  $i$  be a degenerate action. Let  $A_i = \{j : C_j \in \mathcal{C}, C_i \subset C_j\}$ . The following two statements hold:*

1. *If any of the actions in  $A_i$  is eliminated, then action  $i$  is eliminated as well.*
2. *There exists an action  $k_i \in A_i$  such that  $\alpha_{k_i, j^*} \geq \alpha_{i, j^*}$ .*

An immediate implication of the first claim of the lemma is that if action  $k_i$  gets eliminated then action  $i$  gets eliminated as well, that is, the number of times action  $i$  is chosen cannot be greater than that of action  $k_i$ . Hence,  $\tau_i^* \leq \tau_{k_i}^*$ .

Let  $\mathcal{E}$  be the complement of the failure event underlying the stopping rules. As discussed earlier,  $\mathbb{P}(\mathcal{E}^c) \leq N^2\delta$ . Note that on  $\mathcal{E}$ , i.e., when the stopping rules do not fail, no suboptimal action can remain for the final phase. Hence,  $\tau_i \mathbb{I}(\mathcal{E}) \leq \tau_i^* \mathbb{I}(\mathcal{E})$ , where  $\tau_i$  is the number of times action  $i$  is chosen by the algorithm. To upper bound the expected regret we continue from (2) as

$$\begin{aligned}
\sum_{i=1}^N \mathbb{E}[\tau_i] \alpha_{i,j^*} &= \sum_{i=1}^N \mathbb{E}[\mathbb{I}(\mathcal{E})\tau_i] \alpha_{i,j^*} + \mathbb{P}(\mathcal{E}^c)T \quad (\text{because } \sum_{i=1}^N \tau_i = T \text{ and } 0 \leq \alpha_{i,j^*} \leq 1) \\
&\leq \sum_{i=1}^N \mathbb{E}[\mathbb{I}(\mathcal{E})\tau_i^*] \alpha_{i,j^*} + N^2\delta T \\
&\leq \sum_{i: C_i \in \mathcal{C}} \mathbb{E}[\mathbb{I}(\mathcal{E})\tau_i^*] \alpha_{i,j^*} + \sum_{i: C_i \notin \mathcal{C}} \mathbb{E}[\mathbb{I}(\mathcal{E})\tau_i^*] \alpha_{i,j^*} + N^2\delta T \\
&\leq \sum_{i: C_i \in \mathcal{C}} \mathbb{E}[\mathbb{I}(\mathcal{E})\tau_i^*] \alpha_{i,j^*} + \sum_{i: C_i \notin \mathcal{C}} \mathbb{E}[\mathbb{I}(\mathcal{E})\tau_{k_i}^*] \alpha_{k_i,j^*} + N^2\delta T \quad (\text{by Lemma 8}) \\
&\leq \sum_{i: C_i \in \mathcal{C}} T_i \alpha_{i,j^*} + \sum_{i: C_i \notin \mathcal{C}} T_{k_i} \alpha_{k_i,j^*} + N^2\delta T \\
&\leq \sum_{\substack{i: C_i \in \mathcal{C} \\ \alpha_{i,j^*} \geq \alpha_0}} T_i \alpha_{i,j^*} + \sum_{\substack{i: C_i \notin \mathcal{C} \\ \alpha_{k_i,j^*} \geq \alpha_0}} T_{k_i} \alpha_{k_i,j^*} + (\alpha_0 + N^2\delta) T \\
&\leq c(\mathbf{G}) \left( \sum_{\substack{i: C_i \in \mathcal{C} \\ \alpha_{i,j^*} \geq \alpha_0}} \frac{\log \frac{R}{\delta \alpha_{i,j^*}}}{\alpha_{i,j^*}} + \sum_{\substack{i: C_i \notin \mathcal{C} \\ \alpha_{k_i,j^*} \geq \alpha_0}} \frac{\log \frac{R}{\delta \alpha_{k_i,j^*}}}{\alpha_{k_i,j^*}} \right) + (\alpha_0 + N^2\delta) T \\
&\leq c(\mathbf{G}) N \frac{\log \frac{R}{\delta \alpha_0}}{\alpha_0} + (\alpha_0 + N^2\delta) T,
\end{aligned}$$

The above calculation holds for any value of  $\alpha_0 > 0$ . Setting

$$\alpha_0 = \sqrt{\frac{c(\mathbf{G})N}{T}} \quad \text{and} \quad \delta = \sqrt{\frac{c(\mathbf{G})}{TN^3}}, \quad \text{we get}$$

$$\mathbb{E}[R_T] \leq \sqrt{c(\mathbf{G})NT} \log \left( \frac{RTN^2}{c(\mathbf{G})} \right).$$

In conclusion, if we run BALATON with parameter  $\delta = \sqrt{\frac{c(\mathbf{G})}{TN^3}}$ , the algorithm suffers regret of  $\tilde{O}(\sqrt{T})$ , finishing the proof.

## 6. A lower bound for hard games

In this section we prove that for any game that satisfies the condition of Case (d) of Theorem 4, the minimax regret is of  $\Omega(T^{2/3})$ .

**Theorem 9** Let  $\mathbf{G} = (\mathbf{L}, \mathbf{H})$  be an  $N$  by  $M$  partial-monitoring game. Assume that there exist two neighboring actions  $i$  and  $j$  such that  $\ell_i - \ell_j \notin \text{Im } S_{(i,j)}^\top$ . Then there exists a problem dependent constant  $c(\mathbf{G})$  such that for any algorithm  $\mathcal{A}$  and time horizon  $T$  there exists an opponent strategy  $p$  such that the expected regret satisfies

$$\mathbb{E}[R_T(\mathcal{A}, p)] \geq c(\mathbf{G})T^{2/3}.$$

**Proof** Without loss of generality we can assume that the two neighbor cells in the condition are  $C_1$  and  $C_2$ . Let  $C_3 = C_1 \cap C_2$ . For  $i = 1, 2, 3$ , let  $A_i$  be the set of actions associated with cell  $C_i$ . Note that  $A_3$  may be the empty set. Let  $A_4 = A \setminus (A_1 \cup A_2 \cup A_3)$ . By our convention for naming loss vectors,  $\ell_1$  and  $\ell_2$  are the loss vectors for  $C_1$  and  $C_2$ , respectively. Let  $\mathcal{L}_3$  collect the loss vectors of actions which lie on the open segment connecting  $\ell_1$  and  $\ell_2$ . It is easy to see that  $\mathcal{L}_3$  is the set of loss vectors that correspond to the cell  $C_3$ . We define  $\mathcal{L}_4$  as the set of all the other loss vectors. For  $i = 1, 2, 3, 4$ , let  $k_i = |A_i|$ .

Let  $S = S_{i,j}$  the signal matrix of the neighborhood action set of  $C_1$  and  $C_2$ . It follows from the assumption of the theorem that  $\ell_2 - \ell_1 \notin \text{Im}(S^\top)$ . Thus,  $\{\rho(\ell_2 - \ell_1) : \rho \in \mathbb{R}\} \not\subset \text{Im}(S^\top)$ , or equivalently,  $(\ell_2 - \ell_1)^\perp \not\supset \text{Ker } S$ , where we used that  $(\text{Im } M)^\perp = \text{Ker}(M^\top)$ . Thus, there exists a vector  $v$  such that  $v \in \text{Ker } S$  and  $(\ell_2 - \ell_1)^\top v \neq 0$ . By scaling we can assume that  $(\ell_2 - \ell_1)^\top v = 1$ . Note that since  $v \in \text{Ker } S$  and the rowspace of  $S$  contains the vector  $(1, 1, \dots, 1)$ , the coordinates of  $v$  sum up to zero.

Let  $p_0$  be an arbitrary probability vector in the relative interior of  $C_3$ . It is easy to see that for any  $\varepsilon > 0$  small enough,  $p_1 = p_0 + \varepsilon v \in C_1 \setminus C_2$  and  $p_2 = p_0 - \varepsilon v \in C_2 \setminus C_1$ .

Let us fix a deterministic algorithm  $\mathcal{A}$  and a time horizon  $T$ . For  $i = 1, 2$ , let  $R_T^{(i)}$  denote the expected regret of the algorithm under opponent strategy  $p_i$ . For  $i = 1, 2$  and  $j = 1, \dots, 4$ , let  $N_j^i$  denote the expected number of times the algorithm chooses an action from  $A_j$ , assuming the opponent plays strategy  $p_i$ .

From the definition of  $\mathcal{L}_3$  we know that for any  $\ell \in \mathcal{L}_3$ ,  $\ell - \ell_1 = \eta_\ell(\ell_2 - \ell_1)$  and  $\ell - \ell_2 = (1 - \eta_\ell)(\ell_1 - \ell_2)$  for some  $0 < \eta_\ell < 1$ . Let  $\lambda_1 = \min_{\ell \in \mathcal{L}_3} \eta_\ell$  and  $\lambda_2 = \min_{\ell \in \mathcal{L}_3} (1 - \eta_\ell)$  and  $\lambda = \min(\lambda_1, \lambda_2)$  if  $\mathcal{L}_3 \neq \emptyset$  and let  $\lambda = 1/2$ , otherwise. Finally, let  $\beta_i = \min_{\ell \in \mathcal{L}_4} (\ell - \ell_i)^\top p_i$  and  $\beta = \min(\beta_1, \beta_2)$ . Note that  $\lambda, \beta > 0$ .

As the first step of the proof, we lower bound the expected regret  $R_T^{(1)}$  and  $R_T^{(2)}$  in terms of the values  $N_j^i, \varepsilon, \lambda$  and  $\beta$ :

$$\begin{aligned} R_T^{(1)} &\geq N_2^1 \underbrace{(\ell_2 - \ell_1)^\top p_1}_{\varepsilon} + N_3^1 \lambda (\ell_2 - \ell_1)^\top p_1 + N_4^1 \beta \geq \lambda(N_2^1 + N_3^1)\varepsilon + N_4^1 \beta, \\ R_T^{(2)} &\geq N_1^2 \underbrace{(\ell_1 - \ell_2)^\top p_2}_{\varepsilon} + N_3^2 \lambda (\ell_1 - \ell_2)^\top p_2 + N_4^2 \beta \geq \lambda(N_1^2 + N_3^2)\varepsilon + N_4^2 \beta. \end{aligned} \quad (6)$$

For the next step, we need the following lemma.

**Lemma 10** There exists a (problem dependent) constant  $c$  such that the following inequalities hold:

$$\begin{aligned} N_1^2 &\geq N_1^1 - cT\varepsilon\sqrt{N_4^1}, & N_3^2 &\geq N_3^1 - cT\varepsilon\sqrt{N_4^1}, \\ N_2^1 &\geq N_2^2 - cT\varepsilon\sqrt{N_4^2}, & N_3^1 &\geq N_3^2 - cT\varepsilon\sqrt{N_4^2}. \end{aligned}$$

**Proof** (Lemma 10) For any  $1 \leq t \leq T$ , let  $f^t = (f_1, \dots, f_t) \in \Sigma^t$  be a feedback sequence up to time step  $t$ . For  $i = 1, 2$ , let  $p_i^*$  be the probability mass function of feedback sequences of length  $T - 1$  under opponent strategy  $p_i$  and algorithm  $\mathcal{A}$ . We start by upper bounding the difference between values under the two opponent strategies. For  $i \neq j \in \{1, 2\}$  and  $k \in \{1, 2, 3\}$ ,

$$\begin{aligned} N_k^i - N_k^j &= \sum_{f^{T-1}} (p_i^*(f^{T-1}) - p_j^*(f^{T-1})) \sum_{t=0}^{T-1} \mathbb{I}(\mathcal{A}(f^t) \in A_k) \\ &\leq \sum_{\substack{f^{T-1}: \\ p_i^*(f^{T-1}) - p_j^*(f^{T-1}) \geq 0}} (p_i^*(f^{T-1}) - p_j^*(f^{T-1})) \sum_{t=0}^{T-1} \mathbb{I}(\mathcal{A}(f^t) \in A_k) \\ &\leq T \sum_{\substack{f^{T-1}: \\ p_i^*(f^{T-1}) - p_j^*(f^{T-1}) \geq 0}} p_i^*(f^{T-1}) - p_j^*(f^{T-1}) = \frac{T}{2} \|p_1^* - p_2^*\|_1 \\ &\leq T \sqrt{\text{KL}(p_1^* || p_2^*)/2}, \end{aligned} \tag{7}$$

where  $\text{KL}(\cdot || \cdot)$  denotes the Kullback-Leibler divergence and  $\|\cdot\|_1$  is the  $L_1$ -norm. The last inequality follows from Pinsker's inequality (Cover and Thomas, 2006). To upper bound  $\text{KL}(p_1^* || p_2^*)$  we use the chain rule for KL-divergence. By overloading  $p_i^*$  so that  $p_i^*(f^{t-1})$  denotes the probability of feedback sequence  $f^{t-1}$  under opponent strategy  $p_i$  and algorithm  $\mathcal{A}$ , and  $p_i^*(f_t | f^{t-1})$  denotes the conditional probability of feedback  $f_t \in \Sigma$  given that the past feedback sequence was  $f^{t-1}$ , again under  $p_i$  and  $\mathcal{A}$ . With this notation we have

$$\begin{aligned} \text{KL}(p_1^* || p_2^*) &= \sum_{t=1}^{T-1} \sum_{f^{t-1}} p_1^*(f^{t-1}) \sum_{f_t} p_1^*(f_t | f^{t-1}) \log \frac{p_1^*(f_t | f^{t-1})}{p_2^*(f_t | f^{t-1})} \\ &= \sum_{t=1}^{T-1} \sum_{f^{t-1}} p_1^*(f^{t-1}) \sum_{i=1}^4 \mathbb{I}(\mathcal{A}(f^{t-1}) \in A_i) \sum_{f_t} p_1^*(f_t | f^{t-1}) \log \frac{p_1^*(f_t | f^{t-1})}{p_2^*(f_t | f^{t-1})} \end{aligned} \tag{8}$$

Let  $a_{f_t}^\top$  be the row of  $S$  that corresponds to the feedback symbol  $f_t$ .<sup>7</sup> Assume  $k = \mathcal{A}(f^{t-1})$ . If the feedback set of action  $k$  does not contain  $f_t$  then trivially  $p_i^*(f_t | f^{t-1}) = 0$  for  $i = 1, 2$ . Otherwise  $p_i^*(f_t | f^{t-1}) = a_{f_t}^\top p_i$ . Since  $p_1 - p_2 = 2\varepsilon v$  and  $v \in \text{Ker } S$ , we have  $a_{f_t}^\top v = 0$  and thus, if the choice of the algorithm is in either  $A_1, A_2$  or  $A_3$ , then  $p_1^*(f_t | f^{t-1}) = p_2^*(f_t | f^{t-1})$ . It follows that the inequality chain can be continued from (8) by writing

$$\begin{aligned} \text{KL}(p_1^* || p_2^*) &\leq \sum_{t=1}^{T-1} \sum_{f^{t-1}} p_1^*(f^{t-1}) \mathbb{I}(\mathcal{A}(f^{t-1}) \in A_4) \sum_{f_t} p_1^*(f_t | f^{t-1}) \log \frac{p_1^*(f_t | f^{t-1})}{p_2^*(f_t | f^{t-1})} \\ &\leq c_1 \varepsilon^2 \sum_{t=1}^{T-1} \sum_{f^{t-1}} p_1^*(f^{t-1}) \mathbb{I}(\mathcal{A}(f^{t-1}) \in A_4) \\ &\leq c_1 \varepsilon^2 N_4^1. \end{aligned} \tag{9}$$

---

7. Recall that we assumed that different actions have different feedback symbols, and thus a row of  $S$  corresponding to a symbol is unique.

In (9) we used Lemma 12 (see Appendix) to upper bound the KL-divergence of  $p_1$  and  $p_2$ . Flipping  $p_1^*$  and  $p_2^*$  in (7) we get the same result with  $N_4^2$ . Reading together with the bound in (7) we get all the desired inequalities. ■

Now we can continue lower bounding the expected regret. Let  $r = \operatorname{argmin}_{i \in \{1,2\}} N_4^i$ . It is easy to see that for  $i = 1, 2$  and  $j = 1, 2, 3$ ,

$$N_j^i \geq N_j^r - c_2 T \varepsilon \sqrt{N_4^r}.$$

If  $i \neq r$  then this inequality is one of the inequalities from Lemma 10. If  $i = r$  then it is a trivial lower bounding by subtracting a positive value. From (6) we have

$$\begin{aligned} R_T^{(i)} &\geq \lambda(N_{3-i}^i + N_3^i)\varepsilon + N_4^i\beta \\ &\geq \lambda(N_{3-i}^r - c_2 T \varepsilon \sqrt{N_4^r} + N_3^r - c_2 T \varepsilon \sqrt{N_4^r})\varepsilon + N_4^r\beta \\ &= \lambda(N_{3-i}^r + N_3^r - 2c_2 T \varepsilon \sqrt{N_4^r})\varepsilon + N_4^r\beta. \end{aligned}$$

Now assume that, at the beginning of the game, the opponent randomly chooses between strategies  $p_1$  and  $p_2$  with equal probability. The the expected regret of the algorithm is lower bounded by

$$\begin{aligned} R_T &= \frac{1}{2} (R_T^{(1)} + R_T^{(2)}) \\ &\geq \frac{1}{2} \lambda(N_1^r + N_2^r + 2N_3^r - 4c_2 T \varepsilon \sqrt{N_4^r})\varepsilon + N_4^r\beta \\ &\geq \frac{1}{2} \lambda(N_1^r + N_2^r + N_3^r - 4c_2 T \varepsilon \sqrt{N_4^r})\varepsilon + N_4^r\beta \\ &= \frac{1}{2} \lambda(T - N_4^r - 4c_2 T \varepsilon \sqrt{N_4^r})\varepsilon + N_4^r\beta. \end{aligned}$$

Choosing  $\varepsilon = c_3 T^{-1/3}$  we get

$$\begin{aligned} R_T &\geq \frac{1}{2} \lambda c_3 T^{2/3} - \frac{1}{2} \lambda N_4^r c_3 T^{-1/3} - 2\lambda c_2 c_3^2 T^{1/3} \sqrt{N_4^r} + N_4^r\beta \\ &\geq T^{2/3} \left( \left( \beta - \frac{1}{2} \lambda c_3 \right) \frac{N_4^r}{T^{2/3}} - 2\lambda c_2 c_3^2 \sqrt{\frac{N_4^r}{T^{2/3}}} + \frac{1}{2} \lambda c_3 \right) \\ &= T^{2/3} \left( \left( \beta - \frac{1}{2} \lambda c_3 \right) x^2 - 2\lambda c_2 c_3^2 x + \frac{1}{2} \lambda c_3 \right), \end{aligned}$$

where  $x = \sqrt{N_4^r / T^{2/3}}$ . Now we see that  $c_3 > 0$  can be chosen to be small enough, independently of  $T$  so that, for any choice of  $x$ , the quadratic expression in the parenthesis is bounded away from zero, and simultaneously,  $\varepsilon$  is small enough so that the threshold condition in Lemma 12 is satisfied, completing the proof of Theorem 9. ■

## 7. Discussion

In this we paper we classified all finite partial-monitoring games under stochastic environments, based on their minimax regret. We conjecture that our results extend to non-stochastic environments. This is the major open question that remains to be answered.

One question which we did not discuss so far is the computational efficiency of our algorithm. The issue is twofold. The first computational question is how to efficiently decide which of the four classes a given game  $(\mathbf{L}, \mathbf{H})$  belongs to. The second question is the computational efficiency of BALATON for a fixed easy game. Fortunately, in both cases an efficient implementation is possible, *i.e.*, in polynomial time by using a linear program solver (*e.g.*, the ellipsoid method (Papadimitriou and Steiglitz, 1998)).

Another interesting open question is to investigate the dependence of regret on quantities other than  $T$  such as the number of actions, the number of outcomes, and more generally the structure of the loss and feedback matrices.

Finally, let us note that our results can be extended to a more general framework, similar to that of Pallavi et al. (2011), in which a game with  $N$  actions and  $M$ -dimensional outcome space is defined as a tuple  $\mathbf{G} = (\mathbf{L}, S_1, \dots, S_N)$ . The loss matrix is  $\mathbf{L} \in \mathbb{R}^{N \times M}$  as before, but the outcome and the feedback are defined differently. The outcome  $y$  is an arbitrary vector from a bounded subset of  $\mathbb{R}^M$  and the feedback received by the learner upon choosing action  $i$  is  $O_i = S_i y$ .

## References

- Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 263–273. Citeseer, 2008.
- Alekh Agarwal, Peter Bartlett, and Max Dama. Optimal allocation strategies for the dark pool problem. In *13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010), May 12-15, 2010, Chia Laguna Resort, Sardinia, Italy*, 2010.
- András Antos, Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games, 2011. <http://arxiv.org/abs/1102.2041>.
- Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.
- Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.
- Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a Classification of Finite Partial-Monitoring Games. In *Proceedings of the 21st international conference on Algorithmic Learning Theory (ALT 2010)*, pages 224–238. Springer, 2010.
- Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, June 2005.
- Nicoló Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, New York, second edition, 2006.

Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the 16th annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2005)*, page 394. Society for Industrial and Applied Mathematics, 2005.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Proceedings of 44th Annual IEEE Symposium on Foundations of Computer Science 2003 (FOCS 2003)*, pages 594–605. IEEE, 2003.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

Gábor Lugosi and Nicolò Cesa-Bianchi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

V. Mnih. Efficient stopping rules. Master’s thesis, Department of Computing Science, University of Alberta, 2008.

V. Mnih, Cs. Szepesvári, and J.-Y. Audibert. Empirical Bernstein stopping. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, *ICML 2008*, pages 672–679. ACM, 2008.

A. Pallavi, R. Zheng, and Cs. Szepesvári. Sequential learning for optimal monitoring of multi-channel wireless networks. In *INFOCOMM*, 2011.

Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Dover Publications, New York, 1998.

Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Proceedings of the 14th Annual Conference on Computational Learning Theory (COLT 2001)*, pages 208–223. Springer-Verlag, 2001.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of Twentieth International Conference on Machine Learning (ICML 2003)*, 2003.

## Appendix

### Proof (Lemma 8)

1. In an elimination set, we eliminate every action whose cell is contained in a closed half space. Let us assume that  $j \in A_i$  is being eliminated. According to the definition of  $A_i$ ,  $C_i \subset C_j$  and thus  $C_i$  is also contained in the half space.
2. First let us assume that  $p^*$  is not in the affine subspace spanned by  $C_i$ . Let  $p$  be an arbitrary point in the relative interior of  $C_i$ . We define the point  $p' = p + \varepsilon(p - p^*)$ . For a small enough  $\varepsilon > 0$ ,  $p' \in C_k \in A_i$ , and at the same time,  $p' \notin C_i$ . Thus we have

$$\begin{aligned} \ell_k^\top (p + \varepsilon(p - p^*)) &\leq \ell_i^\top (p + \varepsilon(p - p^*)) \\ (1 + \varepsilon)\ell_k^\top p - \varepsilon\ell_k^\top p^* &\leq (1 + \varepsilon)\ell_i^\top p - \varepsilon\ell_i^\top p^* \\ -\varepsilon\ell_k^\top p^* &\leq -\varepsilon\ell_i^\top p^* \\ \ell_k^\top p^* &\geq \ell_i^\top p^* \\ \alpha_{k,j^*} &\geq \alpha_{i,j^*}, \end{aligned}$$

where we used that  $\ell_k^\top p = \ell_i^\top p$ .

For the case when  $p^*$  lies in the affine subspace spanned by  $C_i$ , We take a hyperplane that contains the affine subspace. Then we take an infinite sequence  $(p_n)_n$  such that every element of the sequence is in the same side of the hyperplane,  $p_n \neq p^*$  and the sequence converges to  $p^*$ . Then the statement is true for every element  $p_n$  and, since the value  $\alpha_{r,s}$  is continuous in  $p$ , the limit has the desired property as well. ■

The following lemma concerns the problem of producing an estimate of an unknown mean of some stochastic process with a given relative error bound and with high probability in a sample-efficient manner. The procedure is a simple variation of the one proposed by Mnih et al. (2008). The main differences are that here we deal with martingale difference sequences shifted by an unknown constant, which becomes the common mean, whereas Mnih et al. (2008) considered an i.i.d. sequence. On the other hand, we consider the case when we have a known upper bound on the predictable variance of the process, whereas one of the main contributions of Mnih et al. (2008) was the lifting of this assumption. The proof of the lemma is omitted, as it follows the same lines as the proof of results of Mnih et al. (2008) (the details of these proofs are found in the thesis of (Mnih, 2008)), the only difference being, that here we would need to use Bernstein's inequality for martingales, in place of the empirical Bernstein inequality, which was used by Mnih et al. (2008).

**Lemma 11** *Let  $(\mathcal{F}_t)$  be a filtration on some probability space, and let  $(X_t)$  be an  $\mathcal{F}_t$ -adapted sequence of random variables. Assume that  $(X_t)$  is such that, almost surely, the range of each random variable  $X_t$  is bounded by  $R > 0$ ,  $\mathbb{E}[X_t | \mathcal{F}_{t-1}] = \mu$ , and  $\text{Var}[X_t | \mathcal{F}_{t-1}] \leq \sigma^2$  a.s., where  $R, \mu \neq 0$  and  $\sigma^2$  are non-random constants. Let  $p > 1$ ,  $\epsilon > 0$ ,  $0 < \delta < 1$  and let*

$$L_n = (1 + \varepsilon) \max_{1 \leq t \leq n} \{|\bar{X}_t| - c_t\}, \quad \text{and} \quad U_n = (1 - \varepsilon) \min_{1 \leq t \leq n} \{|\bar{X}_t| + c_t\},$$

where  $c_t = c(\sigma, R, t, \delta)$ , and  $c(\cdot)$  is defined in (1). Define the estimate  $\hat{\mu}_n$  of  $\mu$  as follows:

$$\hat{\mu}_n = \text{sgn}(\bar{X}_n) \frac{(1 + \varepsilon)L_n + (1 - \varepsilon)U_n}{2}.$$

Denote the stopping time  $\tau = \min\{n : L_n \geq U_n\}$ . Then, with probability at least  $1 - \delta$ ,

$$|\hat{\mu}_\tau - \mu| \leq \varepsilon |\mu| \quad \text{and} \quad \tau \leq C \cdot \max\left(\frac{\sigma^2}{\epsilon^2 \mu^2}, \frac{R}{\epsilon |\mu|}\right) \left(\log \frac{1}{\delta} + \log \frac{R}{\epsilon |\mu|}\right),$$

where  $C > 0$  is a universal constant.

**Lemma 12** Fix a probability vector  $p \in \Delta_M$ , and let  $\epsilon \in \mathbb{R}^M$  such that  $p - \epsilon, p + \epsilon \in \Delta_M$  also holds. Then

$$\text{KL}(p - \epsilon || p + \epsilon) = O(\|\epsilon\|_2^2) \quad \text{as } \epsilon \rightarrow 0.$$

The constant and the threshold in the  $O(\cdot)$  notation depends on  $p$ .

**Proof** Since  $p$ ,  $p + \epsilon$ , and  $p - \epsilon$  are all probability vectors, notice that  $|\epsilon(i)| \leq p(i)$  for  $1 \leq i \leq M$ . So if a coordinate of  $p$  is zero then the corresponding coordinate of  $\epsilon$  has to be zero as well. As zero coordinates do not modify the KL divergence, we can assume without loss of generality that all coordinates of  $p$  are positive. Since we are interested only in the case when  $\epsilon \rightarrow 0$ , we can also assume without loss of generality that  $|\epsilon(i)| \leq p(i)/2$ . Also note that the coordinates of  $\epsilon = (p + \epsilon) - \epsilon$  have to sum up to zero. By definition,

$$\text{KL}(p - \epsilon || p + \epsilon) = \sum_{i=1}^M (p(i) - \epsilon(i)) \log \frac{p(i) - \epsilon(i)}{p(i) + \epsilon(i)}.$$

We write the term with the logarithm

$$\log \frac{p(i) - \epsilon(i)}{p(i) + \epsilon(i)} = \log \left(1 - \frac{\epsilon(i)}{p(i)}\right) - \log \left(1 + \frac{\epsilon(i)}{p(i)}\right),$$

so that we can use that, by second order Taylor expansion around 0,  $\log(1-x) - \log(1+x) = -2x + r(x)$ , where  $|r(x)| \leq c|x|^3$  for  $|x| \leq 1/2$  and some  $c > 0$ . Combining these equations, we get

$$\begin{aligned} \text{KL}(p - \epsilon || p + \epsilon) &= \sum_{i=1}^M (p(i) - \epsilon(i)) \left[ -2 \frac{\epsilon(i)}{p(i)} + r\left(\frac{\epsilon(i)}{p(i)}\right) \right] \\ &= \sum_{i=1}^M -2\epsilon(i) + \sum_{i=1}^M 2 \frac{\epsilon^2(i)}{p(i)} + \sum_{i=1}^M (p(i) - \epsilon(i))r\left(\frac{\epsilon(i)}{p(i)}\right). \end{aligned}$$

Here the first term is 0, letting  $\underline{p} = \min_{i \in \{1, \dots, M\}} p(i)$  the second term is bounded by  $2 \sum_{i=1}^M \epsilon^2(i)/\underline{p} = (2/\underline{p})\|\epsilon\|_2^2$ , and the third term is bounded by

$$\begin{aligned} \sum_{i=1}^M (p(i) - \epsilon(i)) \left| r \left( \frac{\epsilon(i)}{p(i)} \right) \right| &\leq c \sum_{i=1}^M \frac{p(i) - \epsilon(i)}{p^3(i)} |\epsilon(i)|^3 \\ &\leq c \sum_{i=1}^M \frac{|\epsilon(i)|}{p^2(i)} \epsilon^2(i) \\ &\leq \frac{c}{2} \sum_{i=1}^M \frac{1}{\underline{p}} \epsilon^2(i) = \frac{c}{2\underline{p}} \|\epsilon\|_2^2. \end{aligned}$$

Hence,  $\text{KL}(p - \epsilon || p + \epsilon) \leq \frac{4+c}{2\underline{p}} \|\epsilon\|_2^2 = \mathcal{O}(\|\epsilon\|_2^2)$ . ■

BARTÓK PÁL SZEPESVÁRI

# Sample Complexity Bounds for Differentially Private Learning

**Kamalika Chaudhuri**

*University of California, San Diego  
9500 Gilman Drive #0404  
La Jolla, CA 92093-0404*

KAMALIKA@CS.UCSD.EDU

**Daniel Hsu**

*Microsoft Research New England  
One Memorial Drive  
Cambridge, MA 02142*

DAHSU@MICROSOFT.COM

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

This work studies the problem of privacy-preserving classification – namely, learning a classifier from sensitive data while preserving the privacy of individuals in the training set. In particular, the learning algorithm is required in this problem to guarantee differential privacy, a very strong notion of privacy that has gained significant attention in recent years.

A natural question to ask is: what is the sample requirement of a learning algorithm that guarantees a certain level of privacy and accuracy? We address this question in the context of learning with infinite hypothesis classes when the data is drawn from a continuous distribution. We first show that even for very simple hypothesis classes, any algorithm that uses a finite number of examples and guarantees differential privacy must fail to return an accurate classifier for at least some unlabeled data distributions. This result is unlike the case with either finite hypothesis classes or discrete data domains, in which distribution-free private learning is possible, as previously shown by Kasiviswanathan et al. (2008).

We then consider two approaches to differentially private learning that get around this lower bound. The first approach is to use prior knowledge about the unlabeled data distribution in the form of a reference distribution  $\mathcal{U}$  chosen independently of the sensitive data. Given such a reference  $\mathcal{U}$ , we provide an upper bound on the sample requirement that depends (among other things) on a measure of closeness between  $\mathcal{U}$  and the unlabeled data distribution. Our upper bound applies to the non-realizable as well as the realizable case. The second approach is to relax the privacy requirement, by requiring only label-privacy – namely, that the only labels (and not the unlabeled parts of the examples) be considered sensitive information. An upper bound on the sample requirement of learning with label privacy was shown by Chaudhuri et al. (2006); in this work, we show a lower bound.

**Keywords:** List of keywords

## 1. Introduction

As increasing amounts of personal data is collected, stored and mined by companies and government agencies, the question of how to learn from sensitive datasets while still maintaining the privacy of individuals in the data has become very important. Over the last few years, the notion of differential privacy (Dwork et al., 2006) has received a significant

amount of attention, and has become the de facto standard for privacy-preserving computation. In this paper, we study the problem of learning a classifier from a dataset, while simultaneously guaranteeing differential privacy of the training data.

The key issue in differentially-private computation is that given a certain amount of resources, there is usually a tradeoff between privacy and utility. In classification, a natural measure of utility is the classification accuracy, and data is a scarce resource. Thus, a key question in differentially-private learning is: how many examples does a learning algorithm need to guarantee a certain level of privacy and accuracy? In this paper, we study this question from an information-theoretic perspective – namely, we are concerned with the sample complexity, and not the computational complexity of the learner.

This question was first considered by Kasiviswanathan et al. (2008), who studied the case of finite hypothesis classes, as well as the case of discrete data domains. They showed that in these two cases, one can obtain any given privacy guarantee and generalization error, regardless of the unlabeled data distribution with a modest increase in the worst-case sample requirement.

In this paper, we consider the sample complexity of differentially private learning in the context of *infinite hypothesis classes on continuous data distributions*. This is a very general class of learning problems, and includes many popular machine-learning tasks such as learning linear classifiers when the examples have real-valued features, which cannot be modeled by finite hypothesis classes or hypothesis classes over discrete data domains.

Surprisingly, we show that the results of Kasiviswanathan et al. (2008) do not extend to infinite hypothesis classes on continuous data distributions. As an example, consider the class of thresholds on the unit interval. This simple learning problem has VC dimension 1, and thus for all unlabeled data distributions, it can be learnt (non-privately) with error  $\epsilon$  given at most  $\tilde{O}(\frac{1}{\epsilon})$  examples<sup>1</sup>. We show that even for this very simple hypothesis class, any algorithm that uses a bounded number of examples and guarantees differential privacy must fail to return an accurate classifier for at least some unlabeled data distributions.

The key intuition behind our proof is that if most of the unlabeled data is concentrated in a small region around the best classifier, then, even slightly perturbing the best classifier will result in a large classification error. As the process of ensuring differential privacy necessarily involves some perturbation – see, for example, Dwork et al. (2006), unless the algorithm has some prior public knowledge about the data distribution, the number of samples required to learn privately grows with growing concentration of the data around the best classifier.

How can we then learn privately in infinite hypothesis classes over continuous data distributions? One approach is to use some prior information about the data distribution that is known independently of the sensitive data. Another approach is to relax the privacy requirements. In this paper, we examine both approaches.

First, we consider the case when the learner has access to some prior information on the unlabeled data. In particular, the learner knows a reference distribution  $\mathcal{U}$  that is close to the unlabeled data distribution. Similar assumptions are common in Bayesian learning, and PAC-Bayes style bounds have also been studied in the learning theory literature, for example, by McAllester (1998).

---

1. Here the  $\tilde{O}$  notation hides factors logarithmic in  $1/\epsilon$

Under this assumption, we provide an algorithm for learning with  $\alpha$ -privacy, excess generalization error  $\epsilon$ , and confidence  $1 - \delta$ , using  $\tilde{O}(d_{\mathcal{U}} \log(\kappa/\epsilon)(\frac{1}{\epsilon^2} + \frac{1}{\epsilon\alpha}))$  samples. Here  $\alpha$  is a privacy parameter (where, lower  $\alpha$  implies a stronger privacy guarantee),  $\mathcal{U}$  is the reference distribution,  $d_{\mathcal{U}}$  is the doubling dimension of its disagreement metric (Bshouty et al., 2009), and  $\kappa$  is a smoothness parameter that we define. The quantity  $d_{\mathcal{U}}$  measures the complexity of the hypothesis class with respect to  $\mathcal{U}$  (see (Bshouty et al., 2009) for a discussion), and we assume that it is finite. The smoothness parameter measures how close the unlabeled data distribution is to  $\mathcal{U}$  (smaller  $\kappa$  means closer), and is motivated by notions of closeness used in Dasgupta (2005) and Freund et al. (1997). Thus the sample requirement of our algorithm grows with increasing distance between  $\mathcal{U}$  and the unlabeled data distribution. Our algorithm works in the non-realizable case, that is, when no hypothesis in the class has zero error; using standard techniques, a slightly better bound of  $\tilde{O}(\frac{d_{\mathcal{U}} \log(\kappa/\epsilon)}{\epsilon\alpha})$  can be obtained in the realizable setting. However, like the results of Kasiviswanathan et al. (2008), our algorithm is computationally inefficient in general.

The main difficulty in reducing the differentially-private learning algorithms of Kasiviswanathan et al. (2008) to infinite hypothesis classes on continuous data distributions is in finding a suitable finite cover of the class with respect to the unlabeled data. This issue is specific to our particular problem: for non-private learning, a finite cover can always be computed based on the (sensitive) data, and for finite hypothesis classes, the entire class is a cover. The main insight behind our upper bound is that when the unlabeled distribution  $\mathcal{D}$  is close to the reference distribution  $\mathcal{U}$ , then a cover of  $\mathcal{U}$  is also a possibly coarser cover of  $\mathcal{D}$ . Since one can compute a private cover of  $\mathcal{U}$  independent of the sensitive data, we simply compute a finer cover of  $\mathcal{U}$ , and learn over this fine cover using standard techniques such as the exponential mechanism (McSherry and Talwar, 2007).

Next we relax the privacy requirement by requiring only *label privacy*. In other words, we assume that the unlabeled part of the examples are not sensitive, and the only private information is the labels. This setting was considered by Chaudhuri et al. (2006). An example when this may be applicable is in predicting income from public demographic information. Here, while the label (income) is private, the demographic information of individuals, such as education, gender, and age may be public.

In this case, we provide lower bounds to characterize the sample requirement of label-private learning. We show two results, based on the value of  $\alpha$  and  $\epsilon$ . For small  $\epsilon$  and  $\alpha$  (that is, for high privacy and accuracy) we show that any learning algorithm for a given hypothesis class that guarantees  $\alpha$ -label privacy and  $\epsilon$  accuracy necessarily requires at least  $\Omega(\frac{d}{\alpha\epsilon})$  examples. Here  $d$  is the doubling dimension of the disagreement metric at a certain scale, and is a measure of the complexity of the hypothesis class on the unlabeled data distribution. This bound holds when the hypothesis class has finite VC dimension. For larger  $\alpha$  and  $\epsilon$ , our bounds are weaker but more general; we show a lower bound of  $\Omega(\frac{d'}{\alpha})$  on the sample requirement that holds for any  $\alpha$  and  $\epsilon$ , and do not require the VC dimension of the hypothesis class to be finite. Here  $d'$  is the doubling dimension of the disagreement metric at a certain scale.

The main idea behind our stronger label privacy lower bounds is to show that differentially private learning algorithms necessarily perform poorly when there is a large set of hypotheses such that every pair in the set labels approximately  $1/\alpha$  examples differently.

We then show that such large sets can be constructed when the doubling dimension of the disagreement metric of the hypothesis class with respect to the data distribution is high.

How do these results fit into the context of non-private learning? For non-private learning, sample requirement bounds based on the doubling dimension of the disagreement metric has been extensively studied by (Bshouty et al., 2009); in the realizable case, they show an upper bound of  $\tilde{O}(\frac{\bar{d}}{\epsilon})$  for learning with accuracy  $\epsilon$ , where  $\bar{d}$  is again the doubling dimension of the disagreement metric at a certain scale. These bounds are incomparable to ours in general, as the doubling dimensions in the two bounds are with respect to different scales; however, we can compare them for hypothesis classes and data distributions for which the doubling dimension of the disagreement metric is equal at all scales. An example is learning halfspaces with respect to the uniform distribution on the sphere. For such problems, on the upper bound side, we need a factor of  $O(\frac{\log(\kappa/\epsilon)}{\alpha})$  times more examples to learn with  $\alpha$ -privacy. On the other hand, our lower bounds indicate that for small  $\alpha$  and  $\epsilon$ , even if we only want  $\alpha$ -label privacy, the sample requirement can be as much as a factor of  $\Omega(\frac{1}{\alpha})$  more than the upper bound for non-private learning.

Finally, one may be tempted to think that we can always discretize a data domain or a hypothesis class, and therefore in practice we are likely to only learn finite hypothesis classes or over discrete data domains. However, there are several issues with such discretization. First, if we discretize either the hypothesis class or the data, then the sample requirement of differentially private learning algorithms will grow as the discretization grows finer, instead of depending on intrinsic properties of the problem. Second, as our  $\alpha$ -privacy lower bound example shows, indiscriminate discretization without prior knowledge of the data can drastically degrade the performance of the best classifier in a class. Finally, infinite hypothesis classes and continuous data domains provide a natural abstraction for designing many machine learning algorithms, such as those based on convex optimization or differential geometry. Understanding the limitations of differentially private learning on such hypothesis classes and data domains is useful in designing differentially private approximations to these algorithms.

The rest of our paper is organized as follows. In Section 2, we define some preliminary notation, and explain our privacy model. In Section 3, we present our  $\alpha$ -privacy lower bound. Our  $\alpha$ -privacy upper bound is provided in Section 4. In Section 5, we provide some lower bounds on the sample requirement of learning with  $\alpha$ -label privacy. Finally, the proofs of most of our results are in the appendix.

### 1.1. Related work

The work most related to ours is Kasiviswanathan et al. (2008), Blum et al. (2008) and Beimel et al. (2010), each of which deals with either finite hypothesis classes or discrete data domains.

Kasiviswanathan et al. (2008) initiated the study of the sample requirement of differentially-private learning. They provided a (computationally inefficient)  $\alpha$ -private algorithm that learns any finite hypothesis class  $\mathcal{H}$  with error at most  $\epsilon$  using at most  $\tilde{O}(\frac{\log |\mathcal{H}|}{\alpha\epsilon})$  examples in the realizable case. For the non-realizable case, they provided an algorithm with a sample requirement of  $\tilde{O}(\log |\mathcal{H}| \cdot (\frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2}))$ . Moreover, using a result from Blum et al. (2008), they provided a computationally inefficient  $\alpha$ -private algorithm that learns a hypothesis class

with VC-dimension  $V$  and data dimension  $n$  with at most  $\tilde{O}(\frac{nV}{\alpha\epsilon^3})$  examples, *provided the data domain is  $\{-1, 1\}^n$* . None of these results apply when the data is drawn from a continuous distribution; moreover, their results cannot be directly extended to the continuous case.

The first work to study lower bounds on the sample requirement of differentially private learning was Beimel et al. (2010). They show that any  $\alpha$ -private algorithm that selects a hypothesis from a specific set  $C_\epsilon$  requires at least  $\tilde{\Omega}(\log(|C_\epsilon|)/\alpha)$  samples to achieve error  $\epsilon$ . Here  $C_\epsilon$  is an  $\epsilon$ -cover as well as an  $\epsilon$ -packing of the hypothesis class  $\mathcal{H}$  with respect to *every* distribution over the discrete data domain. They also show an upper bound of  $\tilde{O}(\log(|C_\epsilon|)/(\alpha\epsilon))$ . Such a cover  $C_\epsilon$  does not exist for continuous data domains; as a result their upper bounds do not apply to our setting. Moreover, unlike our lower bounds, their lower bound only applies to algorithms of a specific form (namely, those that output a hypothesis in  $C_\epsilon$ ), and it also does not apply when we only require the labels to be private.

For the setting of label privacy, Chaudhuri et al. (2006) show an upper bound for PAC-learning in terms of the VC dimension of the hypothesis class. We show a result very similar to theirs in the appendix for completeness, and we show lower bounds for learning with label-privacy which indicate that their bounds are almost tight, in terms of the dependence on  $\alpha$  and  $\epsilon$ .

Zhou et al. (2009) study some issues in defining differential privacy when dealing with continuous outcomes; however, they do not consider the question of learning classifiers on such data.

Finally, a lot of our work uses tools from the theory of generalization bounds. In particular, some of our upper and lower bounds are inspired by Bshouty et al. (2009), which bounds the sample complexity of (non-private) classification in terms of the doubling dimension of the disagreement metric.

**Other related work on privacy.** The issue of privacy in data analysis of sensitive information has long been a source of problems for curators of such data, and much of this is due to the realization that many simple and intuitive mechanisms designed to protect privacy are simply ineffective. For instance, the work of Narayanan and Shmatikov (2008) showed that an anonymized dataset released by Netflix revealed enough information so that an adversary, by knowing just a few of the movies rated by a particular user, would be able to uniquely identify such a user in the data set and determine *all* of his movie ratings. Similar attacks have been demonstrated on private data in other domains as well including social networks (Backstrom et al., 2007) and search engine query logs (Jones et al., 2007). Even releasing coarse statistics without proper privacy safeguards can be problematic. This was recently shown by Wang et al. (2009) in the context of genetic data, where a correlation matrix of genetic markers compiled from a group of individuals contained enough clues to uniquely pinpoint individuals in the dataset and learn of their private information, such as whether or not they had certain diseases.

In order to reason about privacy guarantees (or lack thereof), we need a formal definition of what it means to preserve privacy. In our work, we adopt the notion of *differential privacy* due to Dwork et al. (2006), which has over the last few years gained much popularity. Differential privacy is known to be a very strong notion of privacy: it has strong semantic guarantees (Kasiviswanathan and Smith, 2008) and is resistant to attacks that many earlier privacy definitions are susceptible to (Ganta et al., 2008b).

There has been a significant amount of work on differential privacy applied to a wide variety of data analysis tasks (Dwork et al., 2006; Chaudhuri and Mishra, 2006; Nissim et al., 2007; Barak et al., 2007; McSherry and Mironov, 2009). Some work that is relevant to ours include Blum et al. (2008), which provides a general method for publishing datasets on discrete data domains while preserving differential privacy so that the answers to queries from a function class with bounded VC dimension will be approximately preserved after the applying the sanitization procedure. More work on this line includes Roth (2010) and Gupta et al. (2011). A number of learning algorithms have also been suitably modified to guarantee differential privacy. For instance, both the classes of statistical query algorithms and the class of methods based on  $L_2$ -regularized empirical risk minimization with certain types of convex losses can be made differentially private (Blum et al., 2005; Chaudhuri et al., 2011).

There has also been some prior work on providing lower bounds on the loss of accuracy that any differentially private mechanism would suffer; much of this work is in the context of releasing answers to some  $a$  of queries made on a database of  $n$  individuals. The first such work is by (Blum et al., 2008), which shows that no differentially private mechanism can hope to release with a certain amount of accuracy the answer to a number of median queries when the data lies on a real line. This result is similar in spirit to our Theorem 5, but applies to a much harder problem, namely data release. Other relevant work includes (Hardt and Talwar, 2010), which uses a packing argument similar to ours to provide a lower bound on the amount of noise any differentially private mechanism needs to add to the answer to  $k$  linear queries on a database of  $n$  people.

There has also been a significant amount of prior work on privacy-preserving data mining (Agrawal and Srikant, 2000; Evfimievski et al., 2003; Sweeney, 2002; Machanavajjhala et al., 2006), which spans several communities and uses privacy models other than differential privacy. Many of the models used have been shown to be susceptible to various attacks, such as *composition attacks*, where the adversary has some amount of prior knowledge (Ganta et al., 2008a). An alternative line of privacy work is in the Secure Multiparty Computation setting due to Yao (1982), where the sensitive data is split across several adversarial databases, and the goal is to compute a function on the union of these databases. This is in contrast with our setting, where a single centralized algorithm can access the entire dataset.

## 2. Preliminaries

### 2.1. Privacy model

We use the *differential privacy* model of Dwork et al. (2006). In this model, a private database  $\text{DB} \subseteq \mathcal{Z}$  consists of  $m$  sensitive entries from a domain  $\mathcal{Z}$ ; each entry in  $\text{DB}$  is a record about an individual (*e.g.*, their medical history) that one wishes to keep private.

The database  $\text{DB}$  is accessed by users through a sanitizer  $M$ . The sanitizer, a randomized algorithm, is said to preserve differential privacy if the value of any one individual in the database does not significantly alter the output distribution of  $M$ .

**Definition 1** *A randomized mechanism  $M$  guarantees  $\alpha$ -differential privacy if, for all databases  $\text{DB}_1$  and  $\text{DB}_2$  that differ by the value of at most one individual, and for every set*

$G$  of possible outputs of  $M$ ,

$$\Pr_M[M(\mathbf{DB}_1) \in G] \leq \Pr_M[M(\mathbf{DB}_2) \in G] \cdot e^\alpha.$$

We emphasize that the probability in the definition above is only with respect to the internal randomization of the algorithm; it is independent of all other random sources, including any that may have generated the values of the input database.

Differential privacy is a strong notion of privacy (Dwork et al., 2006; Kasiviswanathan and Smith, 2008; Ganta et al., 2008b). In particular, if a sanitizer  $M$  ensures  $\alpha$ -differential privacy, then, an adversary who knows the private values of all the individuals in the database except for one and has arbitrary prior knowledge about the value of the last individual, cannot gain additional confidence about the private value of the last individual by observing the output of a differentially private sanitizer. The level of privacy is controlled by  $\alpha$ , where a lower value of  $\alpha$  implies a stronger guarantee of privacy.

## 2.2. Learning model

We consider a standard probabilistic learning model for binary classification. Let  $\mathcal{P}$  be a distribution over  $\mathcal{X} \times \{\pm 1\}$ , where  $\mathcal{X}$  is the data domain and  $\{\pm 1\}$  are the possible labels. We use  $\mathcal{D}$  to denote the marginal of  $\mathcal{P}$  over the data domain  $\mathcal{X}$ . The *classification error* of a hypothesis  $h : \mathcal{X} \rightarrow \{\pm 1\}$  with respect to a data distribution  $\mathcal{P}$  is

$$\Pr_{(x,y) \sim \mathcal{P}}[h(x) \neq y].$$

We denote by  $S \sim \mathcal{P}^m$  an i.i.d. draw of  $m$  labeled examples  $S = \{(x_1, y_1), \dots, (x_m, y_m)\} \subseteq \mathcal{X} \times \{\pm 1\}$  from the distribution  $\mathcal{P}$ . This process can equivalently be seen as drawing an unlabeled sample  $X := \{x_1, \dots, x_m\}$  from the marginal  $\mathcal{D}$ , and then, for each  $x \in X$ , drawing the corresponding label  $y$  from the induced conditional distribution.

A learning algorithm is given as input a set of  $m$  labeled examples  $S \sim \mathcal{P}^m$ , a target accuracy parameter  $\epsilon \in (0, 1)$ , and target confidence parameter  $\delta \in (0, 1)$ . Its goal is to return a hypothesis  $h : \mathcal{X} \rightarrow \{\pm 1\}$  such that its *excess generalization error* with respect to a specified hypothesis class  $\mathcal{H}$

$$\Pr_{(x,y) \sim \mathcal{P}}[h(x) \neq y] - \inf_{h' \in \mathcal{H}} \Pr_{(x,y) \sim \mathcal{P}}[h'(x) \neq y]$$

is at most  $\epsilon$  with probability at least  $1 - \delta$  over the random choice of the sample  $S \sim \mathcal{P}^m$ , as well as any internal randomness of the algorithm.

We also occasionally adopt the *realizable assumption* (with respect to  $\mathcal{H}$ ). The realizable assumption states that there exists some  $h^* \in \mathcal{H}$  such that  $\Pr_{(x,y) \sim \mathcal{P}}[h^*(x) \neq y] = 0$ . In this case, the excess generalization error of a hypothesis  $h$  is simply its classification error. Without the realizable assumption, there may be no classifier in the hypothesis class  $\mathcal{H}$  with zero classification error, and we refer to this as the non-realizable case.

## 2.3. Privacy-preserving classification

In privacy-preserving classification, we assume that the database is a training dataset drawn in an i.i.d manner from some data distribution  $\mathcal{P}$ , and that the sanitization mechanism is a learning algorithm that outputs a classifier based on the training data. In this paper, we consider two possible privacy requirements on our learning algorithms.

**Definition 2** A randomized learning algorithm  $\mathcal{A}$  guarantees  $\alpha$ -label privacy ( $\mathcal{A}$  is  $\alpha$ -label private) if, for any two datasets  $S_1 = \{(x_1, y_1), \dots, (x_{m-1}, y_{m-1}), (x_m, y_m)\}$  and  $S_2 = \{(x_1, y_1), \dots, (x_{m-1}, y_{m-1}), (x_m, y'_m)\}$  differing in at most one label  $y'_m$ , and any set of outputs  $G$  of  $\mathcal{A}$ ,

$$\Pr_{\mathcal{A}}[\mathcal{A}(S_1) \in G] \leq \Pr_{\mathcal{A}}[\mathcal{A}(S_2) \in G] \cdot e^\alpha.$$

**Definition 3** A randomized learning algorithm  $\mathcal{A}$  guarantees  $\alpha$ -privacy ( $\mathcal{A}$  is  $\alpha$ -private) if, for any two datasets  $S_1 = \{(x_1, y_1), \dots, (x_{m-1}, y_{m-1}), (x_m, y_m)\}$  and  $S_2 = \{(x_1, y_1), \dots, (x_{m-1}, y_{m-1}), (x'_m, y'_m)\}$  differing in at most one example  $(x'_m, y'_m)$ , and any set of outputs  $G$  of  $\mathcal{A}$ ,

$$\Pr_{\mathcal{A}}[\mathcal{A}(S_1) \in G] \leq \Pr_{\mathcal{A}}[\mathcal{A}(S_2) \in G] \cdot e^\alpha.$$

Note that if the input dataset  $S$  is a random variable, then for any value  $S' \subseteq \mathcal{X} \times \{\pm 1\}$  in the range of  $S$ , the conditional probability distribution of  $\mathcal{A}(S) \mid S = S'$  is determined only by the algorithm  $\mathcal{A}$  and the value  $S'$ ; it is independent of the distribution of the random variable  $S$ . Therefore, for instance,

$$\Pr_{S, \mathcal{A}}[\mathcal{A}(S) \in G \mid S = S'] = \Pr_{\mathcal{A}}[\mathcal{A}(S') \in G].$$

for any  $S' \subseteq \mathcal{X} \times \{\pm 1\}$  and any set of outputs  $G$ .

The difference between the two notions of privacy is that for  $\alpha$ -label privacy, the two databases can differ only in the label of one example; whereas for  $\alpha$ -privacy, the two databases differ can differ in a complete example (both labeled and unlabeled parts). Thus,  $\alpha$ -label privacy only ensures the privacy of the label component of each example; it makes no guarantees about the unlabeled part. If a classification algorithm guarantees  $\alpha$ -privacy, then it also guarantees  $\alpha$ -label privacy. Thus  $\alpha$ -label privacy is a weaker notion of privacy than  $\alpha$ -privacy.

The notion of label privacy was also considered by Chaudhuri et al. (2006), who provided an algorithm for learning with label privacy. For strict privacy, one would require the learning algorithm to guarantee  $\alpha$ -privacy; however, label privacy may also be an useful notion. For example, if the data  $x$  represents public demographic information (e.g., age, zip code, education), while the label  $y$  represents income level, an individual may consider the label to be private but may not mind if others can infer her demographic information (which could be relatively public already) by her inclusion in the database.

Thus, the goal of a  $\alpha$ -private (resp.  $\alpha$ -label private) learning algorithm is as follows. Given a dataset  $S$  of size  $m$ , a privacy parameter  $\alpha$ , a target accuracy  $\epsilon$ , and a target confidence parameter  $\delta$ :

1. guarantee  $\alpha$ -privacy (resp.  $\alpha$ -label privacy) of the training dataset  $S$ ;
2. with probability at least  $1 - \delta$  over both the random choice of  $S \sim \mathcal{P}^m$  and the internal randomness of the algorithm, return a hypothesis  $h : \mathcal{X} \rightarrow \{\pm 1\}$  with excess generalization error

$$\Pr_{(x,y) \sim \mathcal{P}}[h(x) \neq y] - \inf_{h' \in \mathcal{H}} \Pr_{(x,y) \sim \mathcal{P}}[h'(x) \neq y] \leq \epsilon.$$

## 2.4. Additional definitions and notation

We now present some additional essential definitions and notation.

**Metric spaces, doubling dimension, covers, and packings.** A metric space  $(\mathcal{Z}, \rho)$  is a tuple, where  $\mathcal{Z}$  is a set of elements, and  $\rho$  is a distance function from  $\mathcal{Z} \times \mathcal{Z}$  to  $\{0\} \cup \mathbb{R}^+$ . Let  $(\mathcal{Z}, \rho)$  be any arbitrary metric space. For any  $z \in \mathcal{Z}$  and  $r > 0$ , let  $B(z, r) = \{z' \in \mathcal{Z} : \rho(z, z') \leq r\}$  denote the ball centered at  $z$  of radius  $r$ .

The *diameter* of  $(\mathcal{Z}, \rho)$  is  $\sup\{\rho(z, z') : z, z' \in \mathcal{Z}\}$ , the longest distance in the space. An  $\varepsilon$ -*cover* of  $(\mathcal{Z}, \rho)$  is a set  $C \subseteq \mathcal{Z}$  such that for all  $z \in \mathcal{Z}$ , there exists some  $z' \in C$  such that  $\rho(z, z') \leq \varepsilon$ . An  $\varepsilon$ -*packing* of  $(\mathcal{Z}, \rho)$  is a set  $P \subseteq \mathcal{Z}$  such that  $\rho(z, z') > \varepsilon$  for all distinct  $z, z' \in P$ . Let  $\mathcal{N}_\varepsilon(\mathcal{Z}, \rho)$  denote the size of the smallest  $\varepsilon$ -cover of  $(\mathcal{Z}, \rho)$ .

We define the *doubling dimension of  $(\mathcal{Z}, \rho)$  at scale  $\varepsilon$* , denoted as  $\text{ddim}_\varepsilon(\mathcal{Z}, \rho)$ , as the smallest number  $d$  such that each ball  $B(z, \varepsilon) \subseteq \mathcal{Z}$  of radius  $\varepsilon$  can be covered by at most  $\lfloor 2^d \rfloor$  balls of radius  $\varepsilon/2$ , i.e. there exists  $z_1, \dots, z_{\lfloor 2^d \rfloor} \in \mathcal{Z}$  such that  $B(z, \varepsilon) \subseteq B(z_1, \varepsilon/2) \cup \dots \cup B(z_{\lfloor 2^d \rfloor}, \varepsilon/2)$ . Notice that  $\text{ddim}_\varepsilon(\mathcal{Z}, \rho)$  may increase or decrease with  $\varepsilon$ . The *doubling dimension of  $(\mathcal{Z}, \rho)$*  is  $\sup\{\text{ddim}_r(\mathcal{Z}, \rho) : r > 0\}$ .

**Disagreement metrics.** The *disagreement metric* of a hypothesis class  $\mathcal{H}$  with respect to a data distribution  $\mathcal{D}$  over  $\mathcal{X}$  is the metric  $(\mathcal{H}, \rho_{\mathcal{D}})$ , where  $\rho_{\mathcal{D}}$  is the following distance function:

$$\rho_{\mathcal{D}}(h, h') := \Pr_{x \sim \mathcal{D}}[h(x) \neq h'(x)].$$

The *empirical disagreement metric* of a hypothesis class  $\mathcal{H}$  with respect to a data distribution  $\mathcal{D}$  over  $\mathcal{X}$  is the metric  $(\mathcal{H}, \rho_X)$ , where  $\rho_X$  is the following distance function:

$$\rho_X(h, h') := \frac{1}{|X|} \sum_{x \in X} I[h(x) \neq h'(x)].$$

The disagreement metric (or empirical disagreement metric) is the proportion of unlabeled examples on which  $h$  and  $h'$  disagree with respect to  $\mathcal{D}$  (or the uniform distribution over  $X$ ). We use the notation  $B_{\mathcal{D}}(h, r)$  to denote the ball centered at  $h$  of radius  $r$  with respect to  $\rho_{\mathcal{D}}$ , and  $B_X(h, r)$  to denote the ball centered at  $h$  of radius  $r$  with respect to  $\rho_X$ .

**Datasets and empirical error.** For an unlabeled dataset  $X \subseteq \mathcal{X}$  and a hypothesis  $h : \mathcal{X} \rightarrow \{\pm 1\}$ , we denote by  $S_{X,h} := \{(x, h(x)) : x \in X\}$  the labeled dataset induced by labeling  $X$  with  $h$ . The *empirical error* of a hypothesis  $h : \mathcal{X} \rightarrow \{\pm 1\}$  with respect to a labeled dataset  $S \subseteq \mathcal{X} \times \{\pm 1\}$  is  $\text{err}(h, S) := (1/|S|) \sum_{(x,y) \in S} \mathbf{1}[h(x) \neq y]$  the average number of mistakes that  $h$  makes on  $S$ ; note that  $\rho_X(h, S_{X,h'}) = \text{err}(h, S_{X,h'})$ . Finally, we informally use the  $\tilde{O}(\cdot)$  notation to hide  $\log(1/\delta)$  factors, as well as factors that are logarithmic in those that do appear.

## 3. Lower bounds for learning with $\alpha$ -privacy

In this section, we show a lower bound on the sample requirement of learning with  $\alpha$ -privacy. In particular, we show an example that illustrates that when the data is drawn from a continuous distribution, for any  $M$ , all  $\alpha$ -private algorithms that are supplied with at most  $M$  examples fail to output a good classifier for at least one unlabeled data distribution.

Our example hypothesis class is the class of thresholds on  $[0, 1]$ . This simple class has VC dimension 1, and thus can be learnt non-privately with classification error  $\epsilon$  given only  $\tilde{O}(1/\epsilon)$  examples, regardless of the unlabeled data distribution. However, Theorem 5 shows that even in the realizable case, for every  $\alpha$ -private algorithm that is given a bounded number of examples, there is at least one unlabeled data distribution on which the learning algorithm produces a classifier with error  $\geq \frac{1}{5}$ , with probability at least  $1/2$  over its own random coins.

The key intuition behind our example is that if most of the unlabeled data is concentrated in a small region around the best classifier, then, even slightly perturbing the best classifier will result in a large classification error. As the process of ensuring differential privacy necessarily involves some perturbation, unless the algorithm has some prior public knowledge about the data distribution, the number of samples required to learn privately grows with growing concentration of the data around the best classifier. As illustrated by our theorem, this problem is not alleviated if the support of the unlabeled distribution is known; even if the data distribution has large support, a large fraction of the data can still lie in a region close to the best classifier.

Before we describe our example in detail, we first need a definition.

**Definition 4** *The class of thresholds on the unit interval is the class of functions  $h_w : [0, 1] \rightarrow \{-1, 1\}$  such that:*

$$h_w(x) := \begin{cases} 1 & \text{if } x \geq w \\ -1 & \text{otherwise.} \end{cases}$$

**Theorem 5** *Let  $M > 2$  be any number, and let  $\mathcal{H}$  be the class of thresholds on the unit interval  $[0, 1]$ . For any  $\alpha$ -private algorithm  $A$  that outputs a hypothesis  $h \in \mathcal{H}$ , there exists a distribution  $\mathcal{P}$  on labeled examples with the following properties:*

1. *There exists a threshold  $h^* \in \mathcal{H}$  with classification error 0 with respect to  $\mathcal{P}$ .*
2. *For all samples  $S$  of size  $m \leq M$  drawn from  $\mathcal{P}$ , with probability at least  $1/2$  over the random coins of  $A$ , the hypothesis output by  $A(S)$  has classification error at least  $\frac{1}{5}$  with respect to  $\mathcal{P}$ .*
3. *The marginal  $\mathcal{D}$  of  $\mathcal{P}$  over the unlabeled data has support  $[0, 1]$ .*

**Proof** Let  $\eta = \frac{1}{6+4\exp(\alpha M)}$ , and let  $\mathcal{U}$  denote the uniform distribution over  $[0, 1]$ . Let  $Z = \{\eta, 2\eta, \dots, K\eta\}$ , where  $K = \lfloor 1/\eta \rfloor - 1$ . We let  $G_z = [z - \eta/3, z + \eta/3]$  for  $z \in Z$ , and let  $\mathcal{G}_z \subset \mathcal{H}$  be the subset of thresholds:  $\mathcal{G}_z = \{h_\tau | \tau \in G_z\}$ . We note that  $G_z \subseteq [0, 1]$  for all  $z \in Z$ .

For each  $z \in Z$ , we define a distribution  $\mathcal{P}_z$  over labeled examples as follows. First, we describe the marginal  $\mathcal{D}_z$  of  $\mathcal{P}_z$  over the unlabeled data. A sample from  $\mathcal{D}_z$  is drawn as follows. With probability  $\frac{1}{2}$ ,  $x$  is drawn from  $\mathcal{U}$ ; with probability  $\frac{1}{2}$ , it is drawn uniformly from  $G_z$ . Now, an unlabeled example  $x$  drawn from  $\mathcal{D}_z$  is labeled positive if  $x \geq z$ , and negative otherwise. We observe that for every such distribution  $\mathcal{P}_z$ , there exists a threshold, namely,  $h_z$  that has classification error 0; in addition, the support of  $\mathcal{D}_z$  is  $[0, 1]$ . Moreover, there are  $\lfloor \frac{1}{\eta} \rfloor - 1$  such distributions  $\mathcal{P}_z$  in all, and  $\lfloor \frac{1}{\eta} \rfloor - 1 \geq 5$ .

We say that an  $\alpha$ -private algorithm  $A$  *succeeds* on a sample  $S$  with respect to a distribution  $\mathcal{P}$  if with probability  $\frac{1}{2}$  over the random coins of  $A$ , the hypothesis output by  $A(S)$  has classification error  $< \frac{1}{5}$  over  $\mathcal{P}$ .

Suppose for the sake of contradiction that there exists an  $\alpha$ -private algorithm  $A^*$  such that for all distributions  $\mathcal{P}$ , there is at least one sample  $S$  of size  $\leq M$  drawn from  $\mathcal{P}$  such that  $A^*$  succeeds on  $S$  with respect to  $\mathcal{P}$ . Then, for all  $z \in Z$ , and for all  $\mathcal{P}_z$ , there exists a sample  $S_z$  of size  $m \leq M$  drawn from  $\mathcal{P}_z$  such that  $A^*$  succeeds on  $S_z$  with respect to  $\mathcal{P}_z$ .

By construction, the  $G_z$ 's are disjoint, so

$$\Pr_{A^*}[A^*(S_z) \notin G_z] \geq \sum_{z' \in Z \setminus \{z\}} \Pr_{A^*}[A^*(S_z) \in G_{z'}]. \quad (1)$$

Furthermore, any  $S_z$  differs from  $S_{z'}$  by at most  $m$  labeled examples, so because  $A^*$  is  $\alpha$ -private, Lemma 22 that implies for any  $z'$ ,

$$\Pr_{A^*}[A^*(S_z) \in G_{z'}] \geq e^{-\alpha m} \Pr_{A^*}[A^*(S_{z'}) \in G_{z'}]. \quad (2)$$

If  $A^*(S_{z'})$  lies outside  $G_{z'}$ ,  $A^*(S_{z'})$  classifies at least  $1/4$  fraction of the examples from  $\mathcal{P}_{z'}$  incorrectly, and thus  $A^*$  cannot succeed on  $S_{z'}$  with respect to  $\mathcal{P}_{z'}$ . Therefore, by the assumption on  $A^*$ , for any  $z'$ ,

$$\Pr_{A^*}[A^*(S_{z'}) \in G_{z'}] \geq \frac{1}{2}. \quad (3)$$

Combining Equations (1), (2), and (3) gives the inequality

$$\Pr_{A^*}[A^*(S_z) \notin G_z] \geq e^{-\alpha m} \cdot \sum_{z' \in Z \setminus \{z\}} \frac{1}{2} \geq e^{-\alpha m} \cdot \left(\frac{1}{\eta} - 2\right) \cdot \frac{1}{2}.$$

Since  $m \leq M$ , the quantity on the RHS of the above equation is more than  $\frac{2}{3}$ .  $A^*$  therefore does not succeed on  $S_z$  with respect to  $\mathcal{P}_z$ , thus leading to a contradiction. ■

#### 4. Upper bounds for learning with $\alpha$ -privacy

In this section, we show an upper bound on the sample requirement of learning with  $\alpha$ -privacy by presenting a learning algorithm that works on infinite hypothesis classes over continuous data domains, under certain conditions on the hypothesis class and the data distribution. Our algorithm works in the non-realizable case, that is, when there may be no hypothesis in the target hypothesis class with zero classification error.

A natural way to extend the algorithm of Kasiviswanathan et al. (2008) to an infinite hypothesis class  $\mathcal{H}$  is to compute a suitable finite subset  $\mathcal{G}$  of  $\mathcal{H}$  that contains a hypothesis with low excess generalization error, and then use the exponential mechanism of McSherry and Talwar (2007) on  $\mathcal{G}$ . To ensure that a hypothesis with low error is indeed in  $\mathcal{G}$ , we would like  $\mathcal{G}$  to be an  $\epsilon$ -cover of the disagreement metric  $(\mathcal{H}, \rho_{\mathcal{D}})$ . In a non-private or label-private learning, we can compute such a  $\mathcal{G}$  directly based on the unlabeled training examples; in our

setting, the training examples themselves are sensitive, and this approach does not directly apply.

The key idea behind our algorithm is that instead of using the sensitive data to compute  $\mathcal{G}$ , we can use a reference distribution  $\mathcal{U}$  that is known independently of the sensitive data. For instance, if the domain of the unlabeled data is bounded, then a reasonable choice for  $\mathcal{U}$  is the uniform distribution over the domain. Our key observation is that if  $\mathcal{U}$  is close to the unlabeled data distribution  $\mathcal{D}$  according to a certain measure of closeness inspired by Dasgupta (2005) and Freund et al. (1997), then a cover of the disagreement metric on  $\mathcal{H}$  with respect to  $\mathcal{U}$  is a (possibly coarser) cover of the disagreement metric on  $\mathcal{H}$  with respect to  $\mathcal{D}$ . Thus we can set  $\mathcal{G}$  to be a fine cover of  $(\mathcal{H}, \rho_{\mathcal{U}})$ , and this cover can be computed privately as it is independent of the sensitive data.

Our algorithm works when the doubling dimension of  $(\mathcal{H}, \rho_{\mathcal{U}})$  is finite; under this condition, there is always such a finite cover  $\mathcal{G}$ . We note that this is a fairly weak condition that is satisfied by many hypothesis classes and data distributions. For example, any hypothesis class with finite VC dimension will satisfy this condition for any unlabeled data distribution  $\mathcal{U}$ .

Finally, it may be tempting to think that one can further improve the sample requirement of our algorithm by using the sensitive data to privately refine a cover of  $(\mathcal{H}, \rho_{\mathcal{U}})$  to a cover of  $(\mathcal{H}, \rho_{\mathcal{D}})$ . However, our calculations show that naively refining such a cover leads to a much higher sample requirement.

We now define our notion of closeness.

**Definition 6** *We say that a data distribution  $\mathcal{D}$  is  $\kappa$ -smooth with respect to a distribution  $\mathcal{U}$  for some  $\kappa \geq 1$ , if for all measurable sets  $A \subseteq \mathcal{X}$ ,*

$$\Pr_{x \sim \mathcal{D}}[x \in A] \leq \kappa \cdot \Pr_{x \sim \mathcal{U}}[x \in A].$$

This notion of smoothness is very similar to, but weaker than the notions of closeness between distributions that have been used by (Dasgupta, 2005; Freund et al., 1997). We note that if  $\mathcal{D}$  is absolutely continuous with respect to  $\mathcal{U}$  (*i.e.*,  $\mathcal{U}$  assigns zero probability to a set only if  $\mathcal{D}$  does also), then  $\mathcal{D}$  is  $\kappa$ -smooth with respect to  $\mathcal{U}$  for some finite  $\kappa$ .

#### 4.1. Algorithm

Our main algorithm  $\mathcal{A}_1$  is given in Figure 1. The first step of the algorithm calculates the distance scale at which it should construct a cover of  $(\mathcal{H}, \rho_{\mathcal{U}})$ . This scale is a function of  $|S|$ , the size of the input data set  $S$ , and can be computed privately because  $|S|$  is not sensitive information. A suitable cover of  $(\mathcal{H}, \rho_{\mathcal{U}})$  that is also a suitable packing of  $(\mathcal{H}, \rho_{\mathcal{U}})$  is then constructed; note that such a set always exists because of Lemma 13. In the final step, an exponential mechanism (McSherry and Talwar, 2007) is used to select a hypothesis from the cover with low error. As this step of the algorithm is the only one that uses the input data, the algorithm is  $\alpha$ -private as long as this last step guarantees  $\alpha$ -privacy.

#### 4.2. Privacy and learning guarantees

Our first theorem states the privacy guarantee of Algorithm  $\mathcal{A}_1$ .

**Theorem 7** *Algorithm  $\mathcal{A}_1$  preserves  $\alpha$ -privacy.*

**Algorithm  $\mathcal{A}_1$ .**

**Input:** private labeled dataset  $S \subseteq \mathcal{X} \times \{\pm 1\}$ , public reference distribution  $\mathcal{U}$  over  $\mathcal{X}$ , privacy parameter  $\alpha \in (0, 1)$ , accuracy parameter  $\epsilon \in (0, 1)$ , confidence parameter  $\delta \in (0, 1)$ .

**Output:**  $h_{\mathcal{A}} \in \mathcal{H}$ .

1. Solve the following equation to compute  $\hat{\kappa} > 0$ :

$$|S| = C \cdot \left( \frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2} \right) \cdot \left( d_{\mathcal{U}} \cdot \log \frac{\hat{\kappa}}{\epsilon} + \log \frac{1}{\delta} \right),$$

where  $C$  is the constant from Theorem 8; let  $\varepsilon_0 := \epsilon/(4\hat{\kappa})$ .

2. Let  $\mathcal{G}$  be an  $\varepsilon_0$ -packing of  $(\mathcal{H}, \rho_{\mathcal{U}})$  that is also an  $\varepsilon_0$ -cover.
3. Randomly choose  $h_{\mathcal{A}} \in \mathcal{G}$  according to the distribution  $(p_g : g \in \mathcal{G})$ , where  $p_g \propto \exp(-\alpha|S|\text{err}(g, S)/2)$  for each  $g \in \mathcal{G}$ , and return  $h_{\mathcal{A}}$ .

Figure 1: Learning algorithm for  $\alpha$ -privacy.

**Proof** The algorithm only accesses the private dataset  $S$  in the final step. Because changing one labeled example in  $S$  changes  $\text{err}(g, S)$  by at most 1, this step is guarantees  $\alpha$ -privacy (McSherry and Talwar, 2007). ■

The next theorem provides an upper bound on the sample requirement of Algorithm  $\mathcal{A}_1$ . This bound depends on the doubling dimension  $d_{\mathcal{U}}$  of  $(\mathcal{H}, \rho_{\mathcal{U}})$  and the smoothness parameter  $\kappa$ , as well as the privacy and learning parameters  $\alpha, \epsilon, \delta$ .

**Theorem 8** *Let  $\mathcal{P}$  be a distribution over  $\mathcal{X} \times \{\pm 1\}$  whose marginal over  $\mathcal{X}$  is  $\mathcal{D}$ . There exists a universal constant  $C > 0$  such that for any  $\alpha, \epsilon, \delta \in (0, 1)$ , the following holds. If*

1. *the doubling dimension  $d_{\mathcal{U}}$  of  $(\mathcal{H}, \rho_{\mathcal{U}})$  is finite,*
2.  *$\mathcal{D}$  is  $\kappa$ -smooth with respect to  $\mathcal{U}$ ,*
3.  *$S \subseteq \mathcal{X} \times \{\pm 1\}$  is an i.i.d. random sample from  $\mathcal{P}$  such that*

$$|S| \geq C \cdot \left( \frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2} \right) \cdot \left( d_{\mathcal{U}} \cdot \log \frac{\kappa}{\epsilon} + \log \frac{1}{\delta} \right), \quad (4)$$

*then with probability at least  $1 - \delta$ , the hypothesis  $h_{\mathcal{A}} \in \mathcal{H}$  returned by  $\mathcal{A}_1(S, \mathcal{U}, \alpha, \epsilon, \delta)$  satisfies*

$$\Pr_{(x,y) \sim \mathcal{P}}[h_{\mathcal{A}}(x) \neq y] \leq \inf_{h' \in \mathcal{H}} \Pr_{(x,y) \sim \mathcal{P}}[h'(x) \neq y] + \epsilon.$$

The proof of Theorem 8 is stated in Appendix C. If we have prior knowledge that some hypothesis in  $\mathcal{H}$  has zero error (the realizability assumption), then the sample requirement can be improved with a slightly modified version of Algorithm  $\mathcal{A}_1$ . This algorithm, called Algorithm  $\mathcal{A}_{1r}$ , is given in Figure 3 in Appendix C.

**Theorem 9** Let  $\mathcal{P}$  be any probability distribution over  $\mathcal{X} \times \{\pm 1\}$  whose marginal over  $\mathcal{X}$  is  $\mathcal{D}$ . There exists a universal constant  $C > 0$  such that for any  $\alpha, \epsilon, \delta \in (0, 1)$ , the following holds. If

1. the doubling dimension  $d_{\mathcal{U}}$  of  $(\mathcal{H}, \rho_{\mathcal{U}})$  is finite,
2.  $\mathcal{D}$  is  $\kappa$ -smooth with respect to  $\mathcal{U}$ ,
3.  $S \subseteq \mathcal{X} \times \{\pm 1\}$  is an i.i.d. random sample from  $\mathcal{P}$  such that

$$|S| \geq C \cdot \frac{1}{\alpha\epsilon} \cdot \left( d_{\mathcal{U}} \cdot \log(\kappa/\epsilon) + \log \frac{1}{\delta} \right), \quad (5)$$

4. there exists  $h^* \in \mathcal{H}$  with  $\Pr_{(x,y) \sim \mathcal{P}}[h^*(x) \neq y] = 0$ ,

then with probability at least  $1 - \delta$ , the hypothesis  $h_{\mathcal{A}} \in \mathcal{H}$  returned by  $\mathcal{A}_{1r}(S, \mathcal{U}, \alpha, \epsilon, \delta)$  satisfies

$$\Pr_{(x,y) \sim \mathcal{P}}[h_{\mathcal{A}}(x) \neq y] \leq \epsilon.$$

Again, the proof of Theorem 9 is in Appendix C.

### 4.3. Examples

In this section, we give some examples that illustrate the sample requirement of Algorithm  $\mathcal{A}_1$ .

First, we consider the example from the lower bound given in the proof of Theorem 5.

**Example 1** The domain of the data is  $\mathcal{X} := [0, 1]$ , and the hypothesis class is  $\mathcal{H} := \mathcal{H}_{\text{thresholds}} = \{h_t : t \in [0, 1]\}$  (recall,  $h_t(x) = 1$  if and only if  $x \geq t$ ). A natural choice for the reference distribution  $\mathcal{U}$  is the uniform distribution over  $[0, 1]$ ; the doubling dimension of  $(\mathcal{H}, \rho_{\mathcal{U}})$  is 1 because every interval can be covered by two intervals of half the length. Fix some  $M > 0$  and  $\alpha \in (0, 1)$ , and let  $\eta := 1/(6 + 4 \exp(\alpha M))$ . For  $z \in [\eta, 1 - \eta]$ , let  $\mathcal{D}_z$  be the distribution on  $[0, 1]$  with density

$$p_{\mathcal{D}_z}(x) := \begin{cases} \frac{1}{2} + \frac{3}{4\eta} & \text{if } z - \eta/3 \leq x \leq z + \eta/3, \\ \frac{1}{2} & \text{if } 0 \leq x < z - \eta/3 \text{ or } z + \eta/3 < x \leq 1. \end{cases}$$

Clearly,  $\mathcal{D}_z$  is  $\kappa$ -smooth with respect to  $\mathcal{U}$  for  $\kappa = \frac{1}{2} + \frac{3}{4\eta} = O(\exp(\alpha M))$ . Therefore the sample requirement of Algorithm  $\mathcal{A}_1$  to learn with  $\alpha$ -privacy and excess generalization error  $\epsilon$  is at most

$$C \cdot \left( \frac{1}{\epsilon\alpha} + \frac{1}{\epsilon^2} \right) \cdot \left( \alpha M + \log \frac{1}{\delta} \right)$$

which is  $\tilde{O}(M)$  for constant  $\epsilon$ , matching the lower bound from Theorem 5 up to constants.

Next, we consider two examples in which the domain of the unlabeled data  $\mathcal{X} := \mathbb{S}^{n-1}$  is the uniform distribution on the unit sphere in  $\mathbb{R}^n$ :

$$\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\| = 1\}$$

and the target hypothesis class  $\mathcal{H} := \mathcal{H}_{\text{linear}}$  is the class of linear separators that pass through the origin in  $\mathbb{R}^n$ :

$$\mathcal{H}_{\text{linear}} := \{h_w : w \in \mathbb{S}^{n-1}\} \quad \text{where } h_w(x) = 1 \text{ if and only if } w \cdot x \geq 0.$$

The examples will consider two different distributions over  $\mathcal{X}$ .

A natural reference data distribution in this setting is the uniform distribution over  $\mathbb{S}^{n-1}$ ; this will be our reference distribution  $\mathcal{U}$ . It is known that  $d_{\mathcal{U}} := \sup\{\text{ddim}_r(\mathcal{H}, \rho_{\mathcal{U}}) : r \geq 0\} = O(n)$  (Bshouty et al., 2009).

**Example 2** We consider a case where the unlabeled data distribution  $\mathcal{D}$  is concentrated near an equator of  $\mathbb{S}^{n-1}$ . More formally, for some vector  $u \in \mathbb{S}^{n-1}$ , and  $\gamma \in (0, 1)$ , we let  $\mathcal{D}$  be uniform over  $W := \{x \in \mathbb{S}^{n-1} : |u \cdot x| \leq \gamma\}$ ; in other words, the unlabeled data lies in a small band of width  $\gamma$  around the equator.

By Lemma 20 (see Appendix C),  $\mathcal{D}$  is  $\kappa$ -smooth with respect to  $\mathcal{U}$  for  $\kappa = \frac{1}{1 - 2 \exp(-n\gamma^2/2)}$ . Thus the sample requirement of Algorithm  $\mathcal{A}_1$  to learn with  $\alpha$ -privacy and excess excess generalization error  $\epsilon$  is at most

$$C \cdot \left( \frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2} \right) \cdot \left( n \cdot \log \left( \frac{1}{\epsilon} \cdot \frac{1}{1 - 2 \exp(-n\gamma^2/2)} \right) + \log \frac{1}{\delta} \right).$$

When  $n$  is large and  $\gamma \geq 1/\sqrt{n}$ , this bound is  $\tilde{O}(\frac{n}{\alpha\epsilon} + \frac{n}{\epsilon^2})$ , where the  $\tilde{O}$  notation hides factors logarithmic in  $1/\delta$  and  $1/\epsilon$ .

**Example 3** Now we consider the case where the unlabeled data lies on two diametrically opposite spherical caps. More formally, for some vector  $u \in \mathbb{S}^{n-1}$ , and  $\gamma \in (0, 1)$ , we now let  $\mathcal{D}$  be uniform over  $\mathbb{S}^{n-1} \setminus W$ , where  $W := \{x \in \mathbb{S}^{n-1} : |u \cdot x| \leq \gamma\}$ ; in other words, the unlabeled data lies outside a band of width  $\gamma$  around the equator.

By Lemma 21 (see Appendix C),  $\mathcal{D}$  is  $\kappa$ -smooth with respect to  $\mathcal{U}$  for  $\kappa = \left(\frac{2}{1-\gamma}\right)^{\frac{n-1}{2}}$ . Thus the sample requirement of Algorithm  $\mathcal{A}_1$  is to learn with  $\alpha$ -privacy and excess generalization error  $\epsilon$  is at most:

$$C \cdot \left( \frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2} \right) \cdot \left( n^2 \cdot \log \frac{2}{1-\gamma} + n \cdot \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right).$$

Thus, for large  $n$  and constant  $\gamma < 1$ , the sample requirement of Algorithm  $\mathcal{A}_1$  is  $\tilde{O}(\frac{n^2}{\epsilon^2} + \frac{n^2}{\alpha\epsilon})$ . So, even though the smoothness parameter  $\kappa$  is exponential in the dimension  $n$ , the sample requirement remains polynomial in  $n$ .

## 5. Lower bounds for learning with $\alpha$ -label privacy

In this section, we provide two lower bounds on the sample complexity of learning with  $\alpha$ -label privacy. Our first lower bound holds when  $\alpha$  and  $\epsilon$  are small (that is, high privacy and high accuracy), and when the hypothesis class has bounded VC dimension  $V$ . If these conditions hold, then we show a lower bound of  $\Omega(d/\epsilon\alpha)$  where  $d$  is the doubling dimension of the disagreement metric  $(\mathcal{H}, \rho_{\mathcal{D}})$  at some scale.

The main idea behind our bound is to show that differentially private learning algorithms necessarily perform poorly when there is a large set of hypotheses such that every pair in the set labels approximately  $1/\alpha$  examples differently. We then show that such large sets can be constructed when the doubling dimension of the disagreement metric  $(\mathcal{H}, \rho_{\mathcal{D}})$  is high.

### 5.1. Main results

**Theorem 10** *There exists a constant  $c > 0$  such that the following holds. Let  $\mathcal{H}$  be a hypothesis class with VC dimension  $V < \infty$ ,  $\mathcal{D}$  be a distribution over  $\mathcal{X}$ ,  $X$  be an i.i.d. sample from  $\mathcal{D}$  of size  $m$ , and  $\mathcal{A}$  be a learning algorithm that guarantees  $\alpha$ -label privacy and outputs a hypothesis in  $\mathcal{H}$ . Let  $d := \text{ddim}_{12\epsilon}(\mathcal{H}, \rho_{\mathcal{D}}) > 2$ , and  $d' := \inf\{\text{ddim}_{12r}(\mathcal{H}, \rho_{\mathcal{D}}) : \epsilon \leq r < \Delta/6\} > 2$ . If*

$$\epsilon < c \cdot \left( \frac{\Delta}{V(1 + \log(1/\Delta))} \right), \quad \alpha \leq c \cdot \left( \frac{d'}{V \log(1/\epsilon)} \right), \quad \text{and} \quad m < c \cdot \left( \frac{d}{\alpha \epsilon} \right)$$

where  $\Delta$  is the diameter of  $(\mathcal{H}, \rho_{\mathcal{D}})$ , then there exists a hypothesis  $h^* \in \mathcal{H}$  such that with probability at least  $1/8$  over the random choice of  $X$  and internal randomness of  $\mathcal{A}$ , the hypothesis  $h_{\mathcal{A}}$  returned by  $\mathcal{A}(S_{X, h^*})$  has classification error

$$\Pr_{x \sim \mathcal{D}} [h_{\mathcal{A}}(x) \neq h^*(x)] > \epsilon.$$

We note that the conditions on  $\alpha$  and  $\epsilon$  can be relaxed by replacing the VC dimension with other (possibly distribution-dependent) quantities that determine the uniform convergence of  $\rho_X$  to  $\rho_{\mathcal{D}}$ ; we used a distribution-free parameter to simplify the argument. Moreover, the condition on  $\epsilon$  can be reduced to  $\epsilon < c$  for some constant  $c \in (0, 1)$  provided that there exists a lower bound of  $\Omega(V/\epsilon)$  to (non-privately) learn  $\mathcal{H}$  under the distribution  $\mathcal{D}$ .

The proof of Theorem 10, which is in Appendix D, relies on the following lemma (possibly of independent interest) which gives a lower bound on the empirical error of the hypothesis returned by an  $\alpha$ -label private learning algorithm.

**Lemma 11** *Let  $X \subseteq \mathcal{X}$  be an unlabeled dataset of size  $m$ ,  $\mathcal{H}$  be a hypothesis class,  $\mathcal{A}$  be a learning algorithm that guarantees  $\alpha$ -label privacy, and  $s > 0$ . Pick any  $h_0 \in \mathcal{H}$ . If  $P$  is an  $s$ -packing of  $B_X(h_0, 4s) \subseteq \mathcal{H}$ , and*

$$m < \frac{\log \left( \frac{|P|}{2} - 1 \right)}{8\alpha s},$$

then there exists a subset  $Q \subseteq P$  such that

1.  $|Q| \geq |P|/2$ ;
2. for all  $h \in Q$ ,  $\Pr_{\mathcal{A}}[\mathcal{A}(S_{X, h}) \notin B_X(h, s/2)] \geq 1/2$ .

The proof of Lemma 11 is in Appendix D. The next theorem shows a lower bound without restrictions on  $\epsilon$  and  $\alpha$ . Moreover, this bound also applies when the VC dimension of the hypothesis class is unbounded. However, we note that this bound is weaker in that it does not involve a  $1/\epsilon$  factor, where  $\epsilon$  is the accuracy parameter.

**Theorem 12** Let  $\mathcal{H}$  be a hypothesis class,  $\mathcal{D}$  be a distribution over  $\mathcal{X}$ ,  $X$  be an i.i.d. sample from  $\mathcal{D}$  of size  $m$ , and  $\mathcal{A}$  be a learning algorithm that guarantees  $\alpha$ -label privacy and outputs a hypothesis in  $\mathcal{H}$ . Let  $d'' := \text{ddim}_{4\epsilon}(\mathcal{H}, \rho_{\mathcal{D}}) \geq 1$ . If  $\epsilon \leq \Delta/2$  and

$$m \leq \frac{(d'' - 1) \log 2}{\alpha}$$

where  $\Delta$  is the diameter of  $(\mathcal{H}, \rho_{\mathcal{D}})$ , then there exists  $h^* \in \mathcal{H}$  such that with probability at least  $1/2$  over the random choice of  $X$  and internal randomness of  $\mathcal{A}$ , the hypothesis  $h_{\mathcal{A}}$  returned by  $\mathcal{A}(S_{X, h^*})$  has classification error

$$\Pr_{x \sim \mathcal{D}} [h_{\mathcal{A}}(x) \neq h^*(x)] > \epsilon.$$

In other words, any  $\alpha$ -label private algorithm for learning a hypothesis in  $\mathcal{H}$  with error at most  $\epsilon \leq \Delta/2$  must use at least  $(d'' - 1) \log(2)/\alpha$  examples. Theorem 12 uses ideas similar to those in (Beimel et al., 2010), but the result is stronger in that it applies to  $\alpha$ -label privacy and continuous data domains. A detailed proof is provided in Appendix D.

## 5.2. Example: linear separators in $\mathbb{R}^n$

In this section, we show an example that illustrates our label privacy lower bounds. Our example hypothesis class  $\mathcal{H} := \mathcal{H}_{\text{linear}}$  is the class of linear separators over  $\mathbb{R}^n$  that pass through the origin, and the unlabeled data distribution  $\mathcal{D}$  is the uniform distribution over the unit sphere  $\mathbb{S}^{n-1}$ . By Lemma 25 (see Appendix D), the doubling dimension of  $(\mathcal{H}, \rho_{\mathcal{D}})$  at any scale  $r$  is at least  $n - 2$ . Therefore Theorem 10 implies that if  $\alpha$  and  $\epsilon$  are small enough, any  $\alpha$ -label private algorithm  $\mathcal{A}$  that correctly learns all hypotheses  $h \in \mathcal{H}$  with error  $\leq \epsilon$  requires at least  $\Omega(\frac{n}{\epsilon\alpha})$  examples. (In fact, the condition on  $\epsilon$  can be relaxed to  $\epsilon \leq c$  for some constant  $c \in (0, 1)$ , because  $\Omega(n)$  examples are needed to even non-privately learn in this setting (Long, 1995).) We also observe that this bound is tight (except for a  $\log(1/\delta)$  factor): as the doubling dimension of  $\mathcal{D}$  is at most  $n$ , in the realizable case, Algorithm  $\mathcal{A}_{1r}$  using  $\mathcal{U} := \mathcal{D}$  learns linear separators with  $\alpha$ -label privacy given  $\tilde{O}(\frac{n}{\alpha\epsilon})$  examples.

## Acknowledgments

KC would like to thank NIH U54 HL108460 for research support. DH was partially supported by AFOSR FA9550-09-1-0425, NSF IIS-1016061, and NSF IIS-713540.

## References

- R. Agrawal and R. Srikant. Privacy-preserving data mining. *SIGMOD Rec.*, 29(2):439–450, 2000. ISSN 0163-5808. doi: <http://doi.acm.org/10.1145/335191.335438>.
- Lars Backstrom, Cynthia Dwork, and Jon M. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *WWW*, pages 181–190. ACM, 2007. ISBN 978-1-59593-654-7.
- K. Ball. An elementary introduction to modern convex geometry. In Silvio Levy, editor, *Flavors of Geometry*, volume 31. 1997.

- B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *PODS*, pages 273–282, 2007.
- Amos Beimel, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. In *TCC*, pages 437–454, 2010.
- A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In *PODS*, pages 128–138, 2005.
- A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In R. E. Ladner and C. Dwork, editors, *STOC*, pages 609–618. ACM, 2008. ISBN 978-1-60558-047-0.
- Nader H. Bshouty, Yi Li, and Philip M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *J. Comput. Syst. Sci.*, 75(6):323–335, 2009.
- K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Learning concept classes with privacy. Manuscript, 2006.
- K. Chaudhuri, C. Monteleoni, and A. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 2011.
- Kamalika Chaudhuri and Nina Mishra. When random sampling preserves privacy. In Cynthia Dwork, editor, *CRYPTO*, volume 4117 of *Lecture Notes in Computer Science*, pages 198–213. Springer, 2006. ISBN 3-540-37432-9.
- Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, 2005.
- C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *3rd IACR Theory of Cryptography Conference*, pages 265–284, 2006.
- A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS*, pages 211–222, 2003.
- Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.
- Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, pages 265–273, 2008a.
- Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, pages 265–273, 2008b.
- Anupam Gupta, Robert Krauthgamer, and James R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *FOCS*, pages 534–543, 2003.
- Anupam Gupta, Moritz Hardt, Aaron Roth, and Jonathan Ullman. Privately releasing conjunctions and the statistical query barrier. In *STOC*, 2011.

- Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *STOC*, pages 705–714, 2010.
- Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. ”i know what you did last summer”: query logs and user privacy. In *CIKM ’07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 909–914, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9. doi: <http://doi.acm.org/10.1145/1321440.1321573>.
- S. A. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? In *Proc. of Foundations of Computer Science*, 2008.
- Shiva Prasad Kasiviswanathan and Adam Smith. A note on differential privacy: Defining resistance to arbitrary side information. *CoRR*, abs/0803.3946, 2008.
- A. Kolmogorov and V. Tikhomirov.  $\epsilon$ -entropy and  $\epsilon$ -capacity of sets in function spaces. *Translations of the American Mathematical Society*, 17:277–364, 1961.
- P. M. Long. On the sample complexity of pac learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.
- A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proc. of ICDE*, 2006.
- David A. McAllester. Some pac-bayesian theorems. In *COLT*, pages 230–234, 1998.
- Frank McSherry and Ilya Mironov. Differentially private recommender systems: building privacy into the net. In *KDD ’09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–636, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi: <http://doi.acm.org/10.1145/1557019.1557090>.
- Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103, 2007.
- Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125, Oakland, CA, USA., May 2008. IEEE Computer Society.
- Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In David S. Johnson and Uriel Feige, editors, *STOC*, pages 75–84. ACM, 2007. ISBN 978-1-59593-631-8.
- D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.
- Aaron Roth. Differential privacy and the fat-shattering dimension of linear queries. In *APPROX-RANDOM*, pages 683–695, 2010.
- L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. on Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiao yong Zhou. Learning your identity and disease from research papers: information leaks in genome wide association study. In *ACM Conference on Computer and Communications Security*, pages 534–544, 2009.

Andrew Chi-Chih Yao. Protocols for secure computations (extended abstract). In *FOCS*, pages 160–164, 1982.

Shuheng Zhou, Katrina Ligett, and Larry Wasserman. Differential privacy with compression. In *Proc. of ISIT*, 2009.

## Appendix A. Metric spaces

**Lemma 13 (Kolmogorov and Tikhomirov, 1961)** *For any metric space  $(\mathcal{Z}, \rho)$  with diameter  $\Delta$ , and any  $\varepsilon \in (0, \Delta)$ , there exists an  $\varepsilon$ -packing of  $(\mathcal{Z}, \rho)$  that is also an  $\varepsilon$ -cover.*

**Lemma 14 (Gupta, Krauthgamer, and Lee, 2003)** *For any  $\varepsilon > 0$  and  $r > 0$ , if a metric space  $(\mathcal{Z}, \rho)$  has doubling dimension  $d$  and  $z \in \mathcal{Z}$ , then every  $\varepsilon$ -packing of  $(B(z, r), \rho)$  has cardinality at most  $(4r/\varepsilon)^d$ .*

**Lemma 15** *Let  $(\mathcal{Z}, \rho)$  be a metric space with diameter  $\Delta$ , and  $r \in (0, 2\Delta)$ . If  $\text{ddim}_r(\mathcal{Z}, \rho) \geq d$ , then there exists  $z \in \mathcal{Z}$  such that  $B(z, r)$  has an  $(r/2)$ -packing of size at least  $2^d$ .*

**Proof** Fix  $r \in (0, 2\Delta)$  and a metric space  $(\mathcal{Z}, \rho)$  with diameter  $\Delta$ . Suppose that for every  $z \in \mathcal{Z}$ , every  $(r/2)$ -packing of  $B(z, r)$  has size less than  $2^d$ . For each  $z \in \mathcal{Z}$ , let  $P_z$  be an  $(r/2)$ -packing of  $(B(z, r), \rho)$  that is also an  $(r/2)$ -cover—this is guaranteed to exist by Lemma 13. Therefore, for each  $z \in \mathcal{Z}$ ,  $B(z, r) \subseteq \bigcup_{z' \in P_z} B(z', r/2)$ , and  $|P_z| < 2^d$ . This implies that  $\text{ddim}_r(\mathcal{Z}, \rho)$  is less than  $d$ . ■

## Appendix B. Uniform convergence

**Lemma 16 (Vapnik and Chervonenkis, 1971)** *Let  $\mathcal{F}$  be a family of measurable functions  $f : \mathcal{Z} \rightarrow \{0, 1\}$  over a space  $\mathcal{Z}$  with distribution  $\mathcal{D}_{\mathcal{Z}}$ . Denote by  $\mathbb{E}_{\mathcal{Z}}[f]$  the empirical average of  $f$  over a subset  $Z \subseteq \mathcal{Z}$ . Let  $\varepsilon_m := (4/m)(\log(\mathcal{S}_{\mathcal{F}}(2m)) + \log(4/\delta))$ , where  $\mathcal{S}_{\mathcal{F}}(n)$  is the  $n$ -th VC shatter coefficient with respect to  $\mathcal{F}$ . Let  $Z$  be an i.i.d. sample of size  $m$  from  $\mathcal{D}_{\mathcal{Z}}$ . With probability at least  $1 - \delta$ , for all  $f \in \mathcal{F}$ ,*

$$\mathbb{E}[f] \geq \mathbb{E}_{\mathcal{Z}}[f] - \min \left\{ \sqrt{\mathbb{E}_{\mathcal{Z}}[f]\varepsilon_m}, \sqrt{\mathbb{E}[f]\varepsilon_m} + \varepsilon_m \right\}.$$

*Also, with probability at least  $1 - \delta$ , for all  $f \in \mathcal{F}$ ,*

$$\mathbb{E}[f] \leq \mathbb{E}_{\mathcal{Z}}[f] + \min \left\{ \sqrt{\mathbb{E}_{\mathcal{Z}}[f]\varepsilon_m}, \sqrt{\mathbb{E}[f]\varepsilon_m} + \varepsilon_m \right\}.$$

**Lemma 17** Let  $\mathcal{H}$  be a hypothesis class with VC dimension  $V$ . Fix any  $\delta \in (0, 1)$ , and let  $X$  be an i.i.d. sample of size  $m \geq V/2$  from  $\mathcal{D}$ . Let  $\varepsilon_m := (8V \log(2em/V) + 4 \log(4/\delta))/m$ . With probability at least  $1 - \delta$ , for all pairs of hypotheses  $\{h, h'\} \subseteq \mathcal{H}$ ,

$$\rho_{\mathcal{D}}(h, h') \geq \rho_X(h, h') - \min \left\{ \sqrt{\rho_X(h, h')\varepsilon_m}, \sqrt{\rho_{\mathcal{D}}(h, h')\varepsilon_m} + \varepsilon_m \right\}.$$

Also, with probability at least  $1 - \delta$ , for all pairs of hypotheses  $\{h, h'\} \subseteq \mathcal{H}$ ,

$$\rho_X(h, h') \geq \rho_{\mathcal{D}}(h, h') - \sqrt{\rho_{\mathcal{D}}(h, h')\varepsilon_m}.$$

**Proof** This is an immediate consequence of Lemma 16 as applied to the function class  $\mathcal{F} := \{x \mapsto \mathbb{1}[h(x) \neq h'(x)] : h, h' \in \mathcal{H}\}$ , which has VC shatter coefficients  $\mathcal{S}_{\mathcal{F}}(2m) \leq \mathcal{S}_{\mathcal{H}}(2m)^2 \leq (2em/V)^{2V}$  by Sauer's Lemma. ■

## Appendix C. Proofs from Section 4

### C.1. Some lemmas

We first give two simple lemmas. The first one, Lemma 18 states some basic properties of the exponential mechanism.

**Lemma 18 (McSherry and Talwar, 2007)** Let  $I$  be a finite set of indices, and let  $a_i \in \mathbb{R}$  for all  $i \in I$ . Define the probability distribution  $p := (p_i : i \in I)$  where  $p_i \propto \exp(-a_i)$  for all  $i \in I$ . If  $j \in I$  is drawn at random according to  $p$ , then the following holds for any element  $i_0 \in I$  and any  $t \in \mathbb{R}$ .

1. Let  $i \in I$ . If  $a_i \geq t$ , then  $\Pr_{j \sim p}[j = i] \leq \exp(-(t - a_{i_0}))$ .
2.  $\Pr_{j \sim p}[a_j \geq a_{i_0} + t] \leq |I| \exp(-t)$ .

**Proof** Fix any  $i_0 \in I$  and  $t \in \mathbb{R}$ . To show the first part of the lemma, note that for any  $i \in I$  with  $a_i \geq t$ , we have

$$\Pr_{j \sim p}[j = i] = \frac{\exp(-a_i)}{\sum_{i' \in I} \exp(-a_{i'})} \leq \frac{\exp(-t)}{\exp(-a_{i_0})} = \exp(-(t - a_{i_0})).$$

For the second part, we apply the inequality from the first part to all  $i \in I$  such that  $a_i \geq a_{i_0} + t$ , so

$$\Pr_{j \sim p}[a_j \geq a_{i_0} + t] = \sum_{i \in I} \mathbb{1}[a_i \geq a_{i_0} + t] \cdot \Pr_{j \sim p}[j = i] \leq \sum_{i \in I} \mathbb{1}[a_i \geq a_{i_0} + t] \cdot \exp(-t) \leq |I| \exp(-t).$$

■

The next lemma is consequences of smoothness between distributions  $\mathcal{D}$  and  $\mathcal{U}$ .

**Lemma 19** If  $\mathcal{D}$  is  $\kappa$ -smooth with respect to  $\mathcal{U}$ , then for all  $\varepsilon > 0$ , every  $\varepsilon$ -cover of  $(\mathcal{H}, \rho_{\mathcal{U}})$  is a  $\kappa\varepsilon$ -cover of  $(\mathcal{H}, \rho_{\mathcal{D}})$ .

**Proof** Suppose  $C$  is an  $\varepsilon$ -cover of  $(\mathcal{H}, \rho_{\mathcal{U}})$ . Then, for any  $h \in \mathcal{H}$ , there exists  $h' \in C$  such that  $\rho_{\mathcal{U}}(h, h') \leq \varepsilon$ . Fix such a pair  $h, h'$ , and let  $A := \{x \in \mathcal{X} : h(x) \neq h'(x)\}$  be the subset of  $\mathcal{X}$  on which  $h$  and  $h'$  disagree. As  $\mathcal{D}$  is  $\kappa$ -smooth with respect to  $\mathcal{U}$ , by definition of smoothness,

$$\rho_{\mathcal{D}}(h, h') = \Pr_{x \sim \mathcal{D}}[x \in A] \leq \kappa \cdot \Pr_{x \sim \mathcal{D}}[x \in A] = \kappa \cdot \rho_{\mathcal{U}}(h, h') \leq \kappa\varepsilon,$$

and thus  $C$  is a  $\kappa\varepsilon$ -cover of  $(\mathcal{H}, \rho_{\mathcal{D}})$ . ■

## C.2. Proof of Theorem 8

First, because of the lower bound on  $m := |S|$  from (4), the computed value of  $\hat{\kappa}$  in the first step of the algorithm must satisfy  $\hat{\kappa} \geq \kappa$ . Therefore,  $\mathcal{D}$  is also  $\hat{\kappa}$ -smooth with respect to  $\mathcal{U}$ . Combining this with Lemma 19,  $\mathcal{G}$  is an  $(\epsilon/4)$ -cover of  $(\mathcal{H}, \rho_{\mathcal{D}})$ . Moreover, as  $\mathcal{G}$  is also an  $(\epsilon/4\hat{\kappa})$ -packing of  $\mathcal{U}$ , from Lemma 14, the cardinality of  $\mathcal{G}$  is at most  $|\mathcal{G}| \leq (16\hat{\kappa}/\epsilon)^{d_{\mathcal{U}}}$ .

Define  $\text{err}(h) := \Pr_{(x,y) \sim \mathcal{P}}[h(x) \neq y]$ . Suppose that  $h^* \in \mathcal{H}$  minimizes  $\text{err}(h)$  over  $h \in \mathcal{H}$ . Let  $g_0 \in \mathcal{G}$  be an element of  $\mathcal{G}$  such that  $\rho_{\mathcal{D}}(h^*, g_0) \leq \epsilon/4$ ;  $g_0$  exists as  $\mathcal{G}$  is an  $(\epsilon/4)$ -cover of  $(\mathcal{H}, \rho_{\mathcal{D}})$ . By the triangle inequality, we have that:

$$\text{err}(g_0) \leq \text{err}(h^*) + \rho_{\mathcal{D}}(h^*, g_0) \leq \text{err}(h^*) + \epsilon/4 \quad (6)$$

Let  $E$  be the event that  $\max_{g \in \mathcal{G}} |\text{err}(g) - \text{err}(g, S)| > \epsilon/4$ , and  $\bar{E}$  be its complement. By Hoeffding's inequality, a union bound, and the lower bound on  $|S|$ , we have that for a large enough value of the constant  $C$  in Equation (4),

$$\Pr_{S \sim \mathcal{P}^m}[E] \leq |\mathcal{G}| \max_{g \in \mathcal{G}} \Pr_{S \sim \mathcal{P}^m} \left[ |\text{err}(g) - \text{err}(g, S)| > \frac{\epsilon}{4} \right] \leq 2|\mathcal{G}| \exp \left( -\frac{|S|\epsilon^2}{32} \right) \leq \frac{\delta}{2}.$$

In the event  $\bar{E}$ , we have  $\text{err}(h_{\mathcal{A}}) \geq \text{err}(h_{\mathcal{A}}, S) - \epsilon/4$  and  $\text{err}(g_0) \leq \text{err}(g_0, S) + \epsilon/4$  because both  $h_{\mathcal{A}}$  and  $g_0$  are in  $\mathcal{G}$ . Therefore,

$$\begin{aligned} \Pr_{S \sim \mathcal{P}^m, \mathcal{A}_1} [\text{err}(h_{\mathcal{A}}) > \text{err}(h^*) + \epsilon] &\leq \Pr_{S \sim \mathcal{P}^m, \mathcal{A}_1} \left[ \text{err}(h_{\mathcal{A}}) > \text{err}(g_0) + \frac{3\epsilon}{4} \right] \\ &\leq \Pr_{S \sim \mathcal{P}^m} [E] + \Pr_{\mathcal{A}_1} \left[ \text{err}(h_{\mathcal{A}}) > \text{err}(g_0) + \frac{3\epsilon}{4} \mid \bar{E} \right] \\ &\leq \frac{\delta}{2} + \Pr_{\mathcal{A}_1} \left[ \text{err}(h_{\mathcal{A}}, S) > \text{err}(g_0, S) + \frac{\epsilon}{4} \mid \bar{E} \right] \\ &\leq \frac{\delta}{2} + |\mathcal{G}| \exp \left( -\frac{\alpha|S|\epsilon}{8} \right) \\ &\leq \frac{\delta}{2} + \left( \frac{16\hat{\kappa}}{\epsilon} \right)^{d_{\mathcal{U}}} \exp \left( -\frac{\alpha|S|\epsilon}{8} \right) \\ &\leq \delta \end{aligned}$$

Here, the first step follows from (7), and the final three inequalities follow from Lemma 18 (using  $a_g = \alpha|S| \text{err}(g, S)/2$  for  $g \in \mathcal{G}$ ), the upper bound on  $|\mathcal{G}|$ , and the lower bound on  $m$  in (4).

**Algorithm  $\mathcal{A}_{1r}$ .**

**Input:** private labeled dataset  $S \subseteq \mathcal{X} \times \{\pm 1\}$ , public reference distribution  $\mathcal{U}$  over  $\mathcal{X}$ , privacy parameter  $\alpha \in (0, 1)$ , accuracy parameter  $\epsilon \in (0, 1)$ , confidence parameter  $\delta \in (0, 1)$ .

**Output:**  $h_{\mathcal{A}} \in \mathcal{H}$ .

1. Solve the equation  $|S| = f_{\text{realizable}}(\alpha, \epsilon, \delta, \hat{\kappa})$  for  $\hat{\kappa} > 0$ , where

$$f_{\text{realizable}}(\alpha, \epsilon, \delta, \hat{\kappa}) = C \cdot \frac{1}{\alpha \epsilon} \cdot \left( d_{\mathcal{U}} \cdot \log(\hat{\kappa}/\epsilon) + \log \frac{1}{\delta} \right)$$

is the function from (5), and let  $\varepsilon_0 := \epsilon/(4\hat{\kappa})$ .

2. Let  $\mathcal{G}$  be an  $\varepsilon_0$ -packing of  $(\mathcal{H}, \rho_{\mathcal{U}})$  that is also an  $\varepsilon_0$ -cover; the existence of such a set is guaranteed by Lemma 13.
3. Randomly choose  $h_{\mathcal{A}} \in \mathcal{G}$  according to the distribution  $(p_g : g \in \mathcal{G})$ , where  $p_g \propto \exp(-\alpha|S|\text{err}(g, S)/2)$  for each  $g \in \mathcal{G}$ , and return  $h_{\mathcal{A}}$ .

Figure 2: Learning algorithm for  $\alpha$ -privacy under the realizable assumption.

### C.3. Proof of Theorem 9

The proof is very similar to the proof of Theorem 8.

First, because of the lower bound on  $m := |S|$  from (5), the computed value of  $\hat{\kappa}$  in the first step of the algorithm must satisfy  $\hat{\kappa} \geq \kappa$ . Therefore,  $\mathcal{D}$  is also  $\hat{\kappa}$ -smooth with respect to  $\mathcal{U}$ . Combining this with Lemma 19, as  $\mathcal{G}$  is an  $(\epsilon/4\hat{\kappa})$ -cover of  $\mathcal{U}$ ,  $\mathcal{G}$  is an  $(\epsilon/4)$ -cover of  $(\mathcal{H}, \rho_{\mathcal{D}})$ . Moreover, as  $\mathcal{G}$  is also an  $(\epsilon/4\hat{\kappa})$ -packing of  $\mathcal{U}$ , from Lemma 14, the cardinality of  $\mathcal{G}$  is at most  $|\mathcal{G}| \leq (16\hat{\kappa}/\epsilon)^{d_{\mathcal{U}}}$ .

Define  $\text{err}(h) := \Pr_{(x,y) \sim \mathcal{P}}[h(x) \neq y]$ . Suppose that  $h^* \in \mathcal{H}$  minimizes  $\text{err}(h)$  over  $h \in \mathcal{H}$ . Recall that from the realizability assumption,  $\text{err}(h^*) = 0$ . Let  $g_0 \in \mathcal{G}$  be an element of  $\mathcal{G}$  such that  $\rho_{\mathcal{D}}(h^*, g_0) \leq \epsilon/4$ ;  $g_0$  exists as  $\mathcal{G}$  is an  $(\epsilon/4)$ -cover of  $(\mathcal{H}, \rho_{\mathcal{D}})$ . By the triangle inequality, we have that:

$$\text{err}(g_0) \leq \text{err}(h^*) + \rho_{\mathcal{D}}(h^*, g_0) \leq \epsilon/4 \quad (7)$$

We define two events  $E_1$  and  $E_2$ . Let  $\mathcal{G}_1 \subset \mathcal{G}$  be the set of all  $g \in \mathcal{G}$  for which  $\text{err}(g) \geq \epsilon$ . The event  $E_1$  is the event that  $\min_{g \in \mathcal{G}_1} \text{err}(g, S) > 9\epsilon/10$ , and let  $\bar{E}_1$  be its complement. Applying the multiplicative Chernoff bounds, for a specific  $g \in \mathcal{G}_1$ ,

$$\Pr_{S \sim \mathcal{P}^m} \left[ \text{err}(g, S) < \frac{9}{10} \text{err}(g) \right] \leq e^{-|S|\text{err}(g)/400} \leq e^{-|S|\epsilon/400}.$$

The quantity on the right hand side is at most  $\frac{\delta}{4|\mathcal{G}|} \leq \frac{\delta}{4|\mathcal{G}_1|}$  for a large enough constant  $C$  in Equation (5). Applying an union bound over all  $g \in \mathcal{G}_1$ , we get that

$$\Pr_{S \sim \mathcal{P}^m} [\bar{E}_1] \leq \delta/4. \quad (8)$$

We define  $E_2$  as the event that  $\text{err}(g_0, S) \leq 3\epsilon/4$ , and  $\bar{E}_2$  as its complement. From a standard multiplicative Chernoff bound, with probability at least  $1 - \delta/4$ ,

$$\text{err}(g_0, S) \leq \text{err}(g) + \sqrt{\frac{3\text{err}(g) \ln(4/\delta)}{|S|}} \leq \frac{\epsilon}{4} + \sqrt{\frac{3\epsilon}{4} \cdot \frac{\ln(4/\delta)}{|S|}} \leq \frac{\epsilon}{4} + \sqrt{\frac{3\epsilon}{4} \cdot \frac{\epsilon}{3}} = \frac{3\epsilon}{4}$$

Thus, if  $|S| \geq (3/\epsilon) \log(4/\delta)$ , which is the case due to Equation (5),

$$\Pr_{S \sim \mathcal{P}^m}[\bar{E}_2] = \Pr_{S \sim \mathcal{P}^m} \left[ \text{err}(g_0, S) > \frac{3\epsilon}{4} \right] \leq \frac{\delta}{4}. \quad (9)$$

Therefore, we have

$$\begin{aligned} & \Pr_{S \sim \mathcal{P}^m, \mathcal{A}}[\text{err}(h_{\mathcal{A}}) > \epsilon] \\ & \leq \Pr_{S \sim \mathcal{P}^m, \mathcal{A}}[\text{err}(h_{\mathcal{A}}) > \epsilon \mid E_1 \cap E_2] + \Pr_{S \sim \mathcal{P}^m}[\bar{E}_1] + \Pr_{S \sim \mathcal{P}^m}[\bar{E}_2] \\ & \leq \Pr_{S \sim \mathcal{P}^m, \mathcal{A}} \left[ \text{err}(h_{\mathcal{A}}, S) > \text{err}(g_0, S) + \left( \frac{9}{10} - \frac{3}{4} \right) \epsilon \mid E_1 \cap E_2 \right] + \delta/4 + \delta/4 \\ & \leq \Pr_{S \sim \mathcal{P}^m, \mathcal{A}} \left[ \text{err}(h_{\mathcal{A}}, S) > \text{err}(g_0, S) + \frac{3\epsilon}{20} \mid E_1 \cap E_2 \right] + \delta/2 \\ & \leq |\mathcal{G}| \exp \left( -\frac{3\epsilon|S|}{20} \right) + \delta/2 \\ & \leq \left( \frac{16\hat{\kappa}}{\epsilon} \right)^{d_{\mathcal{U}}} \exp \left( -\frac{3\epsilon|S|}{20} \right) + \delta/2 \\ & \leq \delta/2 + \delta/2 = \delta. \end{aligned}$$

Here, the second step follows from the definition of events  $E_1$  and  $E_2$  and from Equations (8) and (9), the third step follows from simple algebra, the fourth step follows from Lemma 18, the fifth step from the bound on  $|\mathcal{G}|$  and the final step from Equation (5).

#### C.4. Examples

**Lemma 20** *Let  $\mathcal{U}$  be uniform over the unit sphere  $\mathbb{S}^{n-1}$ , and let  $\mathcal{D}$  be defined as in Example 2. Then,  $\mathcal{D}$  is*

$$\frac{1}{1 - 2 \exp(-n\gamma^2/2)}\text{-smooth}$$

with respect to  $\mathcal{U}$ .

**Proof** From (Ball, 1997), we know that  $\Pr_{x \sim \mathcal{U}}[x \in W] \geq 1 - 2 \exp(-n\gamma^2/2)$ . Thus, for any set  $A \subseteq \mathbb{S}^{n-1}$ , we have

$$\Pr_{x \sim \mathcal{D}}[x \in A] = \Pr_{x \sim \mathcal{D}}[x \in A \cap W] = \frac{\Pr_{x \sim \mathcal{U}}[x \in A \cap W]}{\Pr_{x \sim \mathcal{U}}[x \in W]} \leq \frac{\Pr_{x \sim \mathcal{U}}[x \in A]}{1 - 2 \exp(-n\gamma^2/2)}.$$

This means  $\mathcal{D}$  is  $\kappa$ -smooth with respect to  $\mathcal{U}$  for

$$\kappa = \frac{1}{1 - 2 \exp(-n\gamma^2/2)}.$$

■

**Lemma 21** Let  $\mathcal{U}$  be uniform over the unit sphere  $\mathbb{S}^{n-1}$  and let  $\mathcal{D}$  be defined as in Example 3. Then,  $\mathcal{D}$  is

$$\left(\frac{2}{1-\gamma}\right)^{\frac{n-1}{2}}\text{-smooth}$$

with respect to  $\mathcal{U}$ .

**Proof** From (Ball, 1997), we know that  $\Pr_{x \sim \mathcal{U}}[x \in \mathbb{S}^{n-1} \setminus W] = \Pr_{x \sim \mathcal{U}}[x \notin W] \geq ((1-\gamma)/2)^{(n-1)/2}$ . Therefore, for any  $A \subseteq \mathbb{S}^{n-1}$ , we have

$$\Pr_{x \sim \mathcal{D}}[x \in A] = \Pr_{x \sim \mathcal{D}}[x \in A \setminus W] = \frac{\Pr_{x \sim \mathcal{U}}[x \in A \setminus W]}{\Pr_{x \sim \mathcal{U}}[x \in \mathbb{S}^{n-1} \setminus W]} \leq \frac{\Pr_{x \sim \mathcal{U}}[x \in A]}{\left(\frac{1-\gamma}{2}\right)^{\frac{n-1}{2}}}.$$

This means  $\mathcal{D}$  is  $\kappa$ -smooth with respect to  $\mathcal{U}$  for

$$\kappa = \left(\frac{2}{1-\gamma}\right)^{\frac{n-1}{2}}.$$

■

## Appendix D. Proofs from Section 5

### D.1. Some lemmas

**Lemma 22** Let  $S := \{(x_1, y_1), \dots, (x_m, y_m)\} \subseteq \mathcal{X} \times \{\pm 1\}$  be a labeled dataset of size  $m$ ,  $\alpha \in (0, 1)$ , and  $k \geq 0$ .

1. If a learning algorithm  $\mathcal{A}$  guarantees  $\alpha$ -privacy and outputs a hypothesis from  $\mathcal{H}$ , then for all  $S' := \{(x'_1, y'_1), \dots, (x'_m, y'_m)\} \subseteq \mathcal{X} \times \{\pm 1\}$  with  $(x_i, y_i) = (x'_i, y'_i)$  for at least  $|S| - k$  such examples,

$$\forall \mathcal{G} \subseteq \mathcal{H} \ . \ \Pr_{\mathcal{A}}[\mathcal{A}(S) \in \mathcal{G}] \geq \Pr_{\mathcal{A}}[\mathcal{A}(S') \in \mathcal{G}] \cdot \exp(-k\alpha).$$

2. If a learning algorithm  $\mathcal{A}$  guarantees  $\alpha$ -label privacy and outputs a hypothesis from  $\mathcal{H}$ , then for all  $S' := \{(x_1, y'_1), \dots, (x_m, y'_m)\} \subseteq \mathcal{X} \times \{\pm 1\}$  with  $y_i = y'_i$  for at least  $|S| - k$  such labels,

$$\forall \mathcal{G} \subseteq \mathcal{H} \ . \ \Pr_{\mathcal{A}}[\mathcal{A}(S) \in \mathcal{G}] \geq \Pr_{\mathcal{A}}[\mathcal{A}(S') \in \mathcal{G}] \cdot \exp(-k\alpha).$$

**Proof** We prove just the first part, as the second part is similar. For a labeled dataset  $S'$  that differs from  $S$  in at most  $k$  pairs, there exists a sequence of datasets  $S^{(0)}, \dots, S^{(\ell)}$  with  $\ell \leq k$  such that  $S^{(0)} = S'$ ,  $S^{(\ell)} = S$ , and  $S^{(j)}$  differs from  $S^{(j+1)}$  in exactly one example for  $1 \leq j < \ell$ . In this case, if  $\mathcal{A}$  guarantees  $\alpha$ -privacy, then for all  $\mathcal{G} \subseteq \mathcal{H}$ ,

$$\begin{aligned} \Pr_{\mathcal{A}}[\mathcal{A}(S^{(0)}) \in \mathcal{G}] &\leq \Pr_{\mathcal{A}}[\mathcal{A}(S^{(1)}) \in \mathcal{G}] \cdot e^{\alpha} \\ &\leq \Pr_{\mathcal{A}}[\mathcal{A}(S^{(2)}) \in \mathcal{G}] \cdot e^{2\alpha} \\ &\leq \dots \\ &\leq \Pr_{\mathcal{A}}[\mathcal{A}(S^{(\ell)}) \in \mathcal{G}] \cdot e^{\ell\alpha} \end{aligned}$$

and therefore

$$\Pr_{\mathcal{A}} [\mathcal{A}(S) \in \mathcal{G}] \geq \Pr_{\mathcal{A}} [\mathcal{A}(S') \in \mathcal{G}] \cdot e^{-\ell\alpha} \geq \Pr_{\mathcal{A}} [\mathcal{A}(S') \in \mathcal{G}] \cdot e^{-k\alpha}.$$

■

**Lemma 23** *There exists a constant  $C > 1$  such that the following holds. Let  $\mathcal{H}$  be a hypothesis class with VC dimension  $V$ , and let  $\mathcal{D}$  be distribution over  $\mathcal{X}$ . Fix any  $r \in (0, 1)$ , and let  $X$  be an i.i.d. sample of size  $m$  from  $\mathcal{D}$ . If*

$$m \geq \frac{CV}{r} \log \frac{C}{r},$$

*then the following holds with probability at least  $1/2$ :*

1. *every pair of hypotheses  $\{h, h'\} \subseteq \mathcal{H}$  for which  $\rho_X(h, h') > 2r$  has  $\rho_{\mathcal{D}}(h, h') > r$ ;*
2. *for all  $h_0 \in \mathcal{H}$ , every  $(6r)$ -packing of  $(B_{\mathcal{D}}(h_0, 12r), \rho_{\mathcal{D}})$  is a  $(4r)$ -packing of  $(B_X(h_0, 16r), \rho_X)$ .*

**Proof** This is a consequence of Lemma 17. To show the first part, we plug in Lemma 17 with  $\varepsilon_m = r/2$ .

To show the second part, we use two applications of Lemma 17. Let  $h$  and  $h'$  be any two hypotheses in any  $(6r)$ -packing of  $(B_{\mathcal{D}}(h_0, 12r), \rho_{\mathcal{D}})$ ; we first use Lemma 17 with  $\varepsilon_m = r/3$  to show that for all such  $h$  and  $h'$ ,  $\rho_X(h, h') > 4r$ . Next we need to show that all  $h$  in any  $(6r)$ -packing of  $(B_{\mathcal{D}}(h_0, 12r), \rho_{\mathcal{D}})$  has  $\rho_X(h, h_0) \leq 16r$ ; we show this through a second application of Lemma 17 with  $\varepsilon_m = r/3$ . ■

## D.2. Proof of Theorem 12

We prove the contrapositive: that if  $\epsilon \leq \Delta/2$  and  $\Pr_{X \sim \mathcal{D}^m, \mathcal{A}} [\mathcal{A}(S_{X,h^*}) \in B_{\mathcal{D}}(h^*, \epsilon)] > 1/2$  for all  $h^* \in \mathcal{H}$ , then  $m > \log(2^{d''-1})/\alpha$ . So pick any  $\epsilon \leq \Delta/2$ . By Lemma 15, there exists an  $h_0 \in \mathcal{H}$  and  $P \subseteq \mathcal{H}$  such that  $P$  is a  $(2\epsilon)$ -packing of  $(B_{\mathcal{D}}(h_0, 4\epsilon), \rho_{\mathcal{D}})$  of size  $\geq 2^{d''}$ . For any  $h, h' \in P$  such that  $h \neq h'$ , we have  $B_{\mathcal{D}}(h, \epsilon) \cap B_{\mathcal{D}}(h', \epsilon) = \emptyset$  by the triangle inequality. Therefore for any  $h \in P$  and any  $X' \subseteq \mathcal{X}$  of size  $m$ ,

$$\begin{aligned} \Pr_{\mathcal{A}} [A(S_{X',h}) \notin B_{\mathcal{D}}(h, \epsilon)] &\geq \sum_{h' \in P \setminus \{h\}} \Pr_{\mathcal{A}} [A(S_{X',h}) \in B_{\mathcal{D}}(h', \epsilon)] \\ &\geq \sum_{h' \in P \setminus \{h\}} \Pr_{\mathcal{A}} [A(S_{X',h'}) \in B_{\mathcal{D}}(h', \epsilon)] \cdot e^{-\alpha m}, \end{aligned}$$

where the second inequality follows by Lemma 22 because  $S_{X',h}$  and  $S_{X',h'}$  can differ in at most (all)  $m$  labels. Now integrating both sides with respect to  $X' \sim \mathcal{D}^m$  shows that if  $\Pr_{X \sim \mathcal{D}^m, \mathcal{A}} [\mathcal{A}(S_{X,h^*}) \in B_{\mathcal{D}}(h^*, \epsilon)] > 1/2$  for all  $h^* \in \mathcal{H}$ , then for any  $h \in P$ ,

$$\begin{aligned} \frac{1}{2} > \Pr_{X \sim \mathcal{D}^m, \mathcal{A}} [A(S_{X,h}) \notin B_{\mathcal{D}}(h, \epsilon)] &\geq \sum_{h' \in P \setminus \{h\}} \Pr_{X \sim \mathcal{D}^m, \mathcal{A}} [A(S_{X',h'}) \in B_{\mathcal{D}}(h', \epsilon)] \cdot e^{-\alpha m} \\ &> (|P| - 1) \cdot \frac{1}{2} \cdot e^{-\alpha m} \end{aligned}$$

which in turn implies  $m > \log(|P| - 1)/\alpha \geq \log(2^{d''} - 1)/\alpha \geq \log(2^{d''-1})/\alpha$ , as  $d''$  is always  $\geq 1$ .

### D.3. Proof of Lemma 11

Let  $h_0 \in \mathcal{H}$  and  $P$  be an  $s$ -packing of  $B_X(h_0, 4s) \subseteq \mathcal{H}$ . Say the algorithm  $\mathcal{A}$  is *good* for  $h$  if  $\Pr_{\mathcal{A}}[\mathcal{A}(S_{X,h}) \in B_X(h, s/2)] \geq 1/2$ . Note that  $\mathcal{A}$  is not good for  $h \in P$  if and only if  $\Pr_{\mathcal{A}}[\mathcal{A}(S_{X,h}) \notin B_X(h, s/2)] > 1/2$ . Therefore, it suffices to show that if  $\mathcal{A}$  is good for at least  $|P|/2$  hypotheses in  $P$ , then  $m \geq (\log((|P|/2) - 1)/(8\alpha s)$ .

By the triangle inequality and the fact that  $P$  is an  $s$ -packing,  $B_X(h, s/2) \cap B_X(h', s/2) = \emptyset$  for all  $h, h' \in P$  such that  $h \neq h'$ . Therefore for any  $h \in P$ ,

$$\Pr_{\mathcal{A}} [\mathcal{A}(S_{X,h}) \notin B_X(h, s/2)] \geq \sum_{h' \in P \setminus \{h\}} \Pr_{\mathcal{A}} [\mathcal{A}(S_{X,h}) \in B_X(h', s/2)].$$

Moreover, for all  $h, h' \in P$ , we have  $\rho_X(h, h') \leq \rho_X(h_0, h) + \rho_X(h_0, h') \leq 8s$  by the triangle inequality, so  $S_{X,h}$  and  $S_{X,h'}$  differ in at most  $8sm$  labels. Therefore Lemma 22 implies

$$\Pr_{\mathcal{A}} [\mathcal{A}(S_{X,h}) \in B_X(h', s/2)] \geq \Pr_{\mathcal{A}} [\mathcal{A}(S_{X,h'}) \in B_X(h', s/2)] \cdot e^{-8sm}$$

for all  $h, h' \in P$ . If  $\mathcal{A}$  is good for at least  $|P|/2$  hypotheses  $h' \in P$ , then for any  $h \in P$  such that  $\mathcal{A}$  is good for  $h$ , we have

$$\begin{aligned} \frac{1}{2} &\geq \Pr_{\mathcal{A}} [\mathcal{A}(S_{X,h}) \notin B_X(h, s/2)] \\ &\geq \sum_{h' \in P \setminus \{h\}} \mathbf{1}[\mathcal{A} \text{ is good for } h] \cdot \Pr_{\mathcal{A}} [\mathcal{A}(S_{X,h'}) \in B_X(h', s/2)] \cdot e^{-8sm} \\ &\geq \sum_{h' \in P \setminus \{h\}} \mathbf{1}[\mathcal{A} \text{ is good for } h] \cdot \frac{1}{2} \cdot e^{-8sm} \\ &\geq \left(\frac{|P|}{2} - 1\right) \cdot \frac{1}{2} \cdot e^{-8sm} \end{aligned}$$

which in turn implies  $m \geq \log((|P|/2) - 1)/(8s)$ .

### D.4. Proof of Theorem 10

We need the following lemma.

**Lemma 24** *There exists a constant  $C > 1$  such that the following holds. Let  $\mathcal{H}$  be a hypothesis class with VC dimension  $V$ ,  $\mathcal{D}$  be a distribution over  $\mathcal{X}$ ,  $X$  be an i.i.d. sample from  $\mathcal{D}$  of size  $m$ ,  $\mathcal{A}$  be a learning algorithm that guarantees  $\alpha$ -label privacy and outputs a hypothesis in  $\mathcal{H}$ , and  $\Delta$  be the diameter of  $(\mathcal{H}, \rho_{\mathcal{D}})$ . If  $r \in (0, \Delta/6)$  and*

$$\frac{CV}{r} \log \frac{C}{r} \leq m < \frac{\log(2^{d-1} - 1)}{32\alpha r}$$

where  $d := \text{ddim}_{12r}(\mathcal{H}, \rho_{\mathcal{D}})$ , then there exists a hypothesis  $h^* \in \mathcal{H}$  such that

$$\Pr_{X \sim \mathcal{D}^m, \mathcal{A}} [\mathcal{A}(S_{X,h^*}) \notin B_{\mathcal{D}}(h^*, r)] \geq \frac{1}{8}.$$

**Proof** First, assume  $r$  and  $m$  satisfy the conditions in the lemma statement, where  $C$  is the constant from Lemma 23. Also, let  $h_0 \in \mathcal{H}$  and  $P \subseteq \mathcal{H}$  be such that  $P$  is a  $(6r)$ -packing of  $(B_{\mathcal{D}}(h_0, 12r), \rho_{\mathcal{D}})$  of size  $|P| \geq 2^d$ ; the existence of such an  $h_0$  and  $P$  is guaranteed by Lemma 15.

We first define some events in the sample space of  $X$  and  $\mathcal{A}$ . For each  $h \in \mathcal{H}$ , and a sample  $X$ , let  $E_1(h, X)$  be the event that

$$\mathcal{A}(S_{X,h}) \text{ makes more than } 2rm \text{ mistakes on } S_{X,h} \text{ (i.e., } \rho_X(h, \mathcal{A}(S_{X,h})) > 2r).$$

Given a sample  $X$ , let  $\phi(X)$  be a 0/1 random variable which is 1 when the following conditions hold:

1. every pair of hypotheses  $\{h, h'\} \subseteq \mathcal{H}$  for which  $\rho_X(h, h') > 2r$  has  $\rho_{\mathcal{D}}(h, h') > r$ ; and
2. for all  $h_0 \in \mathcal{H}$ , every  $(6r)$ -packing of  $(B_{\mathcal{D}}(h_0, 12r), \rho_{\mathcal{D}})$  is a  $(4r)$ -packing of  $(B_X(h_0, 16r), \rho_X)$

(i.e., the conclusion of Lemma 23). Note that conditioned on  $E_1(h, X)$  and  $\phi(X) = 1$ , we have  $\rho_X(h, \mathcal{A}(S_{X,h})) > 2r$  and thus  $\rho_{\mathcal{D}}(h, \mathcal{A}(S_{X,h})) > r$ , so  $\Pr_{X \sim \mathcal{D}^m, \mathcal{A}}[E_1(h, X), (\phi(X) = 1)] \leq \Pr_{X \sim \mathcal{D}^m, \mathcal{A}}[\rho_{\mathcal{D}}(h, \mathcal{A}(S_{X,h})) > r]$ . Therefore it suffices to show that there exists  $h^* \in \mathcal{H}$  such that  $\Pr_{X \sim \mathcal{D}^m, \mathcal{A}}[E_1(h^*, X), (\phi(X) = 1)] \geq 1/8$ .

The lower bound on  $m$  and Lemma 23 ensure that

$$\Pr_{X \sim \mathcal{D}^m}[\phi(X) = 1] \geq \frac{1}{2}. \quad (10)$$

Also, if the unlabeled sample  $X$  is such that  $\phi(X) = 1$  holds, then the set  $P$  is a  $(4r)$ -packing of  $(B_X(h_0, 16r), \rho_X)$ . Therefore, the upper bound on  $m$  and Lemma 11 (with  $s = 4r$ ) imply that for all such  $X$ , there exists  $Q \subseteq P$  of size at least  $|P|/2$  such that  $\Pr_{\mathcal{A}}[E_1(h, X) \mid \phi(X) = 1] \geq 1/2$  for all  $h \in Q$ . In other words,

$$\sum_{h \in P} \Pr_{\mathcal{A}}[E_1(h, X) \mid \phi(X) = 1] = \sum_{h \in P} \mathbb{E}_{\mathcal{A}}[\mathbb{1}[E_1(h, X)] \mid \phi(X) = 1] \geq \frac{|P|}{4}. \quad (11)$$

Combining Equations (10) and (11) gives

$$\begin{aligned} & \sum_{h \in P} \Pr_{X \sim \mathcal{D}^m, \mathcal{A}}[E_1(h, X), \phi(X) = 1] \\ &= \sum_{h \in P} \mathbb{E}_{X \sim \mathcal{D}^m, \mathcal{A}}[\mathbb{1}[E_1(h, X)] \mid \phi(X) = 1] \cdot \Pr_{X \sim \mathcal{D}^m}[\phi(X) = 1] \\ &= \sum_{h \in P} \mathbb{E}_{X \sim \mathcal{D}^m} \mathbb{E}_{\mathcal{A}}[\mathbb{1}[E_1(h, X)] \mid \phi(X) = 1] \cdot \Pr_{X \sim \mathcal{D}^m}[\phi(X) = 1] \\ &= \mathbb{E}_{X \sim \mathcal{D}^m} \left[ \sum_{h \in P} \mathbb{E}_{\mathcal{A}}[\mathbb{1}[E_1(h, X)] \mid \phi(X) = 1] \right] \cdot \Pr_{X \sim \mathcal{D}^m}[\phi(X) = 1] \\ &\geq \frac{|P|}{4} \Pr_{X \sim \mathcal{D}^m}[\phi(X) = 1] \\ &\geq \frac{|P|}{8}. \end{aligned}$$

Here the first step follows because  $\phi(X)$  is a 0/1 random variable, the fourth step follows from Equation (11) and the fifth step follows from Equation (10).

Therefore there exists some  $h^* \in P$  such that  $\Pr_{X \sim \mathcal{D}^m, \mathcal{A}}[E_1(h^*, X), \phi(X) = 1] \geq 1/8$ . ■

**Proof** [Proof of Theorem 10] Assume

$$\epsilon < \frac{\Delta}{24CV \log(6C/\Delta)}, \quad \alpha \leq \frac{\log(2^{d'-1} - 1)}{32CV \log(C/\epsilon)}, \quad \text{and} \quad m < \frac{\log(2^{d-1} - 1)}{32\alpha\epsilon}$$

where  $C$  is the constant from Lemma 24. The proof is by case analysis, based on the value of  $m$ .

*Case 1:*  $m < 1/(4\epsilon)$ .

Since  $\epsilon < \Delta/2$ , Lemma 15 implies that there exists a pair  $\{h, h'\} \subseteq \mathcal{H}$  such that  $\rho_{\mathcal{D}}(h, h') > 2\epsilon$  but  $\rho_{\mathcal{D}}(h, h') \leq 4\epsilon$ . Using the bound on  $m$  and the fact  $\epsilon \leq 1/5$ , we have

$$\Pr_{X \sim \mathcal{D}^m}[\rho_X(h, h') = 0] \geq (1 - 4\epsilon)^m \geq (1 - 4\epsilon)^{\frac{1}{4\epsilon}} > \frac{1}{8}.$$

This means that  $\Pr_{X \sim \mathcal{D}^m}[h_{\mathcal{A}} := \mathcal{A}(S_{X,h}) = \mathcal{A}(S_{X,h'})] \geq 1/8$ . By the triangle inequality,  $B_{\mathcal{D}}(h, \epsilon) \cap B_{\mathcal{D}}(h', \epsilon) = \emptyset$ . So if, say,  $\Pr_{X \sim \mathcal{D}^m, \mathcal{A}}[h_{\mathcal{A}} \in B_{\mathcal{D}}(h, \epsilon)] \geq 1/8$ , then  $\Pr_{X \sim \mathcal{D}^m, \mathcal{A}}[h_{\mathcal{A}} \notin B_{\mathcal{D}}(h', \epsilon)] \geq 1/8$ . Therefore  $\Pr_{X \sim \mathcal{D}^m, \mathcal{A}}[h_{\mathcal{A}} \notin B_{\mathcal{D}}(h^*, \epsilon)] \geq 1/8$  for at least one  $h^* \in \{h, h'\}$ .

*Case 2:*  $1/(4\epsilon) \leq m < (CV/\epsilon) \log(C/\epsilon)$ .

First, let  $r > 0$  be the solution to the equation  $(CV/r) \log(C/r) = m$ , so  $r > \epsilon$ . Moreover, the bound on  $m$  and  $\epsilon$  imply

$$m \geq \frac{1}{4\epsilon} > \frac{CV}{\Delta/6} \log \frac{C}{\Delta/6}$$

so  $r < \Delta/6$ . Finally, using the bound on  $\alpha$ , definition of  $d'$ , and fact  $r > \epsilon$ , we have

$$\alpha \leq \frac{\log(2^{d'-1} - 1)}{32CV \log \frac{C}{\epsilon}} < \frac{\log(2^{d''-1} - 1)}{32CV \log \frac{C}{r}}$$

where  $d'' := \text{ddim}_{12r}(\mathcal{H}, \rho_{\mathcal{D}})$ ; this implies

$$m = \frac{CV}{r} \log \frac{C}{r} < \frac{\log(2^{d''-1} - 1)}{32\alpha r}.$$

The conditions of Lemma 24 are thus satisfied, which means there exists  $h^* \in \mathcal{H}$  such that  $\Pr_{X \sim \mathcal{D}^m, \mathcal{A}}[\mathcal{A}(S_{X,h^*}) \notin B_{\mathcal{D}}(h^*, r)] \geq 1/8$ .

*Case 3:*  $(CV/\epsilon) \log(C/\epsilon) \leq m < \log(2^{d-1} - 1)/(32\alpha\epsilon)$ .

The conditions of Lemma 24 are satisfied in this case with  $r := \epsilon < \Delta/6$ , so there exists  $h^* \in \mathcal{H}$  such that  $\Pr_{X \sim \mathcal{D}^m, \mathcal{A}}[\rho_{\mathcal{D}}(h^*, \mathcal{A}(S_{X,h^*})) > \epsilon] \geq 1/8$ . ■

### D.5. Example

The following lemma shows that if  $\mathcal{D}$  is the uniform distribution on  $\mathbb{S}^{n-1}$ , then  $\text{ddim}_r(\mathcal{H}, \rho_{\mathcal{D}}) \geq n - 2$  for all scales  $r > 0$ .

**Lemma 25** *Let  $\mathcal{H} := \mathcal{H}_{\text{linear}}$  be the class of linear separators through the origin in  $\mathbb{R}^n$  and  $\mathcal{D}$  be the uniform distribution on  $\mathbb{S}^{n-1}$ . For any  $u \in \mathbb{S}^{n-1}$  and any  $r > 0$ , there exists an  $(r/2)$ -packing of  $(B_{\mathcal{D}}(h_u, r), \rho_{\mathcal{D}})$  of size at least  $2^{n-2}$ .*

**Proof** Let  $\mu$  be the uniform distribution over  $\mathcal{H}$ ; notice that this is also the uniform distribution over  $\mathbb{S}^{n-1}$ .

We call a pair hypotheses  $h_v$  and  $h_w$  in  $\mathcal{H}$  *close* if  $\rho_{\mathcal{D}}(h_v, h_w) \leq r/2$ . Observe that if any set of hypotheses has no close pairs, then it is an  $(r/2)$ -packing.

Using a technique due to Long (1995), we now construct an  $(r/2)$ -packing of  $B_{\mathcal{D}}(h_u, r)$  by first randomly choosing hypotheses in  $B_{\mathcal{D}}(h_u, r)$ , and then removing hypotheses until no close pairs remain. First, we bound the probability  $p$  that two hypotheses  $h_v$  and  $h_w$ , chosen independently and uniformly at random from  $B_{\mathcal{D}}(h_u, r)$ , are close:

$$\begin{aligned} p &= \Pr_{(h_v, h_w) \sim \mu^2} [\rho_{\mathcal{D}}(h_v, h_w) \leq r/2 \mid h_v \in B_{\mathcal{D}}(h_u, r) \wedge h_w \in B_{\mathcal{D}}(h_u, r)] \\ &= \frac{\Pr_{(h_v, h_w) \sim \mu^2} [\rho_{\mathcal{D}}(h_v, h_w) \leq r/2 \wedge h_v \in B_{\mathcal{D}}(h_u, r) \mid h_w \in B_{\mathcal{D}}(h_u, r)]}{\Pr_{h_v \sim \mu} [h_v \in B_{\mathcal{D}}(h_u, r)]} \\ &\leq \frac{\Pr_{(h_v, h_w) \sim \mu^2} [\rho_{\mathcal{D}}(h_v, h_w) \leq r/2 \mid h_w \in B_{\mathcal{D}}(h_u, r)]}{\Pr_{h_v \sim \mu} [h_v \in B_{\mathcal{D}}(h_u, r)]} \\ &= \frac{\Pr_{(h_v, h_w) \sim \mu^2} [h_v \in B_{\mathcal{D}}(h_w, r/2) \mid h_w \in B_{\mathcal{D}}(h_u, r)]}{\Pr_{h_v \sim \mu} [h_v \in B_{\mathcal{D}}(h_u, r)]} \\ &= \frac{\Pr_{(h_v, h_w) \sim \mu^2} [h_v \in B_{\mathcal{D}}(h_u, r/2)]}{\Pr_{h_v \sim \mu} [h_v \in B_{\mathcal{D}}(h_u, r)]} \\ &= 2^{-(n-1)}. \end{aligned}$$

where the second-to-last equality follows by symmetry, and the last equality follows by the fact that  $B_{\mathcal{D}}(h_u, r)$  corresponds to a  $(n-1)$ -dimensional spherical cap of  $\mathbb{S}^{n-1}$ . Now, choose  $N := 2^{n-1}$  hypotheses  $h_{v_1}, \dots, h_{v_N}$  independently and uniformly at random from  $B_{\mathcal{D}}(h_u, r)$ . The expected number of close pairs among these  $N$  hypotheses is

$$\begin{aligned} M &:= \mathbb{E} \left[ \sum_{i < j} \mathbb{1}[h_{v_i} \text{ and } h_{v_j} \text{ are close}] \right] \\ &= \sum_{i < j} \Pr [h_{v_i} \text{ and } h_{v_j} \text{ are close}] \\ &= \sum_{i < j} p \\ &\leq \binom{N}{2} \cdot 2^{-(n-1)}. \end{aligned}$$

Therefore, there exists  $N$  hypotheses  $h_{v_1}, \dots, h_{v_N}$  in  $B_{\mathcal{D}}(h_u, r)$  among which there are at most  $M$  close pairs. Removing one hypothesis from each such close pair leaves a set of at

**Algorithm  $\mathcal{A}_2$ .**

**Input:** private labeled dataset  $S \subseteq \mathcal{X} \times \{\pm 1\}$ , privacy parameter  $\alpha \in (0, 1)$ , accuracy parameter  $\epsilon \in (0, 1)$ , confidence parameter  $\delta \in (0, 1)$ .

**Output:**  $h_{\mathcal{A}} \in \mathcal{H}$ .

1. Let  $\mathcal{G}$  be a  $(\epsilon/4)$ -cover for  $(\mathcal{H}, \rho_X)$  that is also an  $(\epsilon/4)$ -packing.
2. Randomly choose  $h_{\mathcal{A}} \in \mathcal{G}$  according to the distribution  $(p_g : g \in \mathcal{G})$ , where  $p_g \propto \exp(-\alpha|S|\text{err}(g, S)/2)$  for each  $g \in \mathcal{G}$ , and return  $h_{\mathcal{A}}$ .

Figure 3: Learning algorithm for  $\alpha$ -label privacy.

least  $N - M$  hypotheses with no close pairs—this is our  $(r/2)$ -packing of  $B_{\mathcal{D}}(h_u, r)$ . Since  $N = 2^{n-1}$ , the cardinality of this packing is at least

$$N - M \geq 2^{n-1} - \frac{2^{n-1} (2^{n-1} - 1)}{2} \cdot 2^{-(n-1)} > 2^{n-2}. \quad \blacksquare$$

## Appendix E. Upper bounds for learning with $\alpha$ -label privacy

Algorithm  $\mathcal{A}_2$  for learning with  $\alpha$ -label privacy, given in Figure 3, differs from the algorithms for learning with  $\alpha$ -privacy in that it is able to use the unlabeled data itself to construct a finite set of candidate hypotheses. The algorithm and its analysis are very similar to work due to Chaudhuri et al. (2006); we give the details for completeness.

**Theorem 26** *Algorithm  $\mathcal{A}_2$  preserves  $\alpha$ -label privacy.*

**Proof** The algorithm only accesses the labels in  $S$  in the final step. It follows from standard arguments in (McSherry and Talwar, 2007) that  $\alpha$ -label privacy is guaranteed.  $\blacksquare$

**Theorem 27** *Let  $\mathcal{P}$  be any probability distribution over  $\mathcal{X} \times \{\pm 1\}$  whose marginal over  $\mathcal{X}$  is  $\mathcal{D}$ . There exists a universal constant  $C > 0$  such that for any  $\alpha, \epsilon, \delta \in (0, 1)$ , the following holds. If  $S \subseteq \mathcal{X} \times \{\pm 1\}$  is an i.i.d. random sample from  $\mathcal{P}$  of size*

$$m \geq C \cdot \left( \frac{\eta}{\epsilon^2} + \frac{1}{\epsilon} \right) \cdot \left( V \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right) + \frac{C}{\alpha \epsilon} \cdot \log \frac{\mathbb{E}_{X \sim \mathcal{D}^m} [\mathcal{N}_{\epsilon/8}(\mathcal{H}, \rho_X)]}{\delta}$$

where  $\eta := \inf_{h' \in \mathcal{H}} \Pr_{(x,y) \sim \mathcal{P}} [h'(x) \neq y]$  and  $V$  is the VC dimension of  $\mathcal{H}$ ; then with probability at least  $1 - \delta$ , the hypothesis  $h_{\mathcal{A}} \in \mathcal{H}$  returned by  $\mathcal{A}_2(S, \alpha, \epsilon, \delta)$  satisfies

$$\Pr_{(x,y) \sim \mathcal{P}} [h_{\mathcal{A}}(x) \neq y] \leq \eta + \epsilon.$$

**Remark 28** The first term in the sample size requirement (which depends on VC dimension) can be replaced by distribution-based quantities used for characterizing uniform convergence such as those based on  $l_1$ -covering numbers (Pollard, 1984).

**Proof** Let  $\text{err}(h) := \Pr_{(x,y) \sim \mathcal{P}}[h(x) \neq y]$ , and let  $h^* \in \mathcal{H}$  minimize  $\text{err}(h)$  over  $h \in \mathcal{H}$ . Let  $S := \{(x_1, y_1), \dots, (x_m, y_m)\}$  be the i.i.d. sample drawn from  $\mathcal{P}^m$ , and  $X := \{x_1, \dots, x_m\}$  be the unlabeled components of  $S$ . Let  $g_0 \in \mathcal{G}$  minimize  $\text{err}(g, S)$  over  $g \in \mathcal{G}$ . Since  $\mathcal{G}$  is an  $(\epsilon/4)$ -cover for  $(\mathcal{H}, \rho_X)$ , we have that  $\text{err}(g_0, S) \leq \inf_{h' \in \mathcal{H}} \text{err}(h', S) + \epsilon/4$ . Since  $\mathcal{G}$  is also an  $(\epsilon/4)$ -packing for  $(\mathcal{H}, \rho_X)$ , we have that  $|\mathcal{G}| \leq \mathcal{N}_{\epsilon/8}(\mathcal{H}, \rho_X)$  (Pollard, 1984). Let  $\mathcal{F} := \{f_h : h \in \mathcal{H}\}$  where  $f_h(x, y) := \mathbb{1}[h(x) \neq y]$ . We have  $\mathbb{E}_{(x,y) \sim \mathcal{P}}[f_h(x, y)] = \text{err}(h)$  and  $m^{-1} \sum_{(x,y) \in S} f_h(x, y) = \text{err}(h, S)$ . Let  $E$  be the event that for all  $h \in \mathcal{H}$ ,

$$\text{err}(h, S) \leq \text{err}(h) + \sqrt{\text{err}(h)\varepsilon_m} + \varepsilon_m \quad \text{and} \quad \text{err}(h) \leq \text{err}(h, S) + \sqrt{\text{err}(h)\varepsilon_m}$$

where  $\varepsilon_m := (8V \log(2em/V) + 4 \log(16/\delta))/m$ . By Lemma 16, the fact  $\mathcal{S}(\mathcal{F}, n) = \mathcal{S}(\mathcal{H}, n)$ , and union bounds, we have  $\Pr_{S \sim \mathcal{P}^m}[E] \geq 1 - \delta/2$ . Now let  $E'$  be the event that

$$\text{err}(h_{\mathcal{A}}, S) \leq \text{err}(g_0, S) + t_m$$

where  $t_m := 2 \log(2\mathbb{E}_{X \sim \mathcal{D}^m}[|\mathcal{G}|]/\delta)/(\alpha m)$ . The probability of  $E'$  can be bounded as

$$\begin{aligned} \Pr_{S \sim \mathcal{P}^m, \mathcal{A}}[E'] &= 1 - \Pr_{S \sim \mathcal{P}^m, \mathcal{A}}[\text{err}(h_{\mathcal{A}}, S) > \text{err}(g_0, S) + t_m] \\ &= 1 - \mathbb{E}_{S \sim \mathcal{P}^m} [\Pr_{\mathcal{A}}[\text{err}(h_{\mathcal{A}}, S) > \text{err}(g_0, S) + t_m \mid S]] \\ &\geq 1 - \mathbb{E}_{S \sim \mathcal{P}^m} \left[ |\mathcal{G}| \exp\left(-\frac{\alpha m t_m}{2}\right) \right] \\ &= 1 - \mathbb{E}_{X \sim \mathcal{D}^m}[|\mathcal{G}|] \cdot \exp\left(-\frac{\alpha m t_m}{2}\right) \\ &\geq 1 - \frac{\delta}{2} \end{aligned}$$

where the first inequality follows from Lemma 18, and the second inequality follows from the definition of  $t_m$ . By the union bound,  $\Pr_{S \sim \mathcal{P}^m, \mathcal{A}}[E \cap E'] \geq 1 - \delta$ . In the event  $E \cap E'$ , we have

$$\begin{aligned} \text{err}(h_{\mathcal{A}}) - \text{err}(h^*) &\leq \text{err}(h_{\mathcal{A}}) - \text{err}(h_{\mathcal{A}}, S) + \text{err}(h^*, S) - \text{err}(h^*) + \text{err}(h_{\mathcal{A}}, S) - \text{err}(h^*, S) \\ &\leq \sqrt{\text{err}(h_{\mathcal{A}})\varepsilon_m} + \sqrt{\text{err}(h^*)\varepsilon_m} + \varepsilon_m + \text{err}(g_0, S) - \text{err}(h^*, S) + t_m \\ &\leq \sqrt{\text{err}(h_{\mathcal{A}})\varepsilon_m} + \sqrt{\text{err}(h^*)\varepsilon_m} + \varepsilon_m + \epsilon/4 + t_m \end{aligned}$$

since  $\text{err}(g_0, S) \leq \inf_{h' \in \mathcal{H}} \text{err}(h', S) + \epsilon/4 \leq \text{err}(h^*, S) + \epsilon/4$ . By various algebraic manipulations, this in turn implies

$$\text{err}(h_{\mathcal{A}}) \leq \text{err}(h^*) + C' \cdot \left( \sqrt{\text{err}(h^*)\varepsilon_m} + \varepsilon_m + t_m \right) + \epsilon/2$$

for some constant  $C' > 0$ . The lower bound on  $m$  now implies the theorem. ■

# Tight conditions for consistent variable selection in high dimensional nonparametric regression

**Laëtitia Comminges**

*Université Paris Est/ ENPC  
LIGM/IMAGINE*

LAETITIA.COMMINGES@IMAGINE.ENPC.FR

**Arnak S. Dalalyan**

*Université Paris Est/ ENPC  
LIGM/IMAGINE*

DALALYAN@IMAGINE.ENPC.FR

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We address the issue of variable selection in the regression model with very high ambient dimension, *i.e.*, when the number of covariates is very large. The main focus is on the situation where the number of relevant covariates, called intrinsic dimension, is much smaller than the ambient dimension. Without assuming any parametric form of the underlying regression function, we get tight conditions making it possible to consistently estimate the set of relevant variables. These conditions relate the intrinsic dimension to the ambient dimension and to the sample size. The procedure that is provably consistent under these tight conditions is simple and is based on comparing the empirical Fourier coefficients with an appropriately chosen threshold value.

**Keywords:** List of keywords

## 1. Introduction

Real-world data such as those obtained from neuroscience, chemometrics, data mining, or sensor-rich environments are often extremely high-dimensional, severely underconstrained (few data samples compared to the dimensionality of the data), and interspersed with a large number of irrelevant or redundant features. Furthermore, in most situations the data is contaminated by noise making it even more difficult to retrieve useful information from the data. Relevant variable selection is a compelling approach for addressing statistical issues in the scenario of high-dimensional and noisy data with small sample size. Starting from Mallows (1973), Akaike (1973); Schwarz (1978) who introduced respectively the famous criteria  $C_p$ , AIC and BIC, the problem of variable selection has been extensively studied in the statistical and machine learning literature both from the theoretical and algorithmic viewpoints. It appears, however, that the theoretical limits of performing variable selection in the context of nonparametric regression are still poorly understood, especially in the case where the ambient dimension of covariates, denoted by  $d$ , is much larger than the sample

size  $n$ . The purpose of the present work is to explore this setting under the assumption that the number of relevant covariates, hereafter called intrinsic dimension and denoted by  $d^*$ , may grow with the sample size but remains much smaller than the ambient dimension  $d$ .

In the important particular case of linear regression, the latter scenario has been the subject of a number of recent studies. Many of them rely on  $\ell_1$ -norm penalization (as for instance in Tibshirani (1996); Zhao and Yu (2006); Meinshausen and Bhlmann (2010)) and constitute an attractive alternative to iterative variable selection procedures proposed by Alquier (2008); Zhang (2009); Ting et al. (2010) and to marginal regression or correlation screening explored in Wasserman and Roeder (2009); Fan et al. (2009). Promising results for feature selection are also obtained by minimax concave penalties in Zhang (2010), by Bayesian approach in Scott and Berger (2010) and by higher criticism in Donoho and Jin (2009). Extensions to other settings including logistic regression, generalized linear model and Ising model have been carried out in Bunea and Barbu (2009); Ravikumar et al. (2010); Fan et al. (2009), respectively. Variable selection in the context of groups of variables with disjoint or overlapping groups has been studied by Jenatton et al. (2009); Lounici et al. (2010); Obozinski et al. (2011). Hierarchical procedures for selection of relevant covariates have been proposed by Bach (2009); Bickel et al. (2010) and Zhao et al. (2009).

It is now well understood that in the high-dimensional linear regression, if the Gram matrix satisfies some variant of irrepresentable condition, then consistent estimation of the pattern of relevant variables—also called the sparsity pattern—is possible under the condition  $d^* \log(d/d^*) = o(n)$  as  $n \rightarrow \infty$ . Furthermore, it is well known that if  $(d^* \log(d/d^*))/n$  remains bounded from below by some positive constant when  $n \rightarrow \infty$ , then it is impossible to consistently recover the sparsity pattern. Thus, a tight condition exists that describes in an exhaustive manner the interplay between the quantities  $d^*$ ,  $d$  and  $n$  that guarantees the existence of consistent estimators. The situation is very different in the case of non-linear regression, since, to our knowledge, there is no result providing tight conditions for consistent estimation of the sparsity pattern.

The papers Lafferty and Wasserman (2008) and Bertin and Lecué (2008), closely related to the present work, consider the problem of variable selection in nonparametric Gaussian regression model. They prove the consistency of the proposed procedures under some assumptions that—in the light of the present work—turn out to be suboptimal. More precisely, in Lafferty and Wasserman (2008), the unknown regression function is assumed to be four times continuously differentiable with bounded derivatives. The algorithm they propose, termed Rodeo, is a greedy procedure performing simultaneously local bandwidth choice and variable selection. Under the assumption that the density of the sampling design is continuously differentiable and strictly positive, Rodeo is shown to converge when the ambient dimension  $d$  is  $O(\log n/\log \log n)$  while the intrinsic dimension  $d^*$  does not increase with  $n$ . On the other hand, Bertin and Lecué (2008) propose a procedure based on the  $\ell_1$ -penalization of local polynomial estimators and prove its consistency when  $d^* = O(1)$  but  $d$  is allowed to be as large as  $\log n$ , up to a multiplicative constant. They also have a weaker assumption on the regression function which is merely assumed to belong to the Holder class with smoothness  $\beta > 1$ .

This brief review of the literature reveals that there is an important gap in consistency conditions for the linear regression and for the non-linear one. For instance, if the intrinsic dimension  $d^*$  is fixed, then the condition guaranteeing consistent estimation of the sparsity pattern is  $(\log d)/n \rightarrow 0$  in linear regression whereas it is  $d = O(\log n)$  in the nonparametric case. While it is undeniable that the nonparametric regression is much more complex than the linear one, it is however not easy to find a justification to such an important gap between two conditions. The situation is even worse in the case where  $d^* \rightarrow \infty$ . In fact, for the linear model with at most polynomially increasing ambient dimension  $d = O(n^k)$ , it is possible to estimate the sparsity pattern for intrinsic dimensions  $d^*$  as large as  $n^{1-\epsilon}$ , for some  $\epsilon > 0$ . In other words, the sparsity index can be almost on the same order as the sample size. In contrast, in nonparametric regression, there is no procedure that is proved to converge to the true sparsity pattern when both  $n$  and  $d^*$  tend to infinity, even if  $d^*$  grows extremely slowly.

In the present work, we fill this gap by introducing a simple variable selection procedure that selects the relevant variables by comparing some well chosen empirical Fourier coefficients to a prescribed significance level. Consistency of this procedure is established under some conditions on the triplet  $(d^*, d, n)$  and the tightness of these conditions is proved. The main take-away messages deduced from our results are the following:

- When the number of relevant covariates  $d^*$  is fixed and the sample size  $n$  tends to infinity, there exist positive real numbers  $c_*$  and  $c^*$  such that (a) if  $(\log d)/n \leq c_*$  the estimator proposed in Section 3 is consistent and (b) no estimator of the sparsity pattern may be consistent if  $(\log d)/n \geq c^*$ .
- When the number of relevant covariates  $d^*$  tends to infinity with  $n \rightarrow \infty$ , then there exist real numbers  $\underline{c}_i$  and  $\bar{c}_i$ ,  $i = 1, \dots, 4$  such that  $\underline{c}_i > 0$ ,  $\bar{c}_i > 0$  for  $i = 1, 2, 3$  and (a) if  $\underline{c}_1 d^* + \underline{c}_2 \log d^* + \underline{c}_3 \log \log d - \log n < \underline{c}_4$  the estimator proposed in Section 3 is consistent and (b) no estimator of the sparsity pattern may be consistent if  $\bar{c}_1 d^* + \bar{c}_2 \log d^* + \bar{c}_3 \log \log d - \log n > \bar{c}_4$ .
- In particular, if  $d$  grows not faster than a polynomial in  $n$ , then there exist positive real numbers  $c_0$  and  $c^0$  such that (a) if  $d^* \leq c_0 \log n$  the estimator proposed in Section 3 is consistent and (b) no estimator of the sparsity pattern may be consistent if  $d^* \geq c^0 \log n$ .

Very surprisingly, the derivation of these results required from us to apply some tools from complex analysis, such as the Jacobi  $\theta$ -function and the saddle point method, in order to evaluate the number of lattice points lying in a ball of an Euclidean space with increasing dimension.

The rest of the paper is organized as follows. The notation and assumptions necessary for stating our main results are presented in Section 2. In Section 3, an estimator of the set of relevant covariates is introduced and its consistency is established. The principal condition required in the consistency result involves the number of lattice points in a ball of a high-dimensional Euclidean space. An asymptotic equivalent for this number is obtained in Section 4 via the Jacobi  $\theta$ -function and the saddle point method. Results on impossibility of consistent estimation of the sparsity pattern are derived in Section 5, while the relation

between consistency and inconsistency results are discussed in Section 6. The technical parts of the proofs are postponed to the Appendix.

## 2. Notation and assumptions

We assume that  $n$  independent and identically distributed pairs of input-output variables  $(\mathbf{X}_i, Y_i)$ ,  $i = 1, \dots, n$  are observed that obey the regression model

$$Y_i = f(\mathbf{X}_i) + \sigma \varepsilon_i, \quad i = 1, \dots, n.$$

The input variables  $\mathbf{X}_1, \dots, \mathbf{X}_n$  are assumed to take values in  $\mathbb{R}^d$  while the output variables  $Y_1, \dots, Y_n$  are scalar. As usual, the noise  $\varepsilon_1, \dots, \varepsilon_n$  is such that  $\mathbf{E}[\varepsilon_i | \mathbf{X}_i] = 0$ ,  $i = 1, \dots, n$ ; some additional conditions will be imposed later. Without requiring from  $f$  to be of a special parametric form, we aim at recovering the set  $J \subset \{1, \dots, d\}$  of its relevant covariates.

It is clear that the estimation of  $J$  cannot be accomplished without imposing some further assumptions on  $f$  and the distribution  $P_X$  of the input variables. Roughly speaking, we will assume that  $f$  is differentiable with a squared integrable gradient and that  $P_X$  admits a density which is bounded from below. More precisely, let  $g$  denote the density of  $P_X$  w.r.t. the Lebesgue measure.

**[C1]** We assume that  $g(\mathbf{x}) = 0$  for any  $\mathbf{x} \notin [0, 1]^d$  and that  $g(\mathbf{x}) \geq g_{\min}$  for any  $\mathbf{x} \in [0, 1]^d$ .

To describe the smoothness assumption imposed on  $f$ , let us introduce the Fourier basis

$$\varphi_{\mathbf{k}}(\mathbf{x}) = \begin{cases} 1, & \mathbf{k} = 0, \\ \sqrt{2} \cos(2\pi \mathbf{k} \cdot \mathbf{x}), & \mathbf{k} \in (\mathbb{Z}^d)_+, \\ \sqrt{2} \sin(2\pi \mathbf{k} \cdot \mathbf{x}), & -\mathbf{k} \in (\mathbb{Z}^d)_+, \end{cases} \quad (1)$$

where  $(\mathbb{Z}^d)_+$  denotes the set of all  $\mathbf{k} \in \mathbb{Z}^d \setminus \{0\}$  such that the first nonzero element of  $\mathbf{k}$  is positive and  $\mathbf{k} \cdot \mathbf{x}$  stands for the the usual inner product in  $\mathbb{R}^d$ . In what follows, we use the notation  $\langle \cdot, \cdot \rangle$  for designing the scalar product in  $L^2([0, 1]^d; \mathbb{R})$ , that is  $\langle \mathbf{h}, \tilde{\mathbf{h}} \rangle = \int_{[0,1]^d} \mathbf{h}(\mathbf{x}) \tilde{\mathbf{h}}(\mathbf{x}) d\mathbf{x}$  for every  $\mathbf{h}, \tilde{\mathbf{h}} \in L^2([0, 1]^d; \mathbb{R})$ . Using this orthonormal Fourier basis, we define

$$\Sigma_L = \left\{ f : \sum_{\mathbf{k} \in \mathbb{Z}^d} k_j^2 \langle f, \varphi_{\mathbf{k}} \rangle^2 \leq L; \quad \forall j \in \{1, \dots, d\} \right\}.$$

To ease notation, we set  $\theta_{\mathbf{k}}[f] = \langle f, \varphi_{\mathbf{k}} \rangle$  for all  $\mathbf{k} \in \mathbb{Z}^d$ . In addition to the smoothness, we need also to require that the relevant covariates are sufficiently relevant for making their identification possible. This is done by means of the following condition.

**[C2( $\kappa, L$ )]** The regression function  $f$  belongs to  $\Sigma_L$ . Furthermore, for some subset  $J \subset \{1, \dots, d\}$  of cardinality  $\leq d^*$ , there exists a function  $\bar{f} : \mathbb{R}^{|J|} \rightarrow \mathbb{R}$  such that  $f(\mathbf{x}) = \bar{f}(\mathbf{x}_J)$ ,  $\forall \mathbf{x} \in \mathbb{R}^d$  and it holds that

$$Q_j[f] \triangleq \sum_{\mathbf{k}: k_j \neq 0} \theta_{\mathbf{k}}[f]^2 \geq \kappa, \quad \forall j \in J. \quad (2)$$

Hereafter, we will refer to  $J$  as the sparsity pattern of  $f$ .

One easily checks that  $Q_j[f] = 0$  for every  $j$  that does not lie in the sparsity pattern. This provides a characterization of the sparsity pattern as the set of indices of nonzero coefficients of the vector  $\mathbf{Q}[f] = (Q_1[f], \dots, Q_d[f])$ .

The next assumptions imposed to the regression function and to the noise require their boundedness in an appropriate sense. These assumptions are needed in order to prove, by means of a concentration inequality, the closeness of the empirical coefficients to the true ones.

**[C3]( $L_\infty, L_2$ )** The  $L^\infty([0, 1]^d, \mathbb{R}, P_X)$  and  $L^2([0, 1]^d, \mathbb{R}, P_X)$  norms of the function  $f$  are bounded from above respectively by  $L_\infty > 0$  and  $L_2$ , i.e.,  $P_X(\mathbf{x} \in [0, 1]^d : |f(\mathbf{x})| \leq L_\infty) = 1$  and  $\int_{[0,1]^d} f(\mathbf{x})^2 g(\mathbf{x}) d\mathbf{x} \leq L_2^2$ .

**[C4]** The noise variables satisfy a.e.  $\mathbf{E}[e^{t\varepsilon_i} | \mathbf{X}_i] \leq e^{t^2/2}$  for all  $t > 0$ .

**Remark 1** *The primary aim of this work is to understand when it is possible to estimate the sparsity pattern (with theoretical guarantees on the convergence of the estimator) and when it is impossible. The estimator that we will define in the next section is intended to show the possibility of consistent estimation, rather than being a practical procedure for recovering the sparsity pattern. Therefore, the estimator will be allowed to depend on the parameters  $g_{\min}$ ,  $L$ ,  $\kappa$  and  $M$  appearing in conditions [C1-C3].*

### 3. Consistent estimation of the set of relevant variables

The estimator of the sparsity pattern  $J$  that we are going to introduce now is based on the following simple observation: if  $j \notin J$  then  $\theta_{\mathbf{k}}[f] = 0$  for every  $\mathbf{k}$  such that  $k_j \neq 0$ . In contrast, if  $j \in J$  then there exists  $\mathbf{k} \in \mathbb{Z}^d$  with  $k_j \neq 0$  such that  $|\theta_{\mathbf{k}}[f]| > 0$ . To turn this observation into an estimator of  $J$ , we start by estimating the Fourier coefficients  $\theta_{\mathbf{k}}[f]$  by their empirical counterparts:

$$\widehat{\theta}_{\mathbf{k}} = \frac{1}{n} \sum_{i=1}^n \frac{\varphi_{\mathbf{k}}(\mathbf{X}_i)}{g(\mathbf{X}_i)} Y_i, \quad \mathbf{k} \in \mathbb{Z}^d.$$

Then, for every  $\ell \in \mathbb{N}$  and for any  $\gamma > 0$ , we introduce the notation  $S_{m,\ell} = \{\mathbf{k} \in \mathbb{Z}^d : \|\mathbf{k}\|_2 \leq m, \|\mathbf{k}\|_0 \leq \ell\}$  and  $N(d^*, \gamma) = \text{Card}\{\mathbf{k} \in \mathbb{Z}^{d^*} : \|\mathbf{k}\|_2^2 \leq \gamma d^* \& k_1 \neq 0\}$ . Finally our estimator is defined by

$$\widehat{J}_n(m, \lambda) = \left\{ j \in \{1, \dots, d\} : \max_{\mathbf{k} \in S_{m,d^*} : k_j \neq 0} |\widehat{\theta}_{\mathbf{k}}| > \lambda \right\}, \quad (3)$$

where  $m$  and  $\lambda$  are some parameters to be defined later. The notation  $a \wedge b$ , for two real numbers  $a$  and  $b$ , stands for  $\min(a, b)$ .

**Theorem 2** *Let conditions [C1-C4] be fulfilled with some known constants  $g_{\min}, L, \kappa$  and  $L_2$ . Assume furthermore that the design density  $g$  and an upper estimate on the noise*

magnitude  $\sigma$  are available. Set  $m = (2Ld^*/\kappa)^{1/2}$  and  $\lambda = 4(\sigma + L_2)(d^*\log(6md)/ng_{\min}^2)^{1/2}$ . If

$$\frac{L_\infty^2 d^* \log(6md)}{n} \leq L_2^2, \quad \text{and} \quad \frac{128(\sigma + L_2)^2 d^* N(d^*, 2L/\kappa) \log(6md)}{ng_{\min}^2} \leq \kappa, \quad (4)$$

then the estimator  $\widehat{J}(m, \lambda)$  satisfies  $\mathbf{P}(\widehat{J}(m, \lambda) \neq J) \leq 3(6md)^{-d^*}$ .

If we take a look at the conditions of Theorem 2 ensuring the consistency of the estimator  $\widehat{J}$ , it becomes clear that the strongest requirement is the second inequality in (4). To some extent, this condition requires that  $(d^*N(d^*, 2L/\kappa)\log d)/n$  is bounded from above by some constant. To further analyze the interplay between  $d^*$ ,  $d$  and  $n$  implied by this condition, we need an equivalent to  $N(d^*, 2L/\kappa)$  as the intrinsic dimension  $d^*$  tends to infinity. As proved in the next section,  $N(d^*, 2L/\kappa)$  diverges exponentially fast, making inequality (4) impossible for  $d^*$  larger than  $\log n$  up to a multiplicative constant.

It is also worth stressing that although we require the  $P_X$ -a.e. boundedness of  $f$  by some constant  $L_\infty$ , this constant is not needed for computing the estimator proposed in Theorem 2. Only constants related to some quadratic functionals of the sequence of Fourier coefficients  $\theta_k[f]$  are involved in the tuning parameters  $m$  and  $\lambda$ . This point might be important for designing practical estimators of  $J$ , since the estimation of quadratic functionals is more realistic, see for instance Laurent and Massart (2000), than the estimation of sup-norm.

The result stated above can be reformulated to provide also a level of relevance  $\kappa$  for the covariates of  $\mathbf{X}$  making their identification possible. In fact, an alternative way of stating Theorem 2 is the following: if conditions [C1-C4] and  $L_\infty^2 d^* \log(6md) \leq nL_2^2$  are fulfilled, then the estimator  $\widehat{J}(m, \lambda)$ —with arbitrary tuning parameters  $m$  and  $\lambda$ —satisfies  $\mathbf{P}(\widehat{J}(m, \lambda) \neq J) \leq 3(6md)^{-d^*}$  provided that the smallest level of relevance  $\kappa$  for components  $X_j$  of  $\mathbf{X}$  with  $j \in J$  is not smaller than  $8\lambda^2 N(d^*, m^2/d^*)$ . This statement can be easily deduced from the proof presented in Appendix A.

#### 4. Counting lattice points in a ball

The aim of the present section is to investigate the properties of the quantity  $N(d^*, m^2/d^*)$  that is involved in the conditions ensuring the consistency of the proposed procedure. Quite surprisingly, the asymptotic behavior of  $N(d^*, m^2/d^*)$  turns out to be related to the Jacobi  $\theta$ -function. In order to show this, let us introduce some notation. For a positive number  $\gamma$ , we set

$$\begin{aligned} \mathcal{C}_1(d^*, \gamma) &= \left\{ \mathbf{k} \in \mathbb{Z}^{d^*} : k_1^2 + \dots + k_{d^*}^2 \leq \gamma d^* \right\}, \\ \mathcal{C}_2(d^*, \gamma) &= \left\{ \mathbf{k} \in \mathbb{Z}^{d^*} : k_2^2 + \dots + k_{d^*}^2 \leq \gamma d^* \& k_1 = 0 \right\} \end{aligned}$$

along with  $N_1(d^*, \gamma) = \text{Card}\mathcal{C}_1(d^*, \gamma)$  and  $N_2(d^*, \gamma) = \text{Card}\mathcal{C}_2(d^*, \gamma)$ . In simple words,  $N_1(d^*, \gamma)$  is the number of (integer) lattice points lying in the  $d^*$ -dimensional ball with radius  $(\gamma d^*)^{1/2}$  and centered at the origin, while  $N_2(d^*, \gamma)$  is the number of (integer) lattice

points with the first coordinate equal to zero and lying in the  $d^*$ -dimensional ball with radius  $(\gamma d^*)^{1/2}$  and centered at the origin. With this notation, the quantity  $N(d^*, 2L/\kappa)$  of Theorem 2 can be written as  $N_1(d^*, 2L/\kappa) - N_2(d^*, 2L/\kappa)$ .

In order to determine the asymptotic behavior of  $N_1(d^*, \gamma)$  and  $N_2(d^*, \gamma)$  when  $d^*$  tends to infinity, we will rely on their integral representation through Jacobi's  $\theta$ -function. Recall that the latter is given by  $h(z) = \sum_{r \in \mathbb{Z}} z^{r^2}$ , which is well defined for any complex number  $z$  belonging to the unit ball  $|z| < 1$ . To briefly explain where the relation between  $N_i(\gamma)$  and the  $\theta$ -function comes from, let us denote by  $\{a_r\}$  the sequence of coefficients of the power series of  $h(z)^{d^*}$ , that is  $h(z)^{d^*} = \sum_{r \geq 0} a_r z^r$ . One easily checks that  $\forall r \in \mathbb{N}$ ,  $a_r = \text{Card}\{\mathbf{k} \in \mathbb{Z}^{d^*} : k_1^2 + \dots + k_{d^*}^2 = r\}$ . Thus, for every  $\gamma$  such that  $\gamma d^*$  is integer, we have  $N_1(d^*, \gamma) = \sum_{r=0}^{\gamma d^*} a_r$ . As a consequence of Cauchy's theorem, we get :

$$N_1(d^*, \gamma) = \frac{1}{2\pi i} \oint \frac{h(z)^{d^*}}{z^{\gamma d^*}} \frac{dz}{z(1-z)}.$$

where the integral is taken over any circle  $|z| = w$  with  $0 < w < 1$ . Exploiting this representation and applying the saddle-point method thoroughly described in Dieudonné (1968), we get the following result.

**Proposition 3** *Let  $\gamma > 0$  be such that  $\gamma d^*$  is an integer and let  $l_\gamma(z) = \log h(z) - \gamma \log z$ .*

1. *There is a unique solution  $z_\gamma$  in  $(0, 1)$  to the equation  $l'_\gamma(z) = 0$ . Furthermore, the function  $\gamma \mapsto z_\gamma$  is increasing and  $l''_\gamma(z) > 0$ .*
2. *The following equivalences hold true:*

$$\begin{aligned} N_1(d^*, \gamma) &= \left( \frac{h(z_\gamma)}{z_\gamma^\gamma} \right)^{d^*} \frac{1 + o(1)}{z_\gamma(1 - z_\gamma)(2l''_\gamma(z_\gamma)\pi d^*)^{1/2}}, \\ N_2(d^*, \gamma) &= \left( \frac{h(z_\gamma)}{z_\gamma^\gamma} \right)^{d^*} \frac{1 + o(1)}{h(z_\gamma)z_\gamma(1 - z_\gamma)(2l''_\gamma(z_\gamma)\pi d^*)^{1/2}}, \end{aligned}$$

as  $d^*$  tends to infinity.

Furthermore, the convergence to zero of the terms replaced by  $o(1)$  in the previous formulae is uniform in  $\gamma$  on any compact set  $[\underline{\gamma}, \bar{\gamma}] \subset (0, \infty)$ .

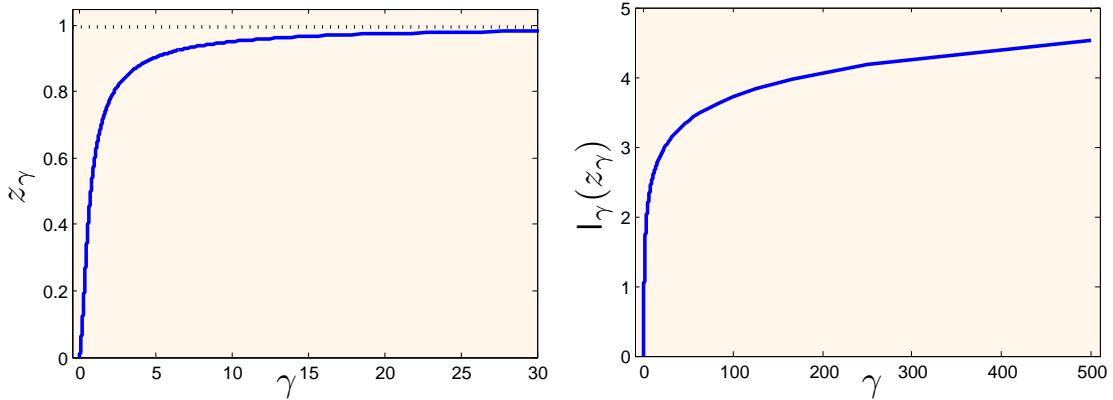
In the sequel, it will be useful to remark that the second part of Proposition 3 yields

$$\log N(d^*, \gamma) = d^* l_\gamma(z_\gamma) - \frac{1}{2} \log d^* - \log \left\{ \frac{h(z_\gamma)z_\gamma(1 - z_\gamma)(2l''_\gamma(z_\gamma)\pi)^{1/2}}{h(z_\gamma) - 1} \right\} + o(1). \quad (5)$$

In order to get an idea of how the terms  $z_\gamma$  and  $l_\gamma(z_\gamma)$  depend on  $\gamma$ , we depicted in Figure 1 the plots of these quantities as functions of  $\gamma > 0$ .

## 5. Tightness of the assumptions

In this section, we assume that the errors  $\varepsilon_i$  are i.i.d. Gaussian with zero mean and variance 1 and we focus our attention on the functional class  $\widehat{\Sigma}(\kappa, L)$  of all functions satisfying

Figure 1: The plots of mappings  $\gamma \mapsto z_\gamma$  and  $\gamma \mapsto l_\gamma(z_\gamma)$ .

assumption [C2( $\kappa, L$ )]. In order to avoid irrelevant technicalities and to better convey the main results, we assume that  $\kappa = 1$  and denote  $\tilde{\Sigma}_L = \tilde{\Sigma}(1, L)$ . Furthermore, we will assume that the design  $\mathbf{X}_1, \dots, \mathbf{X}_n$  is fixed and satisfies

$$\frac{1}{n} \sum_{i=1}^n \varphi_{\mathbf{k}}(\mathbf{X}_i) \varphi_{\mathbf{k}'}(\mathbf{X}_i) \leq \frac{n}{N_1(d^*, L)^2} \quad (6)$$

for all distinct  $\mathbf{k}, \mathbf{k}' \in S_{(d^*L)^{1/2}, d^*} \subset \mathbb{Z}^d$ . The goal in this section is to provide conditions under which the consistent estimation of the sparsity support is impossible, that is there exists a positive constant  $c > 0$  and an integer  $n_0 \in \mathbb{N}$  such that, if  $n \geq n_0$ ,

$$\inf_{\tilde{J}} \sup_{\mathbf{f} \in \tilde{\Sigma}_L} \mathbf{P}_{\mathbf{f}}(\tilde{J} \neq J_{\mathbf{f}}) \geq c,$$

where the inf is over all possible estimators of  $J_{\mathbf{f}}$ . To lower bound the LHS of the last inequality, we introduce a set of  $M + 1$  probability distributions  $\mu_0, \dots, \mu_M$  on  $\tilde{\Sigma}_L$  and use the fact that

$$\inf_{\tilde{J}} \sup_{\mathbf{f} \in \tilde{\Sigma}_L} \mathbf{P}_{\mathbf{f}}(\tilde{J} \neq J_{\mathbf{f}}) \geq \inf_{\tilde{J}} \frac{1}{M+1} \sum_{\ell=0}^M \int_{\tilde{\Sigma}_L} \mathbf{P}_{\mathbf{f}}(\tilde{J} \neq J_{\mathbf{f}}) \mu_{\ell}(d\mathbf{f}). \quad (7)$$

These measures  $\mu_{\ell}$  will be chosen in such a way that for each  $\ell \geq 1$  there is a set  $J_{\ell}$  of cardinality  $d^*$  such that  $\mu_{\ell}\{J_{\mathbf{f}} = J_{\ell}\} = 1$  and all the sets  $J_1, \dots, J_M$  are distinct. The measure  $\mu_0$  is the Dirac measure in 0. Considering these  $\mu_{\ell}$ s as ‘‘prior’’ probability measures on  $\tilde{\Sigma}_L$  and defining the corresponding ‘‘posterior’’ probability measures  $\mathbb{P}_0, \mathbb{P}_1, \dots, \mathbb{P}_M$  by

$$\mathbb{P}_{\ell}(A) = \int_{\tilde{\Sigma}_L} \mathbf{P}_{\mathbf{f}}(A) \mu_{\ell}(d\mathbf{f}), \quad \text{for every measurable set } A \subset \mathbb{R}^n,$$

we can write the inequality (7) as

$$\inf_{\tilde{J}} \sup_{\mathbf{f} \in \tilde{\Sigma}_L} \mathbf{P}_{\mathbf{f}}(\tilde{J} \neq J_{\mathbf{f}}) \geq \inf_{\psi} \frac{1}{M+1} \sum_{\ell=0}^M \mathbb{P}_{\ell}(\psi \neq \ell), \quad (8)$$

where the inf is taken over all random variables  $\psi$  taking values in  $\{0, \dots, M\}$ . The latter inf will be controlled using a suitable version of the Fano lemma, see Fano (1961). In what follows, we denote by  $\mathcal{K}(P, Q)$  the Kullback-Leibler divergence between two probability measures  $P$  and  $Q$  defined on the same probability space.

**Lemma 4 (Cor. 2.6 of Tsybakov (2009))** *Let  $(\mathcal{X}, \mathcal{A})$  be a measurable space and let  $P_0, \dots, P_M$  be probability measures on  $(\mathcal{X}, \mathcal{A})$ . Let us set  $\bar{p}_{e,M} = \inf_{\psi} (M+1)^{-1} \sum_{\ell=0}^M P_\ell(\psi \neq \ell)$ , where the inf is taken over all measurable functions  $\psi : \mathcal{X} \rightarrow \{0, \dots, M\}$ . If for some  $0 < \alpha < 1$*

$$\frac{1}{M+1} \sum_{\ell=0}^M \mathcal{K}(P_\ell, P_0) \leq \alpha \log M,$$

then

$$\bar{p}_{e,M} \geq \frac{\log(M+1) - \log 2}{\log M} - \alpha.$$

It follows from this lemma that one can deduce a lower bound on  $\bar{p}_{e,M}$ , which is the quantity we are interested in, from an upper bound on the average Kullback-Leibler divergence between the measures  $\mathbb{P}_\ell$  and  $\mathbb{P}_0$ . This roughly means that the measures  $\mu_\ell$  should not be very far from  $\mu_0$  but the probability measures  $\mu_\ell$  should be very different one from another in terms of the sparsity pattern of a function  $f$  randomly drawn according to  $\mu_\ell$ . This property is ensured by the following result.

**Lemma 5** *Suppose  $\mu_0 = \delta_0$ , the Dirac measure at  $0 \in \Sigma_L$ . Let  $S$  be a subset of  $\mathbb{Z}^d$  of cardinality  $|S|$  and  $A$  be a constant. Define  $\mu_S$  as a discrete measure supported on the finite set of functions  $\{f_\omega = \sum_{k \in S} A \omega_k \varphi_k : \omega \in \{\pm 1\}^S\}$  such that  $\mu_S(f = f_\omega) = 2^{-|S|}$  for every  $\omega \in \{\pm 1\}^S$ , i.e., the  $\omega_k$ 's are i.i.d. Rademacher random variables under  $\mu_S$ . If, for some  $\epsilon \geq 0$ , the condition*

$$\frac{1}{n} \sum_{i=1}^n \varphi_{\mathbf{k}}(\mathbf{X}_i) \varphi_{\mathbf{k}'}(\mathbf{X}_i) \leq \epsilon \quad \forall \mathbf{k}, \mathbf{k}' \in S$$

is fulfilled, then

$$\mathcal{K}(\mathbb{P}_1, \mathbb{P}_0) \leq \log \left[ \int \left( \frac{d\mathbb{P}_1}{d\mathbb{P}_0}(\mathbf{y}) \right)^2 \mathbb{P}_0(d\mathbf{y}) \right] \leq 4|S|A^4 n^2 \left\{ 1 + \frac{|S|\epsilon}{4nA^2} \right\}.$$

These evaluations lead to the following theorem, that tells us that the conditions to which we have resorted for proving the consistency in Section 3 are nearly optimal.

**Theorem 6** *Let the design  $\mathbf{X}_1, \dots, \mathbf{X}_n \in [0, 1]^d$  be deterministic and satisfy (6). Let  $\gamma^*$  the largest real number such that  $d^* \gamma^*$  is integer and  $L \geq \gamma^*(1+1/2z_{\gamma^*})$ . If for some positive number  $\alpha < (\log 3 - \log 2)/\log 3$*

$$\frac{(N_1(d^*, \gamma^*) - N_2(d^*, \gamma^*))^2 \log \binom{d}{d^*}}{n^2 N_1(d^*, \gamma^*)} \geq \frac{\alpha}{5}, \tag{9}$$

then there exists a positive constant  $c > 0$  and a  $d_0 \in \mathbb{N}$  such that, if  $d^* \geq d_0$ ,

$$\inf_{\tilde{J}} \sup_{f \in \Sigma_L} \mathbf{P}_f(\tilde{J} \neq J_f) \geq c.$$

**Proof** We apply the Fano lemma with  $M = \binom{d}{d^*}$ . We choose  $\mu_0, \dots, \mu_M$  as follows.  $\mu_0$  is the Dirac measure  $\delta_0$ ,  $\mu_1$  is defined as in Lemma 5 with  $S = \mathcal{C}_1(d^*, \gamma^*)$  and  $A = [N_1(d^*, \gamma^*) - N_2(d^*, \gamma^*)]^{-1/2}$ . The measures  $\mu_2, \dots, \mu_M$  are defined similarly and correspond to the  $M - 1$  remaining sparsity patterns of cardinality  $d^*$ .

In view of inequality (8) and Lemma 4, it suffices to show that the measures  $\mu_\ell$  satisfy  $\mu_\ell(\tilde{\Sigma}_L) = 1$  and  $\sum_{\ell=0}^M \mathcal{K}(\mathbb{P}_\ell, \mathbb{P}_0) \leq (M+1)\alpha \log M$ . Combining Lemma 5 with  $\text{Card}(S) = N_1(d^*, \gamma^*)$  and condition (6), one easily checks that equation (9) implies the desired bound on  $\sum_{\ell=0}^M \mathcal{K}(\mathbb{P}_\ell, \mathbb{P}_0)$ .

Let us show now that  $\mu_1(\tilde{\Sigma}_L) = 1$ . By symmetry, this will imply that  $\mu_\ell(\tilde{\Sigma}_L) = 1$  for every  $\ell$ . Since  $\mu_1$  is supported by the set  $\{f_\omega : \omega \in \{\pm 1\}^{\mathcal{C}_1(d^*, \gamma^*)}\}$ , it is clear that

$$\sum_{k_1 \neq 0} \theta_k^2[f_\omega] = A^2[N_1(d^*, \gamma^*) - N_2(d^*, \gamma^*)] = 1$$

and, for every  $j = 1, \dots, d^*$ ,

$$\sum_{k \in \mathbb{Z}^d} k_j^2 \theta_k^2[f_\omega] = \sum_{k \in \mathcal{C}_1(d^*, \gamma^*)} k_j^2 A^2 = \frac{1}{d^*} \sum_{j=1}^{d^*} \sum_{k \in \mathcal{C}_1(d^*, \gamma^*)} k_j^2 A^2 \leq A^2 \gamma^* N_1(d^*, \gamma^*).$$

By virtue of Proposition 3, as  $d^*$  tends to infinity,  $N_1(d^*, \gamma^*)/N_2(d^*, \gamma^*)$  is asymptotically equivalent to  $h(z_{\gamma^*}) > 1 + 2z_{\gamma^*}$ . Hence, for  $d^*$  large enough,

$$A^2 N_1(d^*, \gamma^*) = \frac{N_1(d^*, \gamma^*)}{N_1(d^*, \gamma^*) - N_2(d^*, \gamma^*)} < \frac{1}{2z_{\gamma^*}} + 1.$$

As a consequence, for every  $j = 1, \dots, d^*$ ,

$$\sum_{k \in \mathbb{Z}^d} k_j^2 \theta_k^2[f_\omega] \leq \gamma^* \left( \frac{1}{2z_{\gamma^*}} + 1 \right) \leq L,$$

where the last inequality follows from the definition of  $\gamma^*$ . ■

Note that Theorem 6 is concerned by the case where the intrinsic dimension is not too small, which is the most interesting case in the present context. However, a much simpler result can be established showing that the conditions of Theorem 2 are tight in the case of fixed intrinsic dimension as well.

**Proposition 7** *Let the design  $\mathbf{X}_1, \dots, \mathbf{X}_n \in [0, 1]^d$  be either deterministic or random. If for some positive  $\alpha < (\log 3 - \log 2)/\log 3$ , the inequality*

$$\frac{d^*(\log d - \log d^*)}{n} \geq \alpha^{-1}$$

*holds true, then there is a constant  $c > 0$  such that  $\inf_{\tilde{J}_n} \sup_{f \in \tilde{\Sigma}_L} \mathbf{P}_f(\tilde{J}_n \neq J_f) \geq c$ .*

## 6. Discussion

The results proved in previous sections almost exhaustively answer the questions on the existence of consistent estimators of the sparsity pattern in the problem of nonparametric regression. In fact as far as only rates of convergence are of interest, the result obtained in Theorem 2 is shown in Section 5 to be unimprovable. Thus only the problem of finding sharp constants remains open. To make these statements more precise, let us consider the simplified set-up  $\sigma = \kappa = 1$  and define the following two regimes:

- The regime of fixed sparsity, *i.e.*, when the sample size  $n$  and the ambient dimension  $d$  tend to infinity but the intrinsic dimension  $d^*$  remains constant or bounded.
- The regime of increasing sparsity, *i.e.*, when the intrinsic dimension  $d^*$  tends to infinity along with the sample size  $n$  and the ambient dimension  $d$ . For simplicity, we will assume that  $d^* = O(d^{1-\epsilon})$  for some  $\epsilon > 0$ .

In the fixed sparsity regime, in view of Theorem 2, consistent estimation of the sparsity pattern can be achieved using the estimator  $\widehat{J}$  as soon as  $(\log d)/n \leq c_*$ , where  $c_*$  is the constant defined by

$$c_* = \min \left( \frac{L_2^2}{2d^* L_\infty^2}, \frac{g_{\min}^2}{2^8(1+L_2)^2 d^* N(d^*, 2L)} \right).$$

This follows from the fact that the tuning parameter  $m$  is fixed and that the probability of the error, bounded by  $3(6md)^{d^*}$  tends to zero as  $d \rightarrow \infty$ . On the other hand, by virtue of Proposition 7, consistent estimation of the sparsity pattern is impossible if  $(\log d)/n > c^*$ , where  $c^* = 2 \log 3 / (d^* \log(3/2))$ . Thus, up to multiplicative constants  $c_*$  and  $c^*$  (which are clearly not sharp), the result of Theorem 2 cannot be improved.

In the regime of increasing sparsity, the second inequality in (4) is the most stringent one. Taking the logarithm of both sides and using formula (5) for  $N(d^*, 2L) = N_1(d^*, 2L) - N_2(d^*, 2L)$ , we see that consistent estimation of  $J$  is possible when

$$\underline{c}_1 d^* + \frac{1}{2} \log d^* + \log \log d - \log n < \underline{c}_2, \quad (10)$$

with  $\underline{c}_1 = l_{2L}(z_{2L})$  and  $\underline{c}_2 = 2(\log(g_{\min}) - \log(17(\sigma+L_2))) + \log \left\{ \frac{h(z_{2L}) z_{2L} (1-z_{2L}) (2l''_{2L}(z_{2L})\pi)^{1/2}}{(h(z_{2L})-1)} \right\}$ . On the other hand, by virtue of (5),  $\log \left\{ \frac{[N_1(d^*, \gamma) - N_2(d^*, \gamma)]^2}{N_1(d^*, \gamma)} \right\} = d^* l_\gamma(z_\gamma) - \frac{1}{2} \log d^* - \log \left\{ \frac{h(z_\gamma)^2 z_\gamma (1-z_\gamma) (2l''_\gamma(z_\gamma)\pi)^{1/2}}{(h(z_\gamma)-1)^2} \right\} + o(1)$ . Therefore, Theorem 6 yields that it is impossible to consistently estimate  $J$  if

$$\bar{c}_1 d^* + \frac{1}{2} \log d^* + \log \log d - 2 \log n > \bar{c}_2, \quad (11)$$

where  $\bar{c}_1 = l_{\gamma^*}(z_{\gamma^*})$  and  $\bar{c}_2 = \log \left\{ \frac{h(z_{\gamma^*})^2 z_{\gamma^*} (1-z_{\gamma^*}) (2l''_{\gamma^*}(z_{\gamma^*})\pi)^{1/2}}{(h(z_{\gamma^*})-1)^2} \right\} + \log \log(3/2) - \log 5 - \log \log 3$ . A very simple consequence of inequalities (10) and (11) is that the consistent recovery of the sparsity pattern is possible under the condition  $d^*/\log n \rightarrow 0$  and impossible for  $d^*/\log n \rightarrow \infty$  as  $n \rightarrow \infty$ , provided that  $\log \log d = o(\log n)$ .

Let us stress now that, all over this work, we have deliberately avoided any discussion on the computational aspects of the variable selection in nonparametric regression. The goal in this paper was to investigate the possibility of consistent recovery without paying attention to the complexity of the selection procedure. This lead to some conditions that could be considered a benchmark for assessing the properties of sparsity pattern estimators. As for the estimator proposed in Section 3, it is worth noting that its computational complexity is not always prohibitively large. A recommended strategy is to compute the coefficients  $\hat{\theta}_{\mathbf{k}}$  in a stepwise manner; at each step  $K = 1, 2, \dots, d^*$  only the coefficients  $\hat{\theta}_{\mathbf{k}}$  with  $\|\mathbf{k}\|_0 = K$  need to be computed and compared with the threshold. If some  $\hat{\theta}_{\mathbf{k}}$  exceeds the threshold, then all the covariates  $X^j$  corresponding to nonzero coordinates of  $\mathbf{k}$  are considered as relevant. We can stop this computation as soon as the number of covariates classified as relevant attains  $d^*$ . While the worst-case complexity of this procedure is exponential, there are many functions  $f$  for which the complexity of the procedure will be polynomial in  $d$ . For example, this is the case for additive models in which  $f(\mathbf{x}) = f_1(x_{i_1}) + \dots + f_{d^*}(x_{i_{d^*}})$  for some univariate functions  $f_1, \dots, f_{d^*}$ .

Note also that in the present study we focused exclusively on the consistency of variable selection without paying any attention to the consistency of regression function estimation. A thorough analysis of the latter problem being left to a future work, let us simply remark that in the case of fixed  $d^*$ , under the conditions of Theorem 2, it is straightforward to construct a consistent estimator of the regression function. In fact, it suffices to use a projection estimator with a properly chosen truncation parameter on the set of relevant variables. The situation is much more delicate in the case when the sparsity  $d^*$  grows to infinity along with the sample size  $n$ . Presumably, condition (10) is no longer sufficient for consistently estimating the regression function. The rationale behind this conjecture is that the minimax rate of convergence for estimating  $f$  in our context, if we assume in addition that the set of relevant variables is known, is equal  $n^{-2/(2+d^*)} = \exp(-2 \log n / (2 + d^*))$ . If the left hand side of (10) is equal to a constant and  $\log \log d = o(\log n)$ , then the aforementioned minimax rate does not tend to zero, making thus the estimator inconsistent. This heuristical argument shows that there is still some work to do for getting tight conditions ensuring the consistent estimation of the regression function in the high dimensional set-up.

## Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under the grant PARCIMONIE.

## References

- Hirotsugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.
- Pierre Alquier. Iterative feature selection in least square regression estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 44(1):47–88, 2008.

- Francis Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. Technical report, arXiv:0909.0844, 2009.
- Karine Bertin and Guillaume Lecué. Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electron. J. Stat.*, 2:1224–1241, 2008.
- Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Hierarchical selection of variables in sparse high-dimensional regression. *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown. IMS Collections*, 6:56–69, 2010.
- Florentina Bunea and Adrian Barbu. Dimension reduction and variable selection in case control studies via regularized likelihood optimization. *Electron. J. Stat.*, 3:1257–1287, 2009.
- Jean Dieudonné. *Calcul infinitésimal*. Hermann, Paris, 1968.
- David Donoho and Jiashun Jin. Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 367(1906):4449–4470, 2009. With electronic supplementary materials available online.
- Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.*, 10:2013–2038, 2009.
- Robert M. Fano. *Transmission of information: A statistical theory of communications*. The M.I.T. Press, Cambridge, Mass., 1961.
- Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. Technical report, arXiv:0904.3523, 2009.
- John Lafferty and Larry Wasserman. Rodeo: sparse, greedy nonparametric regression. *Ann. Statist.*, 36(1):28–63, 2008.
- Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.
- Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sara van de Geer. Oracle inequalities and optimal inference under group sparsity. Technical report, arXiv:1007.1771, 2010.
- Colin L. Mallows. Some comments on  $C_p$ . *Technometrics*, 15:661–675, Nov. 1973.
- James Mazo and Andrew Odlyzko. Lattice points in high-dimensional spheres. *Monatsh. Math.*, 110(1):47–61, 1990.
- Nicolai Meinshausen and Peter Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(4):417–473, 2010.
- Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. High-dimensional union support recovery in multivariate. *The Annals of Statistics*, to appear, 2011.

Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using  $\ell_1$ -regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 2010.

Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

James G. Scott and James O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.*, 38(5):2587–2619, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

Jo-Anne Ting, Aaron D’Souza, Sethu Vijayakumar, and Stefan Schaal. Efficient learning and feature selection in high-dimensional regression. *Neural Comput.*, 22(4):831–886, 2010.

Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

Larry Wasserman and Kathryn Roeder. High-dimensional variable selection. *Ann. Statist.*, 37(5A):2178–2201, 2009.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010.

Tong Zhang. On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.*, 10:555–568, 2009.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7: 2541–2563, 2006.

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.*, 37(6A):3468–3497, 2009.

## Appendix A. Proof of Theorem 2

The empirical Fourier coefficients can be decomposed as follows:

$$\hat{\theta}_{\mathbf{k}} = \tilde{\theta}_{\mathbf{k}} + z_{\mathbf{k}}, \quad \text{where} \quad \tilde{\theta}_{\mathbf{k}} = \frac{1}{n} \sum_{i=1}^n \frac{\varphi_{\mathbf{k}}(\mathbf{X}_i)}{g(\mathbf{X}_i)} f(\mathbf{X}_i) \quad \text{and} \quad z_{\mathbf{k}} = \frac{\sigma}{n} \sum_{i=1}^n \frac{\varphi_{\mathbf{k}}(\mathbf{X}_i)}{g(\mathbf{X}_i)} \varepsilon_i. \quad (12)$$

If, for a multi index  $\mathbf{k}$ ,  $\theta_{\mathbf{k}} = 0$ , then the corresponding empirical Fourier coefficient will be close to zero with high probability. To show this, let us first look at what happens with  $z_{\mathbf{k}}$ ’s. We have, for every real number  $x$ ,

$$\mathbf{P}(|z_{\mathbf{k}}| > x \mid \mathbf{X}_1, \dots, \mathbf{X}_n) \leq \exp\left(-\frac{x^2}{2\sigma_{\mathbf{k}}^2}\right) \quad \forall \mathbf{k} \in S_{m,d^*}$$

with

$$\sigma_{\mathbf{k}}^2 = \frac{\sigma^2}{n^2} \sum_{i=1}^n \frac{\varphi_{\mathbf{k}}(\mathbf{X}_i)^2}{g(\mathbf{X}_i)^2} \leq \frac{2\sigma^2}{g_{\min}^2 n}.$$

Therefore, for every  $\mathbf{k} \in S_{m,d^*}$ , it holds that

$$\mathbf{P}(|z_{\mathbf{k}}| > x | \mathbf{X}_1, \dots, \mathbf{X}_n) \leq \exp(-ng_{\min}^2 x^2 / 4\sigma^2).$$

This entails that by setting  $\lambda_1 = (8\sigma^2 d^* \log(6md)/ng_{\min}^2)^{1/2}$  and by using the inequalities

$$\begin{aligned} \text{Card}(S_{m,d^*}) &= \sum_{i=0}^{d^*} \binom{d}{i} (2m)^i \leq (2m)^{d^*} \sum_{i=0}^{d^*} \frac{d^i}{i!} \\ &\leq 3(2md)^{d^*} \leq (6md)^{d^*}, \end{aligned}$$

we get

$$\begin{aligned} \mathbf{P}\left(\max_{\mathbf{k} \in S_{m,d^*}} |z_{\mathbf{k}}| > \lambda_1 | \mathbf{X}_1, \dots, \mathbf{X}_n\right) &\leq \sum_{\mathbf{k} \in S_{m,d^*}} \mathbf{P}(|z_{\mathbf{k}}| > \lambda_1 | \mathbf{X}_1, \dots, \mathbf{X}_n) \\ &\leq \text{Card}(S_{m,d^*}) e^{-ng_{\min}^2 \lambda_1^2 / 4\sigma^2} \leq (6md)^{-d^*}. \end{aligned}$$

Next, we use a concentration inequality for controlling large deviations of  $\tilde{\theta}_{\mathbf{k}}$ 's from  $\theta_{\mathbf{k}}$ 's. Recall that in view of the definition  $\tilde{\theta}_{\mathbf{k}} = \frac{1}{n} \sum_{i=1}^n \frac{\varphi_{\mathbf{k}}(\mathbf{X}_i)}{g(\mathbf{X}_i)} \mathbf{f}(\mathbf{X}_i)$ , we have  $\mathbb{E}(\tilde{\theta}_{\mathbf{k}}) = \theta_{\mathbf{k}}$ . By virtue of the boundedness of  $\mathbf{f}$ , it holds that  $|\frac{\varphi_{\mathbf{k}}(\mathbf{X}_i)}{g(\mathbf{X}_i)} \mathbf{f}(\mathbf{X}_i)| \leq \sqrt{2}L_{\infty}/g_{\min}$ . Furthermore, the bound  $V \triangleq \text{Var}(\frac{\varphi_{\mathbf{k}}(\mathbf{X}_i)}{g(\mathbf{X}_i)} \mathbf{f}(\mathbf{X}_i)) \leq \int f^2(\mathbf{x}) \frac{\varphi_{\mathbf{k}}^2(\mathbf{x})}{g^2(\mathbf{x})} d\mathbf{x} \leq 2L_2^2/g_{\min}^2$  combined with Bernstein's inequality yields

$$\begin{aligned} \mathbf{P}(|\tilde{\theta}_{\mathbf{k}} - \theta_{\mathbf{k}}| > t) &\leq 2 \exp\left(-\frac{nt^2}{2(V + t\sqrt{2}L_{\infty}/3g_{\min})}\right) \\ &\leq 2 \exp\left(-\frac{g_{\min}^2 nt^2}{4L_2^2 + tL_{\infty}g_{\min}}\right), \quad \forall t > 0. \end{aligned}$$

Let us define  $\lambda_2 = 4L_2 \left(\frac{d^* \log(6md)}{ng_{\min}^2}\right)^{1/2}$ . Then,

$$\mathbf{P}(|\tilde{\theta}_{\mathbf{k}} - \theta_{\mathbf{k}}| > \lambda_2) \leq 2 \exp\left(-\frac{4L_2^2 d^* \log(6md)}{L_2^2 + L_{\infty} L_2 \left(\frac{d^* \log(6md)}{n}\right)^{1/2}}\right).$$

The first inequality in condition (4) implies that the denominator in the exponential is not larger than  $2L_2^2$ . Hence,

$$\mathbf{P}\left(\max_{\mathbf{k} \in S_{m,d^*}} |\tilde{\theta}_{\mathbf{k}} - \theta_{\mathbf{k}}| > \lambda_2\right) \leq 2/(6md)^{d^*}.$$

Let  $\mathcal{A}_1 = \{\max_{\mathbf{k} \in S_{m,d^*}} |z_{\mathbf{k}}| \leq \lambda_1\}$  and  $\mathcal{A}_2 = \{\max_{\mathbf{k} \in S_{m,d^*}} |\tilde{\theta}_{\mathbf{k}}| \leq \lambda_2\}$ . One easily checks that

$$\mathbf{P}(J^c \not\subset \widehat{J}^c) \leq \mathbf{P}(\mathcal{A}_1^c) + \mathbf{P}(\mathcal{A}_2^c) \leq 3/(6md)^{d^*}.$$

As for the converse inclusion, we have

$$\begin{aligned}\mathbf{P}(J \not\subset \widehat{J}) &\leq \mathbf{P}\left(\exists j \in J \text{ s.t. } \max_{\mathbf{k} \in S_{m,d^*}: k_j \neq 0} |\widehat{\theta}_{\mathbf{k}}| \leq \lambda\right) \\ &\leq \mathbf{1}\left\{\exists j \in J \text{ s.t. } \max_{\mathbf{k} \in S_{m,d^*}: k_j \neq 0} |\theta_{\mathbf{k}}| \leq 2\lambda\right\} + \mathbf{P}(\mathcal{A}_1^c) + \mathbf{P}(\mathcal{A}_2^c).\end{aligned}$$

We show now that the first term in the last line is equal to zero. If this was not the case, then for some value  $j_0$  we would have  $Q_{j_0} \geq \kappa$  and  $|\theta_{\mathbf{k}}| \leq 2\lambda$ , for all  $\mathbf{k} \in S_{m,d^*}$  such that  $k_{j_0} \neq 0$ . This would imply that

$$Q_{j_0,m,d^*} \triangleq \sum_{\mathbf{k} \in S_{m,d^*}: k_{j_0} \neq 0} \theta_{\mathbf{k}}^2 \leq 4\lambda^2 N(d^*, m^2/d^*).$$

On the other hand,

$$Q_{j_0} - Q_{j_0,m,d^*} \leq \sum_{\|\mathbf{k}\|_2 \geq m} \theta_{\mathbf{k}}^2 \leq m^{-2} \sum_{\|\mathbf{k}\|_2 \geq m} \sum_{j \in J} k_j^2 \theta_{\mathbf{k}}^2 \leq \frac{Ld^*}{m^2}.$$

Remark now that the choice of the truncation parameter  $m$  proposed in the statement of the proposition implies that  $Q_{j_0} - Q_{j_0,m,d^*} \leq \kappa/2$ . Combining these estimates, we get  $Q_{j_0} \leq \frac{\kappa}{2} + 4\lambda^2 N(d^*, m^2/d^*)$ , which is impossible since  $Q_{j_0} \geq \kappa$ .

## Appendix B. Proof of Proposition 3

**Proof of the first assertion.** This proof can be found in Mazo and Odlyzko (1990), we repeat here the arguments therein for the sake of keeping the paper self-contained. Recall that  $N_1(d^*, \gamma)$  admits an integral representation with the integrand:

$$\frac{\mathbf{h}(z)^{d^*}}{z^{\gamma d^*}} \frac{1}{z(1-z)} = \frac{1}{z(1-z)} \exp\left[d^* \log\left(\frac{\mathbf{h}(z)}{z^\gamma}\right)\right].$$

For any real number  $y > 0$ , we define

$$\phi(y) = e^{-y} \mathbf{h}'(e^{-y})/\mathbf{h}(e^{-y}) = \sum_{k=-\infty}^{k=+\infty} k^2 e^{-yk^2} / \sum_{k=-\infty}^{k=+\infty} e^{-yk^2}$$

in such a way that

$$\phi(y) = \gamma \iff \frac{\mathbf{h}'(e^{-y})}{\mathbf{h}(e^{-y})} = \frac{\gamma}{e^{-y}} \iff \mathbf{l}'_\gamma(e^{-y}) = 0.$$

By virtue of the Cauchy-Schwarz inequality, it holds that

$$\sum k^4 e^{-yk^2} \sum e^{-yk^2} > \left(\sum k^2 e^{-yk^2}\right)^2, \quad \forall y \in (0, \infty),$$

implying that  $\phi'(y) < 0$  for all  $y \in (0, \infty)$ , i.e.,  $\phi$  is strictly decreasing. Furthermore,  $\phi$  is obviously continuous with  $\lim_{y \rightarrow 0} \phi(y) = +\infty$  and  $\lim_{y \rightarrow \infty} \phi(y) = 0$ . These properties

imply the existence and the uniqueness of  $y_\gamma \in (0, \infty)$  such that  $\phi(y_\gamma) = \gamma$ . Furthermore, as the inverse of a decreasing function, the function  $\gamma \mapsto y_\gamma$  is decreasing as well. We set  $z_\gamma = e^{-y_\gamma}$  so that  $\gamma \mapsto z_\gamma$  is increasing.

We also have

$$\begin{aligned} l''_\gamma(z_\gamma) &= \frac{h''h - (h')^2}{h^2}(z_\gamma) + \frac{\gamma}{z_\gamma^2} = z_\gamma^{-2} \left\{ \frac{\sum_k (k^4 - k^2)z_\gamma^{k^2}}{\sum_k z_\gamma^{k^2}} - \left( \frac{\sum_k k^2 z_\gamma^{k^2}}{\sum_k z_\gamma^{k^2}} \right)^2 + \gamma \right\} \\ &= z_\gamma^{-2} \{ -\phi'(y_\gamma) - \phi(y_\gamma) + \gamma \} = -z_\gamma^{-2} \phi'(y_\gamma) > 0. \end{aligned}$$

**Proof of the second assertion.** We apply the saddle-point method to the integral representing  $N_1$  see, *e.g.*, Chapter IX in Dieudonné (1968). It holds that

$$N_1(d^*, \gamma) = \frac{1}{2\pi i} \oint_{|z|=z_\gamma} \frac{h(z)^{d^*}}{z^{\gamma d^*}} \frac{dz}{z(1-z)} = \frac{1}{2\pi i} \oint_{|z|=z_\gamma} \{z(1-z)\}^{-1} e^{d^* l_\gamma(z)} dz. \quad (13)$$

The first assertion of the proposition provided us with a real number  $z_\gamma$  such that  $l'_\gamma(z_\gamma) = 0$  and  $l''_\gamma(z_\gamma) > 0$ . The tangent to the steepest descent curve at  $z_\gamma$  is vertical. The path we choose for integration is the circle with center 0 and radius  $z_\gamma$ . As this circle and the steepest descent curve have the same tangent at  $z_\gamma$ , applying formula (1.8.1) of Dieudonné (1968) (with  $\alpha = 0$  since  $l''(z_\gamma)$  is real and positive), we get that

$$\frac{1}{2\pi i} \oint_{|z|=z_\gamma} \{z(1-z)\}^{-1} e^{d^* l_\gamma(z)} dz = \frac{1}{2\pi i} \sqrt{\frac{2\pi}{d^* l''_\gamma(z_\gamma)}} e^{i\pi/2} \{z_\gamma(1-z_\gamma)\}^{-1} e^{d^* l_\gamma(z_\gamma)} (1 + o(1)),$$

when  $d^* \rightarrow \infty$ , as soon as the condition<sup>1</sup>  $\Re[l_\gamma(z) - l_\gamma(z_\gamma)] \leq -\mu$  is satisfied for some  $\mu > 0$  and for any  $z$  belonging to the circle  $|z| = |z_\gamma|$  and lying not too close to  $z_\gamma$ . To check that this is indeed the case, we remark that  $\Re[l_\gamma(z)] = \log \left| \frac{h(z)}{z^\gamma} \right|$ . Hence, if  $z = z_\gamma e^{i\omega}$  with  $\omega \in [\omega_0, 2\pi - \omega_0]$  for some  $\omega_0 \in ]0, \pi[$ , then

$$\begin{aligned} \left| \frac{h(z)}{z^\gamma} \right| &= \frac{|1 + 2z + 2 \sum_{k>1} z^{k^2}|}{z_\gamma^\gamma} \\ &\leq \frac{|1 + z| + z_\gamma + 2 \sum_{k>1} z_\gamma^{k^2}}{z_\gamma^\gamma} \\ &\leq \frac{|1 + e^{i\omega_0} z_\gamma| + z_\gamma + 2 \sum_{k>1} z_\gamma^{k^2}}{z_\gamma^\gamma}. \end{aligned}$$

Therefore  $\Re[l_\gamma(z) - \Re[l_\gamma(z_\gamma)]] \leq -\mu$  with  $\mu = \log \left( \frac{1+2z_\gamma+\sum_{k>1} z_\gamma^{k^2}}{|1+z_\gamma e^{i\omega_0}|+z_\gamma+\sum_{k>1} z_\gamma^{k^2}} \right) > 0$ . This completes the proof for the term  $N_1(d^*, \gamma)$ . The term  $N_2(d^*, \gamma)$  can be dealt in the same way.

## Appendix C. Proof of Lemma 5

Let  $\phi(\cdot)$  be the density of  $\mathcal{N}(0, 1)$  and let

$$p_f(\mathbf{y}) \triangleq \prod_{i=1}^n \phi(y_i - f(\mathbf{X}_i)), \quad \forall \mathbf{y} \in \mathbb{R}^n.$$

---

1.  $\Re u$  stands for the real part of the complex number  $u$ .

Since the errors  $\varepsilon_i$  are Gaussian, the posterior probabilities  $\mathbb{P}_0$  and  $\mathbb{P}_1$  are absolutely continuous w.r.t. the Lebesgue measure on  $\mathbb{R}^n$  and admit the densities

$$p_0(\mathbf{y}) = \prod_{i=1}^n \phi(y_i), \quad \text{and} \quad p_1(\mathbf{y}) = \mathbf{E}_{\mathbf{f} \sim \mu_S} p_{\mathbf{f}}(\mathbf{y}), \quad \forall \mathbf{y} \in \mathbb{R}^n.$$

Simple algebra yields:

$$p_{\mathbf{f}}(\mathbf{y}) = C_{\mathbf{f}} p_0(\mathbf{y}) \prod_{i=1}^n \exp \left\{ y_i \mathbf{f}(\mathbf{X}_i) \right\}, \quad \forall \mathbf{y} \in \mathbb{R}^n,$$

where  $C_{\mathbf{f}} = \prod_{i=1}^n \exp \left\{ -\mathbf{f}(\mathbf{X}_i)^2 / 2 \right\}$ . Thus,

$$\frac{p_1}{p_0}(\mathbf{y}) = \mathbf{E}_{\mathbf{f} \sim \mu_S} \left[ C_{\mathbf{f}} \prod_{i=1}^n \exp \left\{ y_i \mathbf{f}(\mathbf{X}_i) \right\} \right].$$

Therefore,

$$\begin{aligned} \int_{\mathbb{R}^n} \left( \frac{p_1}{p_0}(\mathbf{y}) \right)^2 p_0(\mathbf{y}) d\mathbf{y} &= \mathbf{E}_{(\mathbf{f}, \mathbf{f}') \sim \mu_S \otimes \mu_S} \left[ C_{\mathbf{f}} C_{\mathbf{f}'} \int_{\mathbb{R}^n} \prod_{i=1}^n \left( \exp \left\{ y_i (\mathbf{f} + \mathbf{f}')(\mathbf{X}_i) \right\} \phi(y_i) \right) d\mathbf{y} \right] \\ &= \mathbf{E}_{(\mathbf{f}, \mathbf{f}') \sim \mu_S \otimes \mu_S} \left[ C_{\mathbf{f}} C_{\mathbf{f}'} \prod_{i=1}^n \exp \left( \frac{1}{2} (\mathbf{f} + \mathbf{f}')^2(\mathbf{X}_i) \right) \right] \\ &= \mathbf{E}_{(\mathbf{f}, \mathbf{f}') \sim \mu_S \otimes \mu_S} \left[ \exp \left( \sum_{i=1}^n \mathbf{f}(\mathbf{X}_i) \mathbf{f}'(\mathbf{X}_i) \right) \right] \\ &= \frac{1}{2^{2|S|}} \sum_{\omega, \omega' \in \{\pm 1\}^S} \prod_{\mathbf{k}, \mathbf{k}' \in S} \exp \left( \omega_{\mathbf{k}} \omega'_{\mathbf{k}'} b_{\mathbf{k}\mathbf{k}'} \right), \end{aligned}$$

where  $b_{\mathbf{k}\mathbf{k}'} = A^2 \sum_{i=1}^n \varphi_{\mathbf{k}}(\mathbf{X}_i) \varphi_{\mathbf{k}'}(\mathbf{X}_i)$ , for all  $\mathbf{k}, \mathbf{k}' \in S$ . Note that  $0 \leq b_{\mathbf{k}\mathbf{k}} \leq 2A^2 n$  and  $|b_{\mathbf{k}\mathbf{k}'}| \leq A^2 n \epsilon$ , for all  $\mathbf{k}, \mathbf{k}' \in S$  such that  $\mathbf{k}' \neq \mathbf{k}$ . Now, on the one hand, for a fixed pair  $(\omega, \omega')$ , we have

$$\prod_{\mathbf{k} \neq \mathbf{k}'} \exp \left( \omega_{\mathbf{k}} \omega'_{\mathbf{k}'} b_{\mathbf{k}\mathbf{k}'} \right) \leq \exp(|S|^2 A^2 n \epsilon).$$

On the other hand, if we are given a sequence of numbers  $(b_{\mathbf{k}\mathbf{k}})$  indexed by  $S$ , we have

$$\frac{1}{2^{2|S|}} \sum_{\omega, \omega'} \prod_{\mathbf{k} \in S} e^{\omega_{\mathbf{k}} \omega'_{\mathbf{k}} b_{\mathbf{k}\mathbf{k}}} = \prod_{\mathbf{k} \in S} \frac{e^{b_{\mathbf{k}\mathbf{k}}} + e^{-b_{\mathbf{k}\mathbf{k}}}}{2} \leq \prod_{\mathbf{k} \in S} e^{b_{\mathbf{k}\mathbf{k}}} \leq \exp(4|S| A^4 n^2).$$

From these remarks it results that

$$\int_{\mathbb{R}}^d \left( \frac{p_1}{p_0}(\mathbf{y}) \right)^2 p_0(\mathbf{y}) d\mathbf{y} \leq \exp \left( 4|S| A^4 n^2 \left\{ 1 + \frac{|S|\epsilon}{4nA^2} \right\} \right),$$

and the claim of the lemma follows.

## Appendix D. Proof of Proposition 7

Let  $M = \binom{d}{d^*}$  and let  $\{f_0, f_1, \dots, f_M\}$  be a set included in  $\tilde{\Sigma}_L$ . Let  $I_1, \dots, I_M$  be all the subsets of  $\{1, \dots, d\}$  containing exactly  $d^*$  elements somehow enumerated. Let us set  $f_0 \equiv 0$  and define  $f_\ell$ , for  $\ell \neq 0$ , by its Fourier coefficients  $\{\theta_k^\ell : k \in \mathbb{Z}^d\}$  as follows:

$$\theta_k^\ell = \begin{cases} 1, & k = (k_1, \dots, k_d) = (\mathbf{1}_{1 \in I_\ell}, \dots, \mathbf{1}_{d \in I_\ell}), \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, all the functions  $f_\ell$  belong to  $\Sigma$  and, moreover, each  $f_\ell$  has  $I_\ell$  as sparsity pattern. One easily checks that our choice of  $f_\ell$  implies  $\mathcal{K}(\mathbf{P}_{f_\ell}, \mathbf{P}_{f_0}) = n \|f_\ell - f_0\|_2^2 = n$ . Therefore, if  $\alpha \log M = \alpha \log \binom{d}{d^*} \geq n$ , the desired inequality is satisfied. To conclude it suffices to note that  $\log \binom{d}{d^*}$  is larger than or equal to  $d^* \log(d/d^*) = d^*(\log d - \log d^*)$ .

COMMINGES DALALYAN

# Multiclass Learnability and the ERM principle

**Amit Daniely**

AMIT.DANIELY@MAIL.HUJI.AC.IL

*Dept. of Mathematics, The Hebrew University, Jerusalem, Israel*

**Sivan Sabato**

SIVAN\_SABATO@CS.HUJI.AC.IL

*School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel*

**Shai Ben-David**

SHAI@CS.UWATERLOO.CA

*David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada*

**Shai Shalev-Shwartz**

SIVANS@CS.HUJI.AC.IL

*School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel*

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

Multiclass learning is an area of growing practical relevance, for which the currently available theory is still far from providing satisfactory understanding. We study the learnability of multiclass prediction, and derive upper and lower bounds on the sample complexity of multiclass hypothesis classes in different learning models: batch/online, realizable/unrealizable, full information/bandit feedback. Our analysis reveals a surprising phenomenon: In the multiclass setting, in sharp contrast to binary classification, not all Empirical Risk Minimization (ERM) algorithms are equally successful. We show that there exist hypotheses classes for which some ERM learners have lower sample complexity than others. Furthermore, there are classes that are learnable by some ERM learners, while other ERM learner will fail to learn them. We propose a principle for designing good ERM learners, and use this principle to prove tight bounds on the sample complexity of learning *symmetric* multiclass hypothesis classes (that is, classes that are invariant under any permutation of label names). We demonstrate the relevance of the theory by analyzing the sample complexity of two widely used hypothesis classes: generalized linear multiclass models and reduction trees. We also obtain some practically relevant conclusions.

**Keywords:** List of keywords

## 1. Introduction

The task of multiclass learning, that is learning to classify an object into one of many candidate classes, surfaces in many domains including document categorization, object recognition in computer vision, and web advertisement.

The centrality of the multiclass learning problem has spurred the development of various approaches for tackling the task. Many of the methods define a set of possible multiclass predictors,  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  (where  $\mathcal{X}$  is the data domain and  $\mathcal{Y}$  is the set of labels), called the hypothesis class, and then use the training examples to choose a predictor from  $\mathcal{H}$  (for instance Crammer and Singer, 2003). In this paper we study the sample complexity of such hypothesis classes, namely, how many training examples are needed for learning an accurate predictor. This question has been extensively studied and is quite well understood

for the binary case, where  $|\mathcal{Y}| = 2$ . In contrast, the existing theory of the multiclass case, where  $|\mathcal{Y}| > 2$ , is much less complete.

We study multiclass sample complexity in several learning models. These models vary in three aspects:

- Interaction with the data source (batch vs. online protocols): In the batch protocol, we assume that the training data is generated i.i.d. by some distribution  $\mathcal{D}$  over  $\mathcal{X} \times \mathcal{Y}$ . The goal is to find a predictor  $h$  with a small probability to err,  $\Pr_{(x,y) \sim \mathcal{D}}(h(x) \neq y)$ , with a high probability over training samples. In the online protocol we receive examples one by one and are asked to predict the labels on the fly. Our goal is to make as few prediction mistakes as possible in the worst case (see Littlestone (1987)).
- The underlying labeling mechanism (realizable vs. agnostic): In the realizable case, we assume that the labels of the instances are determined by some  $h^* \in \mathcal{H}$ . In the agnostic case no restrictions on the labeling rule are imposed, and our goal is to make predictions which are not much worse than the best predictor in  $\mathcal{H}$ .
- The type of feedback (full information vs. bandits): In the full information setting, each example is revealed to the learner along with its correct label. In the bandit setting, the learner first sees an unlabeled example, and then outputs its guess for the label. Then a binary feedback is received, indicating only whether the guess was correct or not, but not revealing the correct label in the case of a wrong guess (see for example Auer et al. (2003, 2002); Kakade et al. (2008)).

In Section 2 we consider multiclass sample complexity in the PAC model (namely, the batch protocol with full information). Natarajan (1989) provides a characterization of multiclass PAC learnability in terms of a parameter of  $\mathcal{H}$  known as the Natarajan dimension and denoted  $d_N(\mathcal{H})$  (see section 2.2 for the relevant definitions). For the realizable case we show in Section 2.3 that there are constants  $C_1, C_2$  such that the sample complexity of learning  $\mathcal{H}$  with error  $\epsilon$  and confidence  $1 - \delta$  satisfies

$$C_1 \left( \frac{d + \ln(\frac{1}{\delta})}{\epsilon} \right) \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \left( \frac{d (\ln(\frac{1}{\epsilon}) + \ln(|\mathcal{Y}|) + \ln(d)) + \ln(\frac{1}{\delta})}{\epsilon} \right), \quad (1)$$

where  $d = d_N(\mathcal{H})$ . This improves the best previously known upper bound (theorem 5), in which there is a dependence on  $\ln(|\mathcal{Y}|) \cdot \ln(\frac{1}{\epsilon})$ .

The Natarajan dimension is equal to the VC dimension when  $|\mathcal{Y}| = 2$ . However, for larger label sets  $\mathcal{Y}$ , the bound on the sample complexity is not as tight as the known bound for the binary case, where the gap between the lower and upper bounds is only logarithmic in  $1/\epsilon$ . This invokes the challenge of tightening these sample complexity bounds for the multiclass case. A common approach to proving sample complexity bounds for PAC learning is to carefully analyze the sample complexity of ERM learners. In the case of PAC learning, all ERM learners have the same sample complexity (up to a logarithmic factor, see (Vapnik, 1995)). However, rather surprisingly, this is not the case for multiclass learning<sup>1</sup>.

---

1. Note that Shalev-Shwartz et al. (2010) established gaps between ERM learners in the general learning setting. However, here we consider multiclass learning, which seems very similar to binary classification.

In Section 2.4 we describe a family of concept classes for which there exist “good” ERM learner and “bad” ERM learner with a large gap between their sample complexities. Analyzing these examples, we deduce a rough principle on how to choose a good ERM learner. We also determine the sample complexity of the worst ERM learner for a given concept class,  $\mathcal{H}$ , up to a multiplicative factor of  $O(\ln(\frac{1}{\epsilon}))$ . We further show that if  $|\mathcal{Y}|$  is infinite, then there are hypotheses classes that are learnable by *some* ERM learners but not by *all* ERM learners. In Section 2.5 we employ the suggested principle to derive an improved sample complexity upper bound for *symmetric* classes ( $\mathcal{H}$  is symmetric if  $\phi \circ f \in \mathcal{H}$  whenever  $f \in \mathcal{H}$  and  $\phi$  is a permutation of  $\mathcal{Y}$ ). Symmetric classes are useful, since they are a natural choice when there is no prior knowledge about the relations between the possible labels. Moreover, many popular hypothesis classes that are used in practice are symmetric.

We conjecture that the upper bound obtained for symmetric classes holds for the sample complexity of non-symmetric classes as well. Such a result cannot be implied by uniform convergence alone, since, by the results mentioned above, there always exist bad ERM learners whose sample complexity is higher than this conjectured upper bound. It therefore seems that a proof for our conjecture will require the derivation of new learning rules. We hope that this would lead to new insights in other statistical learning problems as well.

In Section 3 we study multiclass learnability in the online model. We describe a simple generalization of the Littlestone dimension, and derive tight lower and upper bounds on the number, in terms of that dimension, of mistakes the optimal online algorithm will make in the worst case. Section 4 is devoted to a discussion of sample complexity of multiclass learning in the Bandit settings. Finally, in Section 5 we calculate the sample complexity of some popular families of hypothesis classes, which include linear multiclass hypotheses and filter trees, and discuss some practical implications of our bounds.

## 2. Multiclass Learning in the PAC Model

### 2.1. Problem Setting and Notation

For a distribution  $\mathcal{D}$  over  $\mathcal{X} \times \mathcal{Y}$ , the error of a function  $f \in \mathcal{H}$  with respect to  $\mathcal{D}$  is  $\text{Err}(f) = \text{Err}_{\mathcal{D}}(f) = \Pr_{(x,y) \sim \mathcal{D}}(f(x) \neq y)$ . A *learning algorithm* for a class  $\mathcal{H}$  is a function,  $A : \cup_{n=0}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathcal{H}$ . We denote a training sequence by  $S_m = (x_1, y_1), \dots, (x_m, y_m)$ . An *ERM learner* for class  $\mathcal{H}$  is a learning algorithm that for any sample  $S_m$  returns a function  $f \in \mathcal{H}$  that minimizes the number of sample errors  $|\{i \in [m] : f(x_i) \neq y_i\}|$ . This work focuses on statistical properties of the learning algorithms and ignores computatational complexity aspects.

The (*agnostic*) *sample complexity* of an algorithm  $A$  is the function  $m_A^a$  defined as follows: For every  $\epsilon, \delta > 0$ ,  $m_A^a(\epsilon, \delta)$  is the minimal integer such that for every  $m \geq m_A^a(\epsilon, \delta)$  and every distribution  $\mathcal{D}$  on  $\mathcal{X} \times \mathcal{Y}$ ,

$$\Pr_{S_m \sim \mathcal{D}^m} \left( \text{Err}(A(S_m)) > \inf_{f \in \mathcal{H}} \text{Err}(f) + \epsilon \right) \leq \delta. \quad (2)$$

If there is no integer satisfying these requirements, define  $m_A^a(\epsilon, \delta) = \infty$ . The (*agnostic*) *sample complexity* of a class  $\mathcal{H}$  is

$$m_{\mathcal{H}}^a(\epsilon, \delta) = \inf_A m_A^a(\epsilon, \delta),$$

where the infimum is taken over all learning algorithms.

We say that a distribution  $\mathcal{D}$  is realizable by a hypothesis class  $\mathcal{H}$  if there exists some  $f \in \mathcal{H}$  such that  $\text{Err}_{\mathcal{D}}(f) = 0$ . The *realizable sample complexity of an algorithm A* for a class  $\mathcal{H}$ , denoted  $m_A^r$ , is the minimal integer such that for every  $m \geq m_A^r(\epsilon, \delta)$  and every distribution  $\mathcal{D}$  on  $\mathcal{X} \times \mathcal{Y}$  which is realizable by  $\mathcal{H}$ , Equation. (2) holds. The realizable sample complexity of a class  $\mathcal{H}$  is  $m_{\mathcal{H}}^r(\epsilon, \delta) = \inf_A m_A^r(\epsilon, \delta)$  where the infimum is taken over all learning algorithms.

## 2.2. Known Sample Complexity Results

We first survey some known results regarding the sample complexity of multiclass learning. We start with the realizable case and then discuss the agnostic case. Given a subset  $S \subseteq \mathcal{X}$ , we denote  $\mathcal{H}|_S = \{f|_S : f \in \mathcal{H}\}$ . Recall the definition of the Vapnik-Chervonenkis dimension (Vapnik, 1995):

**Definition 1 (VC dimension)** Let  $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$  be a hypothesis class. A subset  $S \subseteq \mathcal{X}$  is shattered by  $\mathcal{H}$  if  $\mathcal{H}|_S = \{0, 1\}^S$ . The VC-dimension of  $\mathcal{H}$ , denoted  $\text{VC}(\mathcal{H})$ , is the maximal cardinality of a subset  $S \subseteq \mathcal{X}$  that is shattered by  $\mathcal{H}$ .

The VC-dimension is cornerstone in statistical learning theory as it characterizes the sample complexity of a *binary* hypothesis class. Namely

**Theorem 2 (Vapnik, 1995)** There are absolute constants  $C_1, C_2 > 0$  such that the realizable sample complexity of every hypothesis class  $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$  satisfies

$$C_1 \left( \frac{\text{VC}(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon} \right) \leq m_{\mathcal{H}}^r(\epsilon, \delta) \leq C_2 \left( \frac{\text{VC}(\mathcal{H}) \ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon} \right).$$

Moreover, the upper bound is attained by any ERM learner.

It is natural to seek a generalization of the VC-Dimension to hypothesis classes of non-binary functions. A straightforward attempt is to redefine shattering of  $S \subset \mathcal{X}$  by the property  $\mathcal{H}|_S = \mathcal{Y}^S$ . However, this requirement is too strong and does not lead to tight bounds on the sample complexity. Instead, we recall two alternative generalizations, introduced by Natarajan (1989). In both definitions, shattering is redefined to require that for any partition of  $S$  into  $T$  and  $S \setminus T$ , there exists a  $g \in \mathcal{H}$  whose behavior on  $T$  differs from its behavior on  $S \setminus T$ . The two definitions differ in how “different behavior” is defined.

**Definition 3 (Graph dimension and Natarajan dimension)** Let  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  be a hypothesis class and let  $S \subseteq \mathcal{X}$ . We say that  $\mathcal{H}$  G-shatters  $S$  if there exists an  $f : S \rightarrow \mathcal{Y}$  such that for every  $T \subseteq S$  there is a  $g \in \mathcal{H}$  such that

$$\forall x \in T, g(x) = f(x), \text{ and } \forall x \in S \setminus T, g(x) \neq f(x).$$

We say that  $\mathcal{H}$  N-shatters  $S$  if there exist  $f_1, f_2 : S \rightarrow \mathcal{Y}$  such that  $\forall y \in S, f_1(y) \neq f_2(y)$ , and for every  $T \subseteq S$  there is a  $g \in \mathcal{H}$  such that

$$\forall x \in T, g(x) = f_1(x), \text{ and } \forall x \in S \setminus T, g(x) = f_2(x).$$

The graph dimension of  $\mathcal{H}$ , denoted  $d_G(\mathcal{H})$ , is the maximal cardinality of a set that is G-shattered by  $\mathcal{H}$ . The Natarajan dimension of  $\mathcal{H}$ , denoted  $d_N(\mathcal{H})$ , is the maximal cardinality of a set that is N-shattered by  $\mathcal{H}$ .

Both of these dimensions coincide with the VC-dimension for  $|\mathcal{Y}| = 2$ . Note also that we always have  $d_N \leq d_G$ .

By reductions from and to the binary case, it is not hard to show, similarly to Natarajan (1989) and Ben-David et al. (1995) (see Appendix A for a full proof), that

**Theorem 4** *For the constants  $C_1, C_2$  from theorem 2, for every  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  we have*

$$C_1 \left( \frac{d_N(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon} \right) \leq m_{\mathcal{H}}^r(\epsilon, \delta) \leq C_2 \left( \frac{d_G(\mathcal{H}) \ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon} \right).$$

Moreover, the upper bound is attained by any ERM learner.

From this theorem it follows that the finiteness of the Natarajan dimension is a necessary condition for learnability, and the finiteness of the graph dimension is a sufficient condition for learnability. In Ben-David et al. (1995) it was proved that for every concept class  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ ,

$$d_N(\mathcal{H}) \leq d_G(\mathcal{H}) \leq 4.67 \log_2(|\mathcal{Y}|) d_N(\mathcal{H}). \quad (3)$$

It follows that if  $|\mathcal{Y}| < \infty$  then the finiteness of the Natarajan dimension is a necessary and sufficient condition for learnability. Incorporating Equation. (3) into theorem 4, it can be seen that the Natarajan dimension, as well as the graph dimension, characterize the sample complexity of  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  up to a multiplicative factor of  $O(\log(|\mathcal{Y}|) \log(\frac{1}{\epsilon}))$ . Precisely,

**Theorem 5** *(Ben-David et al., 1995) For the constants  $C_1, C_2$  from theorem 2,*

$$C_1 \left( \frac{d_N(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon} \right) \leq m_{\mathcal{H}}^r(\epsilon, \delta) \leq C_2 \left( \frac{d_N(\mathcal{H}) \cdot \ln(|\mathcal{Y}|) \cdot \ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon} \right).$$

Moreover, the upper bound is attained by any ERM learner.

A similar analysis can be performed for the agnostic case. For binary classification we have that for every hypothesis class  $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ ,

$$m_{\mathcal{H}}^a(\epsilon, \delta) = \Theta \left( \frac{1}{\epsilon^2} \left( VC(\mathcal{H}) + \ln(\frac{1}{\delta}) \right) \right), \quad (4)$$

and this is attained by any ERM learner. Here too it is possible to obtain by reduction from and to the binary case that for every hypothesis class  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ ,

$$\Omega \left( \frac{1}{\epsilon^2} \left( d_N(\mathcal{H}) + \ln(\frac{1}{\delta}) \right) \right) \leq m_{\mathcal{H}}^a(\epsilon, \delta) \leq O \left( \frac{1}{\epsilon^2} \left( d_G(\mathcal{H}) + \ln(\frac{1}{\delta}) \right) \right). \quad (5)$$

By Equation. (3) we have

$$m_{\mathcal{H}}^a(\epsilon, \delta) = O \left( \frac{1}{\epsilon^2} \left( \log(|\mathcal{Y}|) \cdot d_N(\mathcal{H}) + \ln(\frac{1}{\delta}) \right) \right). \quad (6)$$

Thus in the agnostic case as well, the Natarajan dimension characterizes the agnostic sample complexity up to a multiplicative factor of  $O(\log(|\mathcal{Y}|))$ . Here too, all of these bounds are attained by any ERM learner.

### 2.3. An Improved Result for the Realizable Case

The following theorem provides a sample complexity upper bound which can be better than Theorem 5 when  $\ln(d_N(\mathcal{H})) \ll \ln(|\mathcal{Y}|) \cdot \ln(\frac{1}{\epsilon})$ . The proof of the theorem is given in Appendix A. While the proof is a simple adaptation of previous results, we find it valuable to present this result here, as we could not find it in the research literature.

**Theorem 6** *For every concept class  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ ,*

$$m_{\mathcal{H}}^r(\epsilon, \delta) = O\left(\frac{d_N(\mathcal{H}) (\ln(\frac{1}{\epsilon}) + \ln(|\mathcal{Y}|) + \ln(d_N(\mathcal{H}))) + \ln(\frac{1}{\delta})}{\epsilon}\right).$$

Moreover, the bound is attained by any ERM learner.

Theorem 6 is the departure point of our research. As indicated above, one of our objectives is to prove sample complexity bounds for the multiclass case with a ratio of  $O(\ln(\frac{1}{\epsilon}))$  between the upper bound and the lower bound, as in the binary case. In the next section we show that such an improvement cannot be attained by uniform convergence analysis, since the ratio between the sample complexity of the worst ERM learner and the best ERM learner of a given hypothesis class might be as large as  $\ln(|\mathcal{Y}|)$ .

### 2.4. The Gap between “Good ERM” and “Bad ERM”

The tight bounds in the binary case given in Theorem 2 are attained by *any* ERM learner. In contrast to the binary case, we now show that in the multiclass case there can be a significant sample complexity gap between different ERM learners. Moreover, in the case of classification with an infinite number of classes, there are learnable hypothesis classes that some ERM learners fail to learn. We begin with showing that the graph dimension determines the sample complexity of the worst ERM learner up to a multiplicative factor of  $O(\ln(\frac{1}{\epsilon}))$ .

**Theorem 7** *There are absolute constants  $C_1, C_2 > 0$  such that for every hypothesis class  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  and every ERM learner  $A$ ,*

$$m_A^r(\epsilon, \delta) \leq C_2 \left( \frac{d_G(\mathcal{H}) \ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon} \right).$$

Moreover, there is an ERM learner  $A_{\text{bad}}$  such that

$$m_{A_{\text{bad}}}^r(\epsilon, \delta) \geq C_1 \left( \frac{d_G(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon} \right). \quad (7)$$

**Proof** The upper bound on  $m_A^r$  is just a restatement of theorem 4. It remains to prove that there exists an ERM learner,  $A_{\text{bad}}$ , satisfying (7). We shall first consider the case where  $d = d_G(\mathcal{H}) < \infty$ .

Let  $S = \{x_0, \dots, x_{d-1}\} \subseteq \mathcal{X}$  be a set which is  $G$ -Shattered by  $\mathcal{H}$  using the function  $f_0$ . Let  $A_{\text{bad}}$  be an ERM learner with the property that upon seeing a sample whose instances are in  $T \subseteq S$ , and whose labels are determined by  $f_0$ , it returns  $f \in \mathcal{H}$  such that  $f$  equals

to  $f_0$  on  $T$  and  $f$  is different from  $f_0$  on  $S \setminus T$ . The existence of such an  $f$  follows from the assumption that  $S$  is G-shattered using  $f_0$ .

Fix  $\delta < e^{-1/6}$  and let  $\epsilon$  small enough such that  $1 - 2\epsilon \geq e^{-4\epsilon}$ . Define a distribution on  $\mathcal{X}$  by setting  $\Pr(x_0) = 1 - 2\epsilon$  and for all  $1 \leq i \leq d - 1$ ,  $\Pr(x_i) = \frac{2\epsilon}{d-1}$ . Suppose that the correct hypothesis is  $f_0$  and let the sample size be  $m$ . Clearly, the hypothesis returned by  $A_{\text{bad}}$  will err on all the examples from  $S$  which are not in the sample. By Chernoff's bound, if  $m \leq \frac{d-1}{6\epsilon}$ , then with probability  $\geq e^{-\frac{1}{6}} \geq \delta$ , the sample will include no more than  $\frac{d-1}{2}$  examples from  $S$ . Thus the returned hypothesis will have error  $\geq \epsilon$ . Moreover, the probability that the sample includes only  $x_0$  (and thus  $A_{\text{bad}}$  will return a hypothesis with error  $2\epsilon$ ) is  $(1 - 2\epsilon)^m \geq e^{-4\epsilon m}$ , which is more than  $\delta$  if  $m \leq \frac{1}{4\epsilon} \ln(\frac{1}{\delta})$ . We therefore obtain that

$$m_{A_{\text{bad}}}^r(\epsilon, \delta) \geq \max \left\{ \frac{d-1}{6\epsilon}, \frac{1}{2\epsilon} \ln(1/\delta) \right\} \geq \frac{d-1}{12\epsilon} + \frac{1}{4\epsilon} \ln(1/\delta),$$

as required. If  $d_G(\mathcal{H}) = \infty$  then the argument above can be repeated for a sequence of pairwise disjoint G-shattered sets  $S_n$ ,  $n = 1, 2, \dots$  with  $|S_n| = n$ .  $\blacksquare$

The following example shows that in some cases there are learning algorithms that are much better than the worst ERM:

**Example 1 (A Large Gap Between ERM Learners)** Let  $\mathcal{X}_0$  be any finite or countable domain set and let  $\mathcal{X}$  be some subset of  $\mathcal{X}_0$ . Let  $\mathcal{P}_f(\mathcal{X})$  denote the collection of finite and co-finite subsets  $A \subseteq \mathcal{X}$ . For every  $A \in \mathcal{P}_f(\mathcal{X})$ , define  $f_A : \mathcal{X}_0 \rightarrow \mathcal{P}_f(\mathcal{X}) \cup \{\ast\}$  by

$$f_A(x) = \begin{cases} A & \text{if } x \in A \\ \ast & \text{otherwise,} \end{cases}$$

and consider the concept family  $\mathcal{H}_{\mathcal{X}} = \{f_A : A \in \mathcal{P}_f(\mathcal{X})\}$ . We first note that any ERM learner that sees an example of the form  $(x, A)$  for some  $A \subseteq \mathcal{X}$  must return the hypothesis  $f_A$ , thus to define an ERM learner we only have to specify the hypothesis it returns upon seeing a sample of the form  $S_m = \{(x_1, \ast), \dots, (x_m, \ast)\}$ . Note also that  $\mathcal{X}$  is G-shattered using the function  $f_{\emptyset}$ , and therefore  $d_G(\mathcal{H}_{\mathcal{X}}) \geq |\mathcal{X}|$  (it is easy to see that, in fact  $d_G(\mathcal{H}_{\mathcal{X}}) = |\mathcal{X}|$ ).

We consider two ERM learners –  $A_{\text{good}}$ , which on a sample of the form  $S_m$  returns the hypothesis  $f_{\emptyset}$ , and  $A_{\text{bad}}$ , which, upon seeing  $S_m$ , returns  $f_{\{x_1, \dots, x_m\}^c}$ , thus satisfying the specification of a bad ERM algorithm from the proof of Theorem 7. It follows that the sample complexity of  $A_{\text{bad}}$  is  $\Omega\left(\frac{|\mathcal{X}|}{\epsilon} + \frac{1}{\epsilon} \ln(\frac{1}{\delta})\right)$ . On the other hand,

**Claim 1** The sample complexity of  $A_{\text{good}}$  is at most  $\frac{1}{\epsilon} \ln \frac{1}{\delta}$ .

**Proof** Let  $\mathcal{D}$  be a distribution over  $\mathcal{X}_0$  and suppose that the correct labeling is  $f_A$ . Let  $m$  be the size of the sample. For any sample,  $A_{\text{good}}$  returns either  $f_{\emptyset}$  or  $f_A$ . If it returns  $f_A$  then its generalization error is zero. Thus, it returns a hypothesis with error  $\geq \epsilon$  only if  $\Pr_{\mathcal{D}}(A) \geq \epsilon$  and all the  $m$  examples in the sample are from  $A^c$ . Assume  $m \geq \frac{1}{\epsilon} \ln(\frac{1}{\delta})$ , then probability of the latter event is no more than  $(1 - \epsilon)^m \leq e^{-\epsilon m} \leq \delta$ .  $\blacksquare$

Since  $\mathcal{X}$  can be infinite in the above example we conclude that

**Corollary 8** *There exist sets  $\mathcal{X}, \mathcal{Y}$  and a hypothesis class  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ , such that  $\mathcal{H}$  is learnable by some ERM learner but is not learnable by some other ERM learner.*

What is the crucial feature that makes  $A_{\text{good}}$  better than  $A_{\text{bad}}$ ? If the correct labeling is  $f_A \in \mathcal{H}_{\mathcal{X}}$ , then for *any* sample,  $A_{\text{good}}$  might return at most one of two functions – namely  $f_A$  or  $f_{\emptyset}$ . On the other hand, if the sample is labeled by the function  $f_{\emptyset}$ ,  $A_{\text{bad}}$  might return *every* function in  $\mathcal{H}_{\mathcal{X}}$ . Thus, to return a hypothesis with error  $\leq \epsilon$ ,  $A_{\text{good}}$  needs to reject only one hypothesis while  $A_{\text{bad}}$  needs to reject many more. We conclude the following (rough) principle: *A good ERM is an ERM that, for every target hypothesis, consider a small number of hypotheses.*

Next, we formalize the above intuition by proving a general theorem that enables us to derive sample complexity bounds for ERM learners that are designed using the above principle. Fix a hypothesis class  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ . We view an ERM learner as an operator that for any  $f \in \mathcal{H}, S \subseteq \mathcal{X}$  takes the partial function  $f|_S$  as input and extends it to a function  $g = A(f|_S) \in \mathcal{H}$  such that  $g|_S = f|_S$ . For every  $f \in \mathcal{H}$ , denote by  $F_A(f)$  the set of all the functions that the algorithm  $A$  might return upon seeing a sample of the form  $\{(x_i, f(x_i))\}_{i=1}^m$  for some  $m \geq 0$ . Namely,

$$F_A(f) = \{A(f|_S) : S \subseteq \mathcal{X}, |S| < \infty\}$$

To provide an upper bound on  $m_A^r(\epsilon, \delta)$ , it suffices to show that for every  $f \in \mathcal{H}$ , with probability at least  $1 - \delta$ , all the functions with error at least  $\epsilon$  in  $F_A(f)$  will be rejected after seeing  $m$  examples. This is formalized in the following theorem.

**Theorem 9** *Let  $A$  be an ERM learner for a hypothesis class  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ . Define the growth function of  $A$  by  $\Pi_A(m) = \sup_{f \in \mathcal{H}} \Pi_{F_A(f)}(m)$ , where for  $F \subseteq \mathcal{Y}^{\mathcal{X}}$ ,  $\Pi_F(m) = \sup\{|F|_S| : S \subseteq \mathcal{X}, |S| \leq m\}$ . Then*

$$m_A^r(\epsilon, \delta) \leq \min\{m : \Pi_A(2m) 2^{1-\frac{\epsilon m}{2}} < \delta\}.$$

The theorem immediately follows from the following lemma.

**Lemma 10 (The Double Sampling Lemma)** *Let  $A$  be an ERM learner. Fix a distribution  $\mathcal{D}$  over  $\mathcal{X}$  and a function  $f_0 \in \mathcal{H}$ . Denote by  $A_m$  the event that, after seeing  $m$  i.i.d. examples drawn from  $\mathcal{D}$  and labeled by  $f_0$ ,  $A$  returns a hypothesis with error at least  $\epsilon$ . Then  $\Pr(A_m) \leq 2 \cdot \Pi_A(2m) 2^{-\frac{\epsilon m}{2}}$ .*

**Proof** Let  $S_1$  and  $S_2$  be two samples of  $m$  i.i.d. examples labeled by  $f_0$ . Let  $B_m$  be the event that there exists a function  $f \in \mathcal{H}$  with error at least  $\epsilon$ , such that (1)  $f$  is not rejected by  $S_1$  (i.e.  $f_0(x) = f(x)$  for all examples  $x$  in  $S_1$ ), and (2) there exist at least  $\frac{\epsilon m}{2}$  examples  $(x, f_0(x))$  in  $S_2$  for which  $f(x) \neq f_0(x)$ . By Chernoff's bound, for  $m = \Omega(\frac{1}{\epsilon})$ ,  $\Pr(B_m) = \Pr(B_m|A_m) \Pr(A_m) \geq \frac{1}{2} \Pr(A_m)$ . W.l.o.g., we can assume that  $S_1, S_2$  are generated as follows: First,  $2m$  examples are drawn to create a sample  $U$ . Then  $S_1$  and  $S_2$  are generated by selecting a random partition of  $U$  into two samples of size  $m$ . Now,  $\Pr(B_m)$  is bounded by the probability that there is an  $f \in \mathcal{H}|_U$  such that (1) there are at least  $\frac{\epsilon m}{2}$  examples in  $U$  such that  $f$  disagrees with  $f_0$  on these examples and (2) all of these

examples are located in  $S_2$ . For a single  $f \in \mathcal{H}|_U$  that disagrees with  $f_0$  on  $l \geq \frac{\epsilon m}{2}$  samples, the probability that all these examples are located in  $S_2$  is  $\binom{m}{l}/\binom{2m}{l} \leq 2^{-l} \leq 2^{-\frac{\epsilon m}{2}}$ . Thus, using the union bound we obtain that  $\Pr(B_m) \leq |\mathcal{H}|_U 2^{-\frac{\epsilon m}{2}} \leq \Pi(2m) 2^{-\frac{\epsilon m}{2}}$ . ■

The bound in theorem 6 is based on the (trivial) inequality  $\Pi_A \leq \Pi_{\mathcal{H}}$ . However, as Example 1 shows,  $\Pi_A$  can be much smaller than  $\Pi_{\mathcal{H}}$ . As we shall see in the sequel, we can apply the double sampling lemma to get better sample complexity bounds for “good” ERM learners. The key tool for these sample complexity bounds is Lemma 12, that is, in turn, based on the following combinatorial result:

**Lemma 11** (Natarajan, 1989) *For every hypothesis class  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ ,  $|\mathcal{H}| \leq |\mathcal{X}|^{d_N(\mathcal{H})} |\mathcal{Y}|^{2d_N(\mathcal{H})}$ .*

**Lemma 12** *Let  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  be a class of functions. Assume that for some number  $r$ , for every  $h \in \mathcal{H}$ , the size of the range of  $h$  is at most  $r$ . Let  $A$  be an algorithm such that, for some set of values  $Y' \subseteq \mathcal{Y}$ , for every  $f \in \mathcal{H}$ , and every sample  $S_m = ((x_1, f(x_1)), \dots, (x_m, f(x_m)))$ , the function returned by  $A$  on input  $S_m$  is consistent with  $S_m$  and has its values in the set  $\{f(x_1), \dots, f(x_m)\} \cup Y'$ . Then,*

$$m_A^r(\epsilon, \delta) = O\left(\frac{d_N(\mathcal{H})(\ln(\frac{1}{\epsilon}) + \ln(\max\{r, |Y'|\})) + \ln(\frac{1}{\delta})}{\epsilon}\right).$$

**Proof** The assumptions of the lemma imply that, for every  $f \in \mathcal{H}$ , the range of the functions in  $F_A(f)$  is contained in the union of  $Y'$  and the range of  $f$ . Therefore, using Lemma 11 we obtain that  $\Pi_A(2m) \leq (2m)^{d_N(\mathcal{H})} (|Y'| + r)^{2d_N(\mathcal{H})}$ , and the bound follows from Theorem 9. ■

Note that classes in which each function  $h \in H$  uses at most  $r$  values, for some  $r < d_N(H) \log(|\mathcal{Y}|)$ , can have a large range  $\mathcal{Y}$  and a graph dimension that is significantly larger than their Natarajan dimension. In such cases, we may be able to show a gap between the sample complexity of bad and good ERM learners, by applying the lower bound from Theorem 7. In particular, we get such a result for the following family of hypotheses classes, which generalizes Example 1.

**Corollary 13** *Let  $\mathcal{H}$  be a class of functions from  $\mathcal{X}$  to some range set  $\mathcal{Y}$ , such that, for some value  $y_0 \in \mathcal{Y}$ , for every  $h \in H$ , the range of  $h$  contains at most one value besides  $y_0$ . Assume also that  $\mathcal{H}$  contains the constant  $y_0$  function. Let  $d$  denote the Natarajan dimension of  $\mathcal{H}$ . Then there exists an ERM learning algorithm  $A$  for  $H$  such that the  $(\epsilon, \delta)$  sample complexity of  $A$  is*

$$O\left(\frac{d \cdot \ln(1/\epsilon) + \ln(1/\delta)}{\epsilon}\right).$$

Every class in that family that has a large graph dimension will therefore realize a gap between the sample complexities of different ERM learners.

**Example 2** Consider the set of all balls in  $\mathbb{R}^n$  and, for each such ball,  $B = B(z, r)$  with center  $z$  and radius  $r$ , let  $h_B$  be the function defined by  $h_B(x) = z$  if  $x \in B$  and  $h_B(x) = \star$

otherwise. Let  $\mathcal{H}_{\mathcal{B}^n} = \{h_B : B = B(z, r) \text{ for some } z \in \mathbb{R}^n, r \in \mathbb{R}\} \cup \{h_\star\}$  (where  $h_\star$  is the constant  $\star$  function). It is not hard to see that  $d_N(\mathcal{H}_{\mathcal{B}^n}) = 1$  and  $d_G(\mathcal{H}_{\mathcal{B}^n}) = n+1$ . Furthermore, let  $A_{\text{good}}$  be the ERM learner that for every sample  $S = (x_1, f(x_1)), \dots, (x_m, f(x_m))$ , returns  $h_{B_S}$ , where  $B_S$  is the minimal ball that is consistent with the sample. Note that this algorithm uses, for every  $f \in \mathcal{H}_{\mathcal{B}^n}$  and every sample  $S$  labeled by such  $f$ , at most one value (the value  $\star$ ) on top of the values  $\{f(x_1), \dots, f(x_m)\}$ .

In this case, Theorem 7 implies that for some constant  $C_1$ , there exists a bad ERM learner,  $A_{\text{bad}}$  such that

$$m_{A_{\text{bad}}}^r(\epsilon, \delta) \geq C_1 \left( \frac{n + \ln(1/\delta)}{\epsilon} \right).$$

On the other hand, Lemma 12 implies that there is a good ERM learner,  $A_{\text{good}}$  and a constant  $C_2$  for which

$$m_{A_{\text{good}}}^r(\epsilon, \delta) \leq C_2 \left( \frac{\ln(1/\epsilon) + \ln(1/\delta)}{\epsilon} \right).$$

Note that, if one restricts the hypothesis class to allow only balls that have their centers in some finite set of grid points, the class uses only a finite range of labels. However, if such a grid is sufficiently dense, the sample complexities of both algorithms,  $A_{\text{bad}}$  and  $A_{\text{good}}$ , would not change.

## 2.5. Symmetric Classes

The principle for choosing a good ERM leads to tight bounds on the sample complexity of *symmetric classes*. Recall that a class  $\mathcal{H}$  is called symmetric if for any  $f \in \mathcal{H}$  and any permutation  $\phi$  on labels, we have that  $\phi \circ f \in \mathcal{H}$  as well.

**Theorem 14** *There are absolute constants  $C_1, C_2$  such that for every symmetric hypothesis class  $\mathcal{H} \subseteq \mathcal{Y}^\mathcal{X}$*

$$C_1 \left( \frac{d_N(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon} \right) \leq m_{\mathcal{H}}^r(\epsilon, \delta) \leq C_2 \left( \frac{d_N(\mathcal{H}) (\ln(\frac{1}{\epsilon}) + \ln(d_N(\mathcal{H}))) + \ln(\frac{1}{\delta})}{\epsilon} \right)$$

A key observation that enables us to employ our principle in this case is:

**Lemma 15** *Let  $\mathcal{H} \subseteq \mathcal{Y}^\mathcal{X}$  be a symmetric hypothesis class of Natarajan dimension  $d$ . Then, the range of any  $f \in \mathcal{H}$  is of size at most  $2d+1$ .*

**Proof** If  $|\mathcal{Y}| \leq 2d+1$  we are done. Thus assume that there are  $2d+2$  distinct elements  $y_1, \dots, y_{2d+2} \in \mathcal{Y}$ . Assume to the contrary that there is a hypothesis  $f \in \mathcal{H}$  with a range of more than  $d$  values. Thus there is a set  $S = \{x_1, \dots, x_{d+1}\} \subseteq \mathcal{X}$  such that  $f|_S$  has  $d+1$  values in its range. It follows that  $\mathcal{H}$  N-shatters  $S$ , thus reaching a contradiction. Indeed, since  $\mathcal{H}$  is symmetric, there are functions  $f_0, f_1 \in \mathcal{H}$  such that  $f_j(x_i) = y_{j(d+1)+i}$ . Similarly, for every  $T \subseteq S$ , there is a  $g \in \mathcal{H}$  such that  $g(x) = f_0(x)$  for every  $x \in T$  and  $g(x) = f_1(x)$  for every  $x \in S \setminus T$ .  $\blacksquare$

We are now ready to prove Theorem 14.

**Proof** (of Theorem 14) The lower bound is a restatement of Theorem 4. For the upper bound, we define an algorithm  $A$  that conforms to the conditions in Lemma 12: Fix a set  $\mathcal{Y}' \subseteq \mathcal{Y}$  of size  $|\mathcal{Y}'| = \min\{|\mathcal{Y}|, 2d_N(\mathcal{H}) + 1\}$ . Given a sample  $(x_1, f(x_1)), \dots, (x_m, f(x_m))$ ,  $A$  returns a hypothesis that is consistent with the sample and that attains only values in  $\{f(x_1), \dots, f(x_m)\} \cup \mathcal{Y}'$ . It is possible due to symmetry and Lemma 15. ■

A similar analysis can be performed for the agnostic case. Let  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  be a symmetric hypothesis class. Let  $\mathcal{Y}' \subseteq \mathcal{Y}$  be an arbitrary set of size  $\min\{|\mathcal{Y}|, 4d_N(\mathcal{H}) + 2\}$ . Denote  $\mathcal{H}' = \{f \in \mathcal{H} : f(\mathcal{X}) \subseteq \mathcal{Y}'\}$ . Using lemma 15 and symmetry, it is easy to see that  $d_G(\mathcal{H}) = d_G(\mathcal{H}')$  and  $d_N(\mathcal{H}) = d_N(\mathcal{H}')$ . By equation 3, we conclude that  $d_G(\mathcal{H}) = O(\log(d_N(\mathcal{H})) \cdot d_N(\mathcal{H}))$ . Using equation 5 we obtain a sample complexity bound of

$$m_{\mathcal{H}}^a(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2} \left( \log(\min\{d_N(\mathcal{H}), |\mathcal{Y}|\}) \cdot d_N(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right) \right)\right),$$

which is better than Equation. (6). Moreover, the ratio between this bound and the lower bound (Equation. (5)) is  $O(\log(d_N(\mathcal{H})))$  regardless of  $|\mathcal{Y}|$ . Note that this bound is attained by any ERM. We present the following open question:

**Open question 16** *Examples 1 and 2 show that there are (non-symmetric) hypothesis classes with a ratio of  $\Omega(\ln(|\mathcal{Y}|))$  between the sample complexities of the worst ERM learner and the best ERM learner. How large can this gap be for symmetric hypothesis classes?*

### 3. Multiclass Learning in the Online Model

Learning in the online model is conducted in a sequence of consecutive rounds. On each round  $t = 1, 2, \dots$ , the environment presents a sample  $x_t \in \mathcal{X}$ , the algorithm should predict a value  $\hat{y}_t \in \mathcal{Y}$ , and then the environment reveals the correct value  $y_t \in \mathcal{Y}$ . The prediction at time  $t$  can be based only on the examples  $x_1, \dots, x_t$  and the previous outcomes  $y_1, \dots, y_{t-1}$ . We start with the realizable case, in which we assume that for some function  $f \in \mathcal{H}$ , all the outcomes are evaluations of  $f$ , namely,  $y_t = f(x_t)$ . Given an online learning algorithm,  $A$ , define its (*realizable*) *sample complexity*,  $\mathcal{M}(A)$ , to be the maximal number of wrong predictions that it might make on a legal sequence of any length.

The sample complexity of online learning has been studied by Littlestone (1987), who showed that a combinatorial measure, called the Littlestone dimension, characterizes the sample complexity of online learning. We now propose a generalization of the Littlestone dimension to classes of non-binary functions.

Consider a rooted tree  $T$  whose internal nodes are labeled by  $\mathcal{X}$  and whose edges are labeled by  $\mathcal{Y}$ , such that the labels on edges from a parent to its child nodes are all different from each other. The tree  $T$  is *shattered* by  $\mathcal{H}$  if, for every path from root to leaf  $x_1, \dots, x_k$ , there is a function  $f \in \mathcal{H}$  such that  $f(x_i)$  equals the label of  $(x_i, x_{i+1})$ . The *Littlestone dimension*,  $L\text{-dim}(\mathcal{H})$ , of  $\mathcal{H}$  is the maximal depth of a complete binary tree that is shattered by  $\mathcal{H}$ .

It is not hard to see that, given a shattered tree of depth  $l$ , the environment can force any online learning algorithm to make  $l$  mistakes. Thus, for any algorithm  $A$ ,  $\mathcal{M}(A) \geq L\text{-Dim}(\mathcal{H})$ . We shall now present an algorithm whose sample complexity is upper bounded by  $L\text{-Dim}(\mathcal{H})$ .

**Algorithm:** Standard Optimal Algorithm (SOA)

Initialization:  $V_0 = \mathcal{H}$ .

For  $t = 1, 2, \dots$ ,

```

        receive  $x_t$ 
        for  $y \in \mathcal{Y}$ , let  $V_t^{(y)} = \{f \in V_{t-1} : f(x_t) = y\}$ 
        predict  $\hat{y}_t \in \arg \max_y \text{L-Dim}(V_t^{(y)})$ 
        receive true answer  $y_t$ 
        update  $V_t = V_t^{(t_t)}$ 
```

**Theorem 17**  $\mathcal{M}(\text{SOA}) = \text{L-Dim}(\mathcal{H})$ .

The proof is a simple adaptation of the proof of the binary case (see Littlestone, 1987). The idea is to note that for each  $t$  there is at most one  $y \in \mathcal{Y}$  with  $\text{L-Dim}(V_t^{(y)}) = \text{L-Dim}(V_t)$ , and for the rest of the labels we have  $\text{L-Dim}(V_t^{(y)}) < \text{L-Dim}(V_t)$ . Thus, whenever the algorithm errs, the Littlestone dimension of  $V_t$  decreases by at least 1, so after  $\text{L-Dim}(\mathcal{H})$  mistakes,  $V_t$  is composed of a single function.

Note that we only considered deterministic algorithms. However, allowing the algorithm to make randomized predictions does not substantially improve its sample complexity. It is easy to see that given a shattered tree of depth  $l$ , the environment can enforce any randomized online learning algorithm to make at least  $l/2$  mistakes on average.

In the agnostic case, the sequence of outcomes,  $y_1, \dots, y_m$ , is not necessarily realizable by some target function  $f \in \mathcal{H}$ . In that case, our goal is to have a *regret* of at most  $\epsilon$ , where the regret is defined as

$$\frac{1}{m} |\{t \in [m] : \hat{y}_t \neq y_t\}| - \min_{f \in \mathcal{H}} \frac{1}{m} |\{t \in [m] : f(x_t) \neq y_t\}|.$$

We denote by  $m_A^a(\epsilon)$  the number of examples required so that the regret of an algorithm  $A$  will be at most  $\epsilon$  and by  $m^a(\epsilon)$  the infimum, over all algorithms  $A$ , of  $m_A^a(\epsilon)$ .

Online learnability in the agnostic case, for classes of binary-output functions, has been studied in Ben-David et al. (2009), who showed that the Littlestone dimension characterizes the sample complexity in the agnostic case as well. The basic idea is to construct a set of experts by running the SOA algorithm on all sub-sequences of the examples whose length is at most  $\text{L-Dim}(\mathcal{H})$ , and then to run an online algorithm for learning with experts. This idea can be generalized to the multiclass case, but we leave this generalization to a longer version of this manuscript.

#### 4. The Bandit Setting

So far we have assumed that each learning example is comprised of an instance and its corresponding label. In this section we deal with the so-called bandit setting. In the bandit model, the learner does not get to see the correct label of a training example. Instead, the learner first receives an instance  $x \in \mathcal{X}$ , and should guess a label,  $\hat{y}$ . The learner then receives a binary feedback, indicating whether its guess is correct or not.

#### 4.1. Bandit vs Full Information in the Batch Model

Let  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  be a hypothesis class. Our goal is to analyze the *realizable bandit sample complexity* of  $\mathcal{H}$ , which we denote by  $m_{\mathcal{H}}^{r,b}(\epsilon, \delta)$ , and the *agnostic bandit sample complexity* of  $\mathcal{H}$ , which we denote by  $m_{\mathcal{H}}^{a,b}(\epsilon, \delta)$ . The following theorem provides upper bounds on the sample complexity.

**Theorem 18** *Let  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  be a hypothesis class. Then,*

$$m_{\mathcal{H}}^{r,b}(\epsilon, \delta) = O\left(|\mathcal{Y}| \cdot \frac{d_G(\mathcal{H}) \cdot \ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)}{\epsilon}\right) \text{ and } m_{\mathcal{H}}^{a,b}(\epsilon, \delta) = O\left(|\mathcal{Y}| \cdot \frac{d_G(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)}{\epsilon^2}\right).$$

**Proof** Since the claim is trivial if  $|\mathcal{Y}| = \infty$ , we can assume that  $k := |\mathcal{Y}| < \infty$ . Let  $A_{\text{full}}$  be a (full information) ERM learner for  $\mathcal{H}$ . Consider the following algorithm for the bandit setting: Given a sample  $(x_i, y_i)_{i=1}^m$ , for each  $i$  the algorithm guesses a label  $\hat{y}_i \in \mathcal{Y}$  drawn uniformly at random. Then the algorithm returns the hypothesis returned by  $A_{\text{full}}$  with the input sample which consists of the pairs  $(x_i, y_i)$  for which  $\hat{y}_i = y_i$ . We claim that  $m_{A_{\text{bandit}}}(\epsilon, \delta) \leq 3k \cdot m_{A_{\text{full}}}(\epsilon, \frac{\delta}{2})$  (for both the agnostic and the realizable case), so the theorem is implied by the bounds in the full information setting (theorem 7 and equation 5). Indeed, suppose that  $m$  examples suffice for  $A_{\text{full}}$  to return, with probability at least  $1 - \frac{\delta}{2}$  a hypothesis with regret at most  $\epsilon$ . Let  $(x_i, y_i)_{i=1}^{3km}$  be a sample for the bandit algorithm. By Chernoff bound, with probability at least  $1 - \frac{\delta}{2}$ , the sample  $A_{\text{bandit}}$  transfers to  $A_{\text{full}}$  consist of at least  $m$  examples. Note that the sample that  $A_{\text{full}}$  receives is an i.i.d. sample according to the same distribution from which the original sample was sampled. Thus, with probability at least  $1 - \frac{\delta}{2}$ ,  $A_{\text{full}}$  (and, consequently,  $A_{\text{bandit}}$ ) returns a hypothesis with regret at most  $\epsilon$ . ■

**The price of bandit information in the batch model:** Let  $\mathcal{H}$  be a hypotheses class. Define  $PBI_{\mathcal{H}}(\epsilon, \delta) = \frac{m_{\mathcal{H}}^{r,b}(\epsilon, \delta)}{m_{\mathcal{H}}^r(\epsilon, \delta)}$ . By Theorems 18,4 and Equation 3 we see that,  $PBI_{\mathcal{H}}(\epsilon, \delta) = O(\ln(|\mathcal{Y}|) \cdot \ln(\frac{1}{\epsilon}) \cdot |\mathcal{Y}|)$ . This is essentially tight since it is not hard to see that if both  $\mathcal{X}, \mathcal{Y}$  are finite and we let  $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$ , then  $PBI_{\mathcal{H}} = \Omega(|\mathcal{Y}|)$ .

Using Theorems 18,4 and Equations 5,3 we see that, as in the full information case, the finiteness of the Natarajan dimension is necessary and sufficient for learnability in the bandit setting as well. However, the ratio between the upper and the lower bounds is  $\Omega(\ln(|\mathcal{Y}|) \cdot |\mathcal{Y}|)$ . It would be interesting to find a more tight characterization of the sample complexity in the bandit setting. The Natarajan dimension (as well as the graph dimension and other known notions of dimension defined in (Ben-David et al., 1995), as they are all closely related to the Natarajan dimension) is deemed to fail for the following reason: For every  $k, d$ , there are classes  $\mathcal{H} \subseteq [k]^{[d]}$  of Natarajan dimension  $d$  where the realizable bandit sample complexity is  $O\left(\frac{d}{\epsilon} + \frac{\ln(\frac{1}{\delta})}{\epsilon}\right)$  (e.g. every class  $\mathcal{H}$  such that  $d_N(\mathcal{H}) = d$  and for every  $x \in [d]$ ,  $\#\{f(x) : f \in \mathcal{H}\} = 2$ ). On the other hand, the realizable bandit sample complexity of  $[k]^{[d]}$  is  $\Omega\left(k \cdot \left(\frac{d}{\epsilon} + \frac{\ln(\frac{1}{\delta})}{\epsilon}\right)\right)$ .

#### 4.2. Bandit vs Full Information in the Online Model

We now consider Bandits in the online learning model. We focus on the realizable case, in which the feedback provided to the learner is consistent with some function  $f_0 \in \mathcal{H}$ . We

define a new notion of dimension of a class, that determines the sample complexity in this setting. Let  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  be a hypothesis class and denote  $k = |\mathcal{Y}|$ . Consider a rooted tree  $T$  whose internal nodes are labeled by  $\mathcal{X}$  and such that the labels on edges from a parent to its child nodes are all different from each other. The tree  $T$  is *BL-shattered* by  $\mathcal{H}$  if, for every path from root to leaf  $x_1, \dots, x_k$ , there is a function  $f \in \mathcal{H}$  such that for every  $i$ ,  $f(x_i)$  is different from the label of  $(x_i, x_{i+1})$ . The **bandit Littlestone dimension** of  $\mathcal{H}$ , denoted  $\text{BL-dim}(\mathcal{H})$ , is the maximal depth of a complete  $k$ -ary tree that is BL-shattered by  $\mathcal{H}$ .

**Theorem 19** *Let  $\mathcal{H}$  be a hypothesis class with  $L = \text{BL-Dim}(\mathcal{H})$ . The sample complexity of every deterministic online learning algorithm for  $\mathcal{H}$  is at least  $L$ . Moreover, there is an online learning algorithm whose sample complexity is exactly  $L$ .*

**Proof** First, let  $T$  be a BL-shattered tree of depth  $L$ . We first show that for every deterministic learning algorithm there is a sequence  $x_1, \dots, x_L$  and a labeling function  $f_0 \in \mathcal{H}$  such that the algorithm makes  $L$  mistakes on this sequence. The sequence consists of the instances attached to nodes of  $T$ , when traversing the tree from the root to one of its leaves, such that the label of each edge  $(x_i, x_{i+1})$  is equal to the algorithm's prediction  $\hat{y}_i$ . The labeling function  $f_0 \in \mathcal{H}$  is one such that for all  $i$ ,  $f_0(x_i)$  is different from the label of edge  $(x_i, x_{i+1})$ . Such a function exists since  $T$  is BL-shattered.

Second, the following online learning algorithm makes at most  $L$  mistakes.

**Algorithm:** Bandit Standard Optimal Algorithm (BSOA)

Initialization:  $V_0 = \mathcal{H}$ .

For  $t = 1, 2, \dots$ ,

```

receive  $x_t$ 
for  $y \in \mathcal{Y}$ , let  $V_t^{(y)} = \{f \in V_{t-1} : f(x_t) \neq y\}$ 
predict  $\hat{y}_t \in \arg \min_y \text{BL-Dim}(V_t^{(y)})$ 
receive an indication whether  $\hat{y}_t = f(x_t)$ 
if the prediction is wrong, update  $V_t = V_t^{(\hat{y}_t)}$ 

```

To see that  $\mathcal{M}(\text{BSOA}) \leq L$ , note that at each time  $t$ , there is at least one  $V_t^{(y)}$  with  $\text{BL-Dim}(V_t^{(y)}) < \text{BL-Dim}(V_{t-1})$ . Thus, whenever the algorithm errs, the dimension of  $V_t$  decreases by one. Thus, after  $L$  mistakes, the dimension is 0, which means that there is a single function that is consistent with the sample, so no more mistakes can occur. ■

We conclude with an open question on the price of bandit information in the online model:

**Open question 20** *Let  $PBI(\mathcal{H}) = \frac{\text{BL-Dim}(\mathcal{H})}{\text{L-Dim}(\mathcal{H})}$  and fix  $k \geq 2$ . How large can  $PBI(\mathcal{H})$  be when  $\mathcal{H}$  is a class of functions from a domain  $\mathcal{X}$  to a range  $\mathcal{Y}$  of cardinality  $k$ ?*

## 5. The Sample Complexity of Known Multiclass Hypothesis Classes

In this section we analyze the sample complexity of two families of hypothesis classes for multiclass classification: the generalized linear construction (Duda and Hart, 1973; Vapnik,

1998; Hastie and Tibshirani, 1995; Freund and Schapire, 1997; Schapire and Singer, 1999; Collins, 2002; Taskar et al., 2003), and multiclass reduction trees (Beygelzimer et al., 2007, 2009; Fox, 1997). In particular, a special case of the generalized linear construction is the multi-vector construction (e.g. Crammer and Singer, 2003; Fink et al., 2006). We show that the sample complexity of the multi-vector construction and the reduction trees construction is similar and depends approximately linearly on the number of class labels. Due to the lack of space, proofs are omitted and can be found in the appendix.

### 5.1. The Generalized Linear Multiclass Construction

A generalized linear multiclass hypothesis class is defined with respect to a class specific feature mapping  $\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^t$ , for some integer  $t$ . For any such  $\phi$  define the hypothesis class  $\mathcal{M}_\phi^t = \{h[w] \mid w \in \mathbb{R}^t\}$ , where

$$h[w](x) = \operatorname{argmax}_{y \in \mathcal{Y}} \langle w, \phi(x, y) \rangle,$$

where we ignore tie-breaking issues w.l.o.g. . A popular special case is the linear construction used in multiclass SVM (Crammer and Singer, 2003) where  $\mathcal{X} = \mathbb{R}^d$ ,  $\mathcal{Y} = [k]$ ,  $t = dk$ , and  $\phi = \psi_{d,k}$ , defined by

$$\psi_{d,k}(x, i) \triangleq (0, \dots, 0, x[1], \dots, x[d], 0, \dots, 0),$$

where  $x[1]$  is in coordinate  $d(i - 1) + 1$ . We abbreviate  $\mathcal{L}_d^k \triangleq \mathcal{M}_{\psi_{d,k}}^{dk}$ . We first consider a general  $\phi$  and show that the sample complexity for any  $\phi$  is upper-bounded by a function of  $t$ .

**Theorem 21** *Let  $d_N$  be the Natarajan-dimension of  $\mathcal{M}_\phi^t$ . Then  $d_N \leq O(t \log(t))$ .*

For the linear construction a matching lower bound on the Natarajan dimension is shown in the following theorem. Thus, as one might expect, the sample complexity of learning with  $\mathcal{L}_k^d$  is of the order of  $dk$ .

**Theorem 22** *For  $d \geq 0$  and  $k \geq 2$ , let  $d_N$  be the Natarajan-dimension of  $\mathcal{L}_k^d$ . Then*

$$\Omega(dk) \leq d_N \leq O(dk \log(dk)).$$

### 5.2. Reduction trees

Reduction trees provide a way of constructing multiclass hypotheses from binary classifiers. A reduction tree consists of a tree structure, where each internal node is mapped to a binary classifier and each leaf is mapped to one of the multiclass labels. Classification of an example is done by traversing the tree, starting from the root and ending in one of the leaves, where in each node the result of the binary classifier determines whether to go left or right.

It has been shown that by using appropriate learning algorithms, one can guarantee a multiclass classification error of no more than  $\log_2(k)\epsilon$ , where  $k$  is the number of classes, and  $\epsilon$  is the average error of the binary classifiers (Fox, 1997; Beygelzimer et al., 2009). However,

this result does not directly provide sample complexity guarantees for these algorithms, since the value of  $\epsilon$  itself depends on the sample and on the learning algorithm.

In the following we analyze the sample complexity of any fixed reduction tree, under the assumption that the binary classifiers all belong to some fixed hypothesis class with a finite VC-dimension  $d$ . We provide bounds on the Natarajan dimension of the resulting multiclass hypothesis class, and show that it can be as large as  $\Omega(dk)$  for some hypothesis classes. We further analyze the special case where the binary hypothesis class is the class of linear separators in  $\mathbb{R}^d$ , and show that a similar result, though slightly weaker, holds for this class as well.

We now formally define a reduction tree and the hypothesis class related to it (see Figure 1 in the appendix for illustration). Let  $\mathcal{X}$  be the domain of examples and let  $[k]$  be the set of possible labels. A reduction tree is a full binary tree  $T$ . Denote the head node of  $T$  by  $H(T)$ . The sub-tree which is the left child of  $H(T)$  is denoted by  $L(T)$  and the sub-tree which is the right child of  $H(T)$  is denoted by  $R(T)$ . The set of internal nodes of  $T$  is denoted by  $N(T)$ , and the set of leaf nodes of  $T$  is denoted by  $\text{leaf}(T)$ . A multiclass classifier is a triplet  $[T, \lambda, C]$  where  $T$  is a reduction tree,  $\lambda$  is a one-to-one mapping  $\lambda[\cdot] : \text{leaf}(T) \rightarrow [k]$ , and  $C[\cdot] : N(T) \rightarrow \{0, 1\}^{\mathcal{X}}$  is a mapping from the internal nodes of  $T$  to binary classifiers on the domain  $\mathcal{X}$ .  $[T, \lambda, C] : \mathcal{X} \rightarrow [k]$  is defined recursively as follows:

$$[T, \lambda, C](x) = \begin{cases} [L(T), \lambda, C](x) & H(T) \notin \text{leaf}(T) \text{ and } C[H(T)](x) = 0, \\ [R(T), \lambda, C](x) & H(T) \notin \text{leaf}(T) \text{ and } C[H(T)](x) = 1, \\ \lambda[H(T)](x) & H(T) \in \text{leaf}(T). \end{cases}$$

Unless otherwise mentioned, we assume a fixed  $\lambda$ , and identify  $T$  with the pair  $(T, \lambda)$ . Accordingly,  $[T, \lambda, C]$  is abbreviated to  $[T, C]$ . Let  $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$  be a hypothesis class of binary classifiers on  $\mathcal{X}$ . The hypothesis class induced by  $\mathcal{H}$  on the tree  $T$  with label mapping  $\lambda$ , denoted by  $\mathcal{H}_{(T, \lambda)}$ , is the set of multiclass classifiers which can be generated on  $T$  using binary classifiers from  $\mathcal{H}$ . Formally,

$$\mathcal{H}_{(T, \lambda)} = \{[T, \lambda, C] \mid \forall n \in N(T), C[n] \in \mathcal{H}\}.$$

We abbreviate  $\mathcal{H}_{(T, \lambda)}$  to  $\mathcal{H}_T$  when the labeling  $\lambda$  is fixed.

Suppose that the VC-dimension of  $\mathcal{H}$  is  $d$ . What can be said about the sample complexity of  $\mathcal{H}_T$  for a given tree  $T$ ? First, a simple counting argument provides an upper bound on the graph-dimension and the Natarajan-dimension of  $\mathcal{H}_T$ : Any hypothesis in  $\mathcal{H}_T$  is a function of the values of  $|N(T)| = k - 1$  binary hypotheses from  $\mathcal{H}$ . Therefore, the number of possible labelings of  $A$  by  $\mathcal{H}_T$  for any  $A \subseteq \mathcal{X}$  is bounded by  $|\mathcal{H}|_A|^{k-1}$ . By Sauer's lemma,  $|\mathcal{H}|_A \leq |A|^d$ . Thus  $|\mathcal{H}_T|_A \leq |A|^{d(k-1)}$ . If  $A$  is G-shattered or N-shattered by  $\mathcal{H}_T$ , then  $|\mathcal{H}_T|_A \geq 2^{|A|}$ . Thus  $2^{|A|} \leq |A|^{d(k-1)}$ . It follows that  $|A| \leq O(dk \log(dk))$ , thus the same upper bound holds for the graph-dimension and the Natarajan-dimension. A closely matching lower bound is provided in the following theorem.

**Theorem 23** *Let  $k \geq 2$  and  $d \geq 2$  be integers. For any reduction tree  $T$  with  $k \geq 2$  leafs, there exists a binary hypothesis class  $\mathcal{H}$  with VC-dimension  $d$  such that  $\mathcal{H}_T$  has Natarajan dimension  $d(k - 1)$ .*

Theorem 23 shows that for every tree there exists a binary hypothesis class which induces a high sample complexity on the resulting multiclass hypothesis class. The following theorem shows that moreover, the popular hypothesis class of linear separators in  $\mathbb{R}^d$  induces reduction trees with a sample complexity which is almost as large, up to a logarithmic factor.

Let  $\mathcal{W}^d$  be the class of non-homogeneous linear separators in  $\mathbb{R}^d$ , that is  $\mathcal{W}^d = \{x \rightarrow \text{sign}(\langle x, w \rangle + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$ . For a full binary tree  $T$  with  $k$  leaves, denote by  $n_1(T)$  the number of internal nodes with one leaf child and one non-leaf child, and by  $n_2(T)$  the number of internal nodes with two leaf children.

**Theorem 24** *For any multiclass-to-binary tree  $T$  with  $k$  leaves, the graph dimension of  $\mathcal{W}_T^d$  is at least  $(d+1) \cdot n_2(T) + d \cdot n_1(T) \geq dk/2$ . Consequently the Natarajan dimension is  $\Omega(dk/\log(k))$ .*

We conclude that the sample complexity of different reduction trees is similar, and that this sample complexity is also similar to that of the multi-vector construction. This implies that when choosing between the different hypothesis classes, considerations other than the sample complexity should determine the choice. One such important consideration is the approximation error. Since sample complexity analysis bounds only the estimation error, one wishes to have the approximation error as low as possible. Thus if there is some prior knowledge on the match between the hypothesis class and the source distribution, this might guide the choice of the hypothesis class. The following theorem shows, however, that for fairly balanced reduction trees this match is highly dependent on the assignment of labels to leaf nodes. For any reduction tree  $T$  denote by  $\Lambda$  the set of one-to-one mappings from the leaf( $T$ ) to  $[k]$ , and let  $U$  be the uniform distribution over  $\Lambda$ .

**Theorem 25** *Let  $T$  be a full binary tree with  $k$  leaves, and let  $n$  be the number of leaves on the left sub-tree. For any hypothesis class  $\mathcal{H}$  with VC-dimension  $d$ , and for any distribution  $D$  over  $\mathcal{X} \times [k]$  which assigns non-zero probability to each label in  $[k]$ ,*

$$\Pr_{\lambda \sim U} [\mathcal{H}_{(T,\lambda)} \text{ separates } D] \leq \left(\frac{ek}{d}\right)^d \binom{k}{n}^{-1}.$$

*Thus if  $k \gg d$  and  $n$  is a constant fraction of  $k$ , this probability decreases exponentially with  $k$ .*

## 6. Conclusions and Open Problems

In this paper we have studied several new aspects of multiclass sample complexity. Many interesting questions arise and some are listed below.

Consider the two example classes from section 2.4. It is interesting to note that, in both cases,  $d_N(\mathcal{H}) = 1$ , and  $m_{\mathcal{H}}^r(\epsilon, \delta) = \Theta(\frac{1}{\epsilon} \ln(\frac{1}{\delta}))$ . It seems like the Natarajan dimension is the parameter that controls the sample complexity for those examples. That is also the case for symmetric classes as well as some other classes that we have examined but did not include in this paper. We therefore raise:

**Conjecture 26** *There exists a constant  $C$  such that, for every hypothesis class  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ ,*

$$m_{\mathcal{H}}^r(\epsilon, \delta) \leq C \left( \frac{d_N(\mathcal{H}) \ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon} \right)$$

In light of theorem 7 and the fact that there are cases where  $d_G \geq \log_2(|\mathcal{Y}| - 1)d_N$ , in order to prove the conjecture we will have to find a learning algorithm that is not just an *arbitrary* ERM learner. So far, all the general upper bounds that we are aware of are valid for *any* ERM learner. Understanding how to select among ERM learners is fundamental as it teaches us what is the correct way to learn. We suspect that such an understanding might lead to improved bounds in the binary case as well. We hope that our examples from section 2.4 and our result for symmetric classes will prove to be the first steps in the search for the best ERM.

Another direction is the study of learnability conditions for additional hypotheses classes. Section 5 shows that some well known multiclass constructions have surprisingly similar sample complexity properties. It is of practical significance and theoretical interest to study learnability conditions for other constructions, and especially to develop a fuller understanding of the relationship between different constructions, in a manner that could guide an informed choice of a hypothesis class.

## Acknowledgments

Sivan Sabato is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities.

## References

- P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.
- P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SICOMP: SIAM Journal on Computing*, 32, 2003.
- S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. Long. Characterizations of learnability for classes of  $\{0, \dots, n\}$ -valued functions. *Journal of Computer and System Sciences*, 50: 74–86, 1995.
- S. Ben-David, D. Pal, , and S. Shalev-Shwartz. Agnostic online learning. In *COLT*, 2009.
- A. Beygelzimer, J. Langford, and P. Ravikumar. Multiclass classification with filter trees. *Preprint, June*, 2007.
- Alina Beygelzimer, John Langford, and Pradeep Ravikumar. Error-correcting tournaments. *CoRR*, 2009.
- M. Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Conference on Empirical Methods in Natural Language Processing*, 2002.

- K. Crammer and Y. Singer. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991, 2003.
- R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.
- Michael Fink, Shai Shalev-Shwartz, Yoram Singer, and Shimon Ullman. Online multi-class learning by interclass hypothesis sharing. In *International Conference on Machine Learning*, 2006.
- J. Fox. *Applied Regression Analysis, Linear Models, and Related Methods*. SAGE Publications, 1997.
- Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.
- T.J. Hastie and R.J. Tibshirani. *Generalized additive models*. Chapman & Hall, 1995.
- S.M. Kakade, S. Shalev-Shwartz, and A. Tewari. Efficient bandit algorithms for online multiclass prediction. In *International Conference on Machine Learning*, 2008.
- N. Littlestone. Learning when irrelevant attributes abound. In *FOCS*, pages 68–77, October 1987.
- B. K. Natarajan. On learning sets and functions. *Mach. Learn.*, 4:67–97, 1989.
- R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):1–40, 1999.
- S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.
- B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *NIPS*, 2003.
- V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.
- V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

## Appendix A. Proofs Omitted from the Text

**Proof** (of theorem 4)

**The lower bound:** Let  $\mathcal{H} \subseteq \mathcal{Y}^X$  be a hypothesis class of Natarajan dimension  $d$  and Let  $\mathcal{H}_d := \{0, 1\}^{[d]}$ . We claim that  $m_{\mathcal{H}_d} \leq m_{\mathcal{H}}$ , so the lower bound is obtained by theorem 2. Let  $A$  be a learning algorithm for  $\mathcal{H}$ . Consider the learning algorithm,  $\bar{A}$ , for  $\mathcal{H}_d$  defined as follows. Let  $S = \{s_1, \dots, s_d\} \subseteq X$ ,  $f_0, f_1$  be a set and functions that indicate that  $d_N(\mathcal{H}) = d$ . Given a sample  $(x_i, y_i) \in [d] \times \{0, 1\}$ ,  $i = 1, \dots, m$ , let  $g = A((s_{x_i}, f_{y_i}(s_{x_i})))_{i=1}^m$ . Define  $f = \bar{A}((x_i, y_i)_{i=1}^m)$  by setting  $f(i) = 1$  if and only if  $g(s_i) = f_1(s_i)$ . It is not hard to see that  $m_{\bar{A}} \leq m_A^r$ , thus,  $m_{\mathcal{H}_d} \leq m_{\mathcal{H}}$ .

**The upper bound:** Let  $\mathcal{H} \subseteq \mathcal{Y}^X$  be a hypothesis class of graph dimension  $d$ . For every

$f \in \mathcal{H}$  define  $\bar{f} : \mathcal{X} \times \mathcal{Y} \rightarrow \{0, 1\}$  by setting  $\bar{f}(x, y) = 1$  if and only if  $f(x) = y$  and let  $\bar{\mathcal{H}} = \{\bar{f} : f \in \mathcal{H}\}$ . It is not hard to see that  $VC(\bar{\mathcal{H}}) = d_G(\mathcal{H})$ .

Suppose that  $f \in \mathcal{H}$  is consistent with a sample  $(x_i, f_0(x_i))_{i=1}^m$  of  $m = \Omega(\frac{d}{\epsilon} \ln(\frac{1}{\epsilon}) + \frac{1}{\epsilon} \ln(\frac{1}{\delta}))$  examples, drawn i.i.d. according to some distribution  $\mathcal{D}$  on  $\mathcal{X}$ . We must show that, with probability  $\geq 1 - \delta$ ,  $\text{Err}_{\mathcal{D}, f_0}(f) \leq \epsilon$ . However, by theorem 2,

$$\text{Err}_{\mathcal{D}, f_0}(f) = \Pr_{x \sim \mathcal{D}} (\bar{f}(x, f_0(x)) \neq 1) \leq \epsilon$$

With probability  $\geq 1 - \delta$ . ■

**Proof** (of Theorem 6) Let  $A$  be an ERM learner. Since  $F_A(f) \subseteq \mathcal{H}$  for every  $f$ , it follows that  $\Pi_A \leq \Pi_{\mathcal{H}}$ . By lemma 11,  $\Pi_{\mathcal{H}}(m) \leq m^{d_N(\mathcal{H})} |\mathcal{Y}|^{2d_N(\mathcal{H})}$ . Incorporating it into Theorem 9 we get the desired bound. ■

**Proof** (of Theorem 21) Let  $S = \{x_1, \dots, x_{d_N}\} \subseteq \mathbb{R}^d$  be a set which is N-shattered by  $\mathcal{M}_\phi^t$ , and let  $f_1, f_2 : S \rightarrow \mathcal{Y}$  be the functions that witness the shattering. For every  $i \in [d_N]$  let  $z_i = \phi(x_i, f_1(x_i)) - \phi(x_i, f_2(x_i)) \in \mathbb{R}^t$ . Denote  $Z = \{z_i\}_{i \in [d_N]}$ . Consider the hypothesis class of homogeneous linear separators in  $\mathbb{R}^t$ , defined by  $\{z \rightarrow \text{sign}(\langle w, z \rangle) \mid w \in \mathbb{R}^t\}$ . Since the VC-dimension of this class is  $t$ , by Sauer's lemma the number of possible labelings of  $Z$  with this class is upper-bounded by  $(d_N)^t$ . We now show that there is a one-to-one mapping from subsets  $T \subseteq S$  to labelings of  $Z$ : For any  $T \subseteq S$ , let  $w \in \mathbb{R}^t$  such that

$$\{x \in S \mid h[w](x) = f_1(x)\} = T, \text{ and } \{x \in S \mid h[w](x) = f_2(x)\} = S \setminus T.$$

Then  $T = \{x \in S \mid \langle w, \phi(x, f_1(x)) \rangle \geq \langle w, \phi(x, f_2(x)) \rangle\} = \{x_i \mid \langle w, z_i \rangle \geq 0\}$ . Thus every  $T$  induces a different labeling of  $Z$ . It follows that the number of subsets of  $S$  is bounded by the number of labelings of  $Z$ , thus  $2^{d_N} \leq (d_N)^t$ . It follows that  $d_N \leq O(t \log(t))$ . ■

**Proof** (of Theorem 22) The upper bound is a direct consequence of Theorem 21. For the lower bound, we show that there exists an N-shattered set of size  $\lfloor d/2 \rfloor \cdot \lfloor k/2 \rfloor$ . Let  $b = \lfloor k/2 \rfloor$ . Let  $x_1, \dots, x_b \in \mathbb{R}^2$  be  $b$  different vectors such that  $\forall i \in [b], \|x_i\| = 1$ . Let  $S = \{y_{i,j}\}_{i \in [b], j \in [\lfloor d/2 \rfloor]} \subseteq \mathbb{R}^d$ , where for  $s \in [d]$ :

$$y_{i,j}[s] = \begin{cases} x_i[1] & s = 2j - 1 \\ x_i[2] & s = 2j \\ 0 & \text{otherwise.} \end{cases}$$

We show that  $S$  is N-shattered, thus  $d_N \geq |S| = \lfloor k/2 \rfloor \cdot \lfloor d/2 \rfloor$ . Define functions  $f_1, f_2 : S \rightarrow [k]$  such that for  $y_{i,j} \in S$ ,  $f_1(y_{i,j}) = i$  and  $f_2(y_{i,j}) = b + i$ . For a subset  $T \subseteq S$ , let  $w \in \mathbb{R}^{dk}$  such that for  $i \in [b], s \in [d]$

$$w[d(i-1) + s] = \begin{cases} x_i[1] & y_{i,j} \in Z \text{ and } s = 2j - 1, \\ x_i[2] & y_{i,j} \in Z \text{ and } s = 2j, \\ 0 & \text{otherwise.} \end{cases}$$

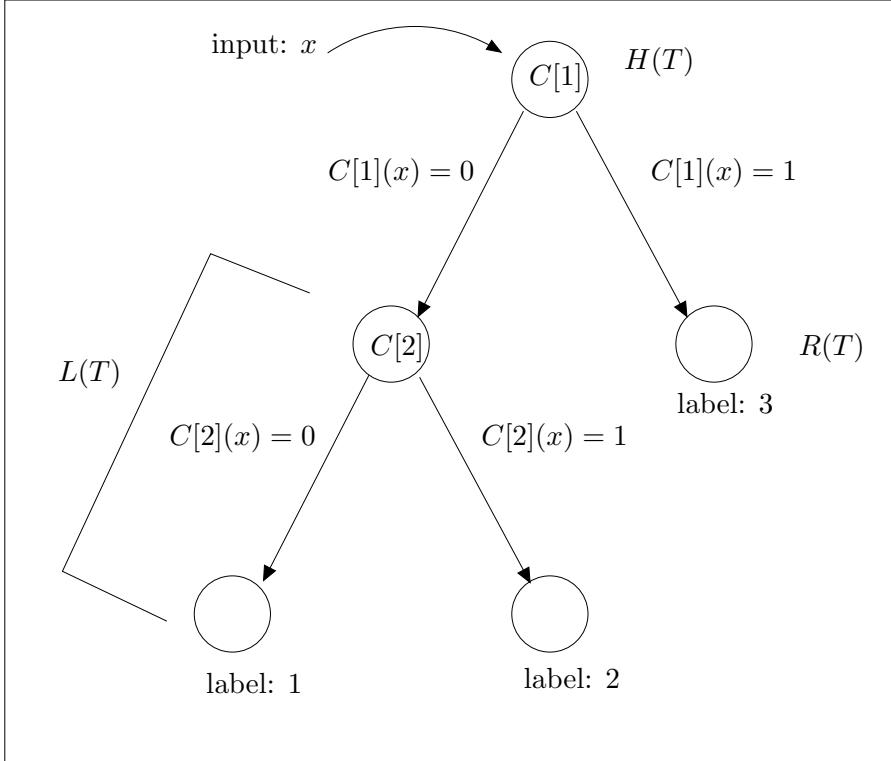


Figure 1: Illustration of a reduction tree

and for  $i \in \{b+1, \dots, 2b\}, s \in [d]$ ,

$$w[d(i-1)+s] = \begin{cases} x_i[1] & y_{i-b,j} \notin Z \text{ and } s = 2j-1, \\ x_i[2] & y_{i-b,j} \notin Z \text{ and } s = 2j, \\ 0 & \text{otherwise.} \end{cases}$$

Then  $h[w] = f_1(y)$  for  $y \in T$  and  $h[w] = f_2(y)$  for  $y \in S \setminus T$ . Thus  $S$  is N-shattered. ■

**Proof** (of Theorem 23) Let  $\mathcal{H}(T)$  be a binary hypothesis class for tree  $T$ . We construct  $\mathcal{H}(T)$  inductively on the structure of the tree. For every tree  $T$ , the domain of the binary hypotheses in  $\mathcal{H}(T)$  will be  $[d] \times N(T)$ .

**Induction basis:** Assume that both  $L(T)$  and  $R(T)$  are leafs, thus  $k = 2$  and  $|N(T)| = 1$ . Define  $\mathcal{H}(T) = \{h \mid h : [d] \times \{H(T)\} \rightarrow \{0, 1\}\}$ .

**Inductive step:** Assume  $T$  has two children  $L(T)$  and  $R(T)$ , and at least one of them is not a leaf. By the induction hypothesis, if  $L(T)$  is a non-leaf then  $\mathcal{H}(L(T))$  is a set of binary hypotheses with domain  $[d] \times N(L(T))$ .  $\mathcal{H}(L(T))$  has VC-dimension  $d$ , and the Natarajan dimension of  $\mathcal{H}(L(T))_{L(T)}$  is  $d \cdot |N(L(T))|$ . The same holds for  $R(T)$ . Define  $\mathcal{H}(T) = \{h_0, h_1\} \cup \mathcal{H}_L \cup \mathcal{H}_R \cup \mathcal{H}_H$ , where:

- $h_0(x) = 0$  and  $h_1(x) = 1$  for all  $x \in [d] \times N(T)$ ,

- If  $L(T)$  is a leaf,  $\mathcal{H}_L = \emptyset$ . Otherwise,

$$\mathcal{H}_L = \left\{ h : [d] \times N(T) \rightarrow \{0, 1\} \mid \exists h_L \in \mathcal{H}(L(T)), \forall x \in [d] \times N(T), \right.$$

$$h(x) = \begin{cases} h_L(x) & x \in [d] \times N(L(T)), \\ 0 & \text{otherwise.} \end{cases} \quad \left. \right\}$$

- $\mathcal{H}_R$  is defined similarly, for  $R(T)$  instead of  $L(T)$ .
- $\mathcal{H}_H$  is defined as follows:

$$\mathcal{H}_H = \{ h : [d] \times N(T) \rightarrow \{0, 1\} \mid \forall x \in [d] \times N(L(T)), h(x) = 0, \forall x \in [d] \times N(R(T)), h(x) = 1 \}.$$

We now prove by induction that for every tree  $T$  the following claims hold:

- $\mathcal{H}(T)$  has VC-dimension  $d$ ,
- $\mathcal{H}(T)_T$  has Natarajan dimension  $d \cdot |N(T)|$ .
- An auxiliary claim:  $\mathcal{H}(T)$  includes the hypotheses  $h_0$  and  $h_1$ .

**Induction Basis:** If both  $L(T)$  and  $R(T)$  are leaves, then the VC-dimension of  $\mathcal{H}(T)$  is clearly  $d$ . The induced multiclass hypothesis class  $\mathcal{H}(T)_T$  is in fact a set of binary hypotheses which is isomorphic to  $\mathcal{H}(T)$ , thus its Natarajan dimension is also  $d = d(k - 1)$ . The zero hypothesis is clearly in  $\mathcal{H}(T)$  by construction.

**Induction Step:** Assume  $T$  has two children  $L(T)$  and  $R(T)$ , and at least one of them is not a leaf. By the construction of  $\mathcal{H}(T)$ , the auxiliary claim clearly holds. The following lemmas, whose proofs follows, prove the two other claims:

**Lemma 27**  $\mathcal{H}(T)$  has VC-dimension  $d$ .

**Lemma 28**  $\mathcal{H}(T)_T$  has Natarajan dimension  $d|N(T)|$ .

Thus the induction hypothesis holds. ■

**Proof** (of Lemma 27) The VC-dimension of  $\mathcal{H}(T)$  is at least  $d$ , since the VC-dimension of at least one of  $\mathcal{H}(L(T))$  and  $\mathcal{H}(R(T))$  is  $d$ . Assume to the contrary that it is larger than  $d$ , then there exists a set  $A = \{x_1, \dots, x_{d+1}\} \subseteq [d] \times N(T)$  which is shattered by  $\mathcal{H}(T)$ . Denote for brevity  $S_L = [d] \times N(L(T))$ ,  $S_R = [d] \times N(R(T))$  and  $S_H = [d] \times H(T)$ . By the construction of  $\mathcal{H}(T)$  and the auxiliary claim,  $\mathcal{H}(T)|_{S_L} = \mathcal{H}(L(T))$  and  $\mathcal{H}(T)|_{S_R} = \mathcal{H}(R(T))$  whenever  $L(T)$  and  $R(T)$  are not leaves respectively. In addition, since  $|S_H| = d$ ,  $A \not\subseteq S_H$ . Since  $|A| \geq 3$ , there exist three different elements in  $x, y, z \in A$  such that at least two of them are in different sets out of  $S_L, S_H, S_R$ . We consider the different cases (where names of elements are w.l.o.g.) and show for each case a labeling  $l_x, l_y, l_z$  for  $x, y, z$  that cannot be achieved with a hypothesis in  $\mathcal{H}(T)$ :

- If  $x \in S_H$ ,  $y \in S_R$  then  $l_x = 1, l_y = 0$  cannot be achieved.

- If  $x, y \in S_L, z \in S_H \cup S_R$  then  $l_x = 1, l_y = 0, l_z = 1$  cannot be achieved.
- If  $x \in S_L, y, z \in S_R$  then  $l_x = 1, l_y = 0, l_z = 1$  cannot be achieved.
- If  $x \in S_L, y, z \in S_H$  then  $l_x = 1, l_y = 0, l_z = 1$  cannot be achieved.

We have reached a contradiction, therefore no such  $A$  exists.  $\blacksquare$

**Proof** (of Lemma 28) The Natarajan dimension is upper bounded by the size of the domain, which is  $d|N(T)|$ . By the induction hypothesis,  $\mathcal{H}(L(T))_{L(T)}$  and  $\mathcal{H}(R(T))_{R(T)}$  have Natarajan dimension  $d_L = d|L(T)|$  and  $d_R = d|R(T)|$  respectively. Thus  $[d] \times N(L(T))$  and  $[d] \times N(R(T))$  are N-shattered by  $\mathcal{H}(L(T))_{L(T)}$  and  $\mathcal{H}(R(T))_{R(T)}$  respectively. Let  $f_1^L, f_2^L$ , and  $f_1^R, f_2^R$  be the pairs of functions that witness the N-shattering of  $\mathcal{H}(L(T))_{L(T)}$  and  $\mathcal{H}(R(T))_{R(T)}$  respectively. Let  $c_L$  be the class of the left-most child in  $L(T)$ , and let  $c_R$  be the class of the left-most child in  $R(T)$ . Define  $g_1$  and  $g_2$  as follows:

$$g_1(x) = \begin{cases} f_1^L(x) & x \in [d] \times N(L(T)) \\ f_1^R(x) & x \in [d] \times N(R(T)) \\ c_L & x \in [d] \times \{H(T)\} \end{cases}$$

$$g_2(x) = \begin{cases} f_2^L(x) & x \in [d] \times N(L(T)) \\ f_2^R(x) & x \in [d] \times N(R(T)) \\ c_R & x \in [d] \times \{H(T)\} \end{cases}$$

It is easy to verify that  $[d] \times N(T)$  is N-shattered using  $g_1$  and  $g_2$ .  $\blacksquare$

**Proof** (of Theorem 24) The proof is by induction on the structure of the tree.

**Induction basis:** Assume that  $T$  is a tree with one internal node and two leaf children. Then  $\mathcal{W}_T^d$  is isomorphic up to label names to  $\mathcal{W}^d$ . Thus the graph dimension of  $\mathcal{W}_T^d$  is equal to the VC-dimension of  $\mathcal{W}^d$ , that is  $d + 1 = (d + 1) \cdot n_1(T)$ .

**Inductive step:** We consider two cases: Either both  $R(T)$  and  $L(T)$  are non-leaves or one is a leaf and one is not.

**Case 1:** Let  $T$  be a tree where both  $L(T)$  and  $R(T)$  are non-leaves. By the induction hypothesis, the graph dimension of  $\mathcal{W}_{L(T)}^d$  is at least  $d_L = (d + 1) \cdot n_2(L(T)) + d \cdot n_1(L(T))$  and the graph dimension of  $\mathcal{W}_{R(T)}^d$  is at least  $d_R = (d + 1) \cdot n_2(R(T)) + d \cdot n_1(R(T))$ . Thus there exist sets  $A_L = \{a_1, \dots, a_{d_L}\}$  and  $B_R = \{b_1, \dots, b_{d_R}\}$  which are G-shattered by  $L(T)$  and  $R(T)$  respectively, using functions  $f_L$  and  $f_R$  respectively. Let

$$a_L = (\min_{i \in [d_L]} \{a_i[1]\} + 1, 0, \dots, 0) \in \mathbb{R}^d$$

$$b_R = (-\max_{i \in [d_R]} \{b_i[1]\} - 1, 0, \dots, 0) \in \mathbb{R}^d$$

Let  $\tilde{A}_L = \{a_1 + a_L, \dots, a_{d_L} + a_L\}$  and let  $\tilde{B}_R = \{b_1 + b_R, \dots, b_{d_R} + b_R\}$ . Then  $\forall x \in \tilde{A}_L, x[1] > 0$ , and  $\forall x \in \tilde{B}_R, x[1] < 0$ .

We show that the set  $\tilde{A}_L \cup \tilde{B}_R$  is G-shattered by  $\mathcal{W}_T^d$ : Define

$$f(x) = \begin{cases} f_R(x) & x[1] > 0 \\ f_L(x) & \text{otherwise.} \end{cases}$$

Let  $Z \subseteq \tilde{A}_L \cup \tilde{B}_R$ . We construct a mapping  $C : N(T) \rightarrow \mathcal{H}$  such that

$$\{x \in \tilde{A}_L \cup \tilde{B}_R \mid [T, C](x) = f(x)\} = Z.$$

Let  $Y \subseteq A_L \cap B_R = \{a_i \mid a_i + a_L \in Z\} \cap \{b_i \mid b_i + b_R \in Z\}$ . Since  $A_L$  and  $B_R$  are G-shattered with  $f_L$  and  $f_R$ , there exist mappings  $C_L : N(L(T)) \rightarrow \mathcal{W}^d$  and  $C_R : N(R(T)) \rightarrow \mathcal{W}^d$  such that

$$\begin{aligned} \{x \in A_L \mid [L(T), C_L](x) = f_L(x)\} &= Y \cap A_L, \\ \{x \in B_R \mid [R(T), C_R](x) = f_R(x)\} &= Y \cap B_R. \end{aligned}$$

Define the mapping  $C$  as a translation of the mappings  $C_L$  and  $C_R$ , defined by:

$$\begin{aligned} \forall n \in L(T), C_L[n] = (w, b) \Rightarrow C[n] &= (w, b - \langle w, a_L \rangle), \\ \forall n \in R(T), C_R[n] = (w, b) \Rightarrow C[n] &= (w, b - \langle w, b_R \rangle). \end{aligned}$$

Then

$$\begin{aligned} \{x \in \tilde{A}_L \mid [L(T), C](x) = f_L(x)\} &= Z \cap \tilde{A}_L, \\ \{x \in \tilde{B}_R \mid [R(T), C](x) = f_R(x)\} &= Z \cap \tilde{B}_R. \end{aligned}$$

Now, set  $C[H(T)](x) = \text{sign}(\langle x, w \rangle + b)$  where  $w = (1, 0, \dots, 0)$  and  $b = 0$ . Then

$$\begin{aligned} \forall x \in \tilde{A}_L, [T, C](x) &= [L(T), C](x) = f_L(x) = f(x), \\ \forall x \in \tilde{B}_R, [T, C](x) &= [R(T), C](x) = f_R(x) = f(x). \end{aligned}$$

Thus  $\tilde{A}_L \cup \tilde{B}_R$  is G-shattered by  $\mathcal{W}_T^d$ . It follows that the graph dimension of  $\mathcal{W}_T^d$  is at least  $|\tilde{A}_L \cup \tilde{B}_R| = d_L + d_R = (d+1) \cdot n_2(T) + d \cdot n_1(T)$ .

**Case 2:** Assume w.l.o.g. that  $T$  is a tree where  $L(T)$  is not a leaf node and  $R(T)$  is a leaf node with  $\lambda[R(T)] = t$ . By the induction hypothesis, the graph dimension of  $\mathcal{W}_{L(T)}^d$  is at least  $d_L = (d+1) \cdot n_2(L(T)) + d \cdot n_1(L(T))$ . Thus there exists a set  $A = \{a_1, \dots, a_{d_L}\}$  which is G-shattered by  $L(T)$  using the function  $f_L$ .

Denote by  $e_i$  the  $i$ 'th unit vector in  $\mathbb{R}^d$ , and let  $q > 0$  be large enough such that  $\{(0, \dots, 0), qe_1, \dots, qe_d\}$  is shattered with a margin of  $2M$ , where  $M = \max_{x \in A} \|x\|_2$ . Let  $B = A \cup \{qe_1, \dots, qe_d\}$ . Then we show  $B$  is G-shattered using the following function  $f$ :

$$f(x) = \begin{cases} f_L & \|x\| \leq q \\ t & \text{otherwise.} \end{cases}$$

Let  $Z \subseteq B$ . We construct a mapping  $C : N(T) \rightarrow \mathcal{H}$  such that

$$\{x \in B \mid [T, C](x) = f(x)\} = Z. \quad (8)$$

Since  $A$  is G-shattered using  $f_L$ , there exists a mapping  $C_L : N(L(T)) \rightarrow \mathcal{W}^d$  such that  $\{x \in A \mid [L(T), C_L](x) = f_L(x)\} = Z \cap A$ . Define  $C$  such that  $\forall n \in N(L(T)), C[n] = C_L[n]$ . In addition, Let  $C[H(T)] \in \mathcal{W}^d$  be a hypothesis such that  $\forall i, e_i \in Z \iff h(e_i) = 1$ , and  $\forall x, \|x\|_2 \leq M \rightarrow h(0) = 0$ . Then Equation. (8) holds. Thus the graph dimension of  $\mathcal{W}_T^d$  is at least  $|B| = d_L + d \geq (d+1) \cdot n_2(T) + d \cdot n_1(T)$ . ■

**Proof** (of Theorem 25) It suffices to consider distributions with deterministic labeling, such that the correct label is a function  $f : \mathcal{X} \rightarrow [k]$ . Let  $A = \{x_1, \dots, x_k\} \in \mathcal{X}$  such that for all  $i \in [k], f(x_i) = i$ . For any labeling  $\lambda \in \Lambda$ , let  $f_\lambda : A \rightarrow \{0, 1\}$  be the indicator function of the set of labels assigned to leaves in  $L(T)$ , that is  $f_\lambda(x_i) = \mathbf{1}[\exists n \in \text{leaf}(L(T)), \lambda[n] = i]$ . If  $D$  is separable with  $\mathcal{H}_{(T, \lambda)}$  then  $f_\lambda = C[H((T, \lambda))]|_A \in \mathcal{H}|_A$ . By Sauer's lemma,  $|\mathcal{H}|_A \leq (\frac{ek}{d})^d$ . There are  $\binom{k}{n}$  possible indicator functions  $f_\lambda$  for a labeling  $\lambda$ , and they all have equal probability for  $\lambda \sim U$ . Thus  $\mathbb{P}_{\lambda \sim U}[f_\lambda \in \mathcal{H}|_A] \leq (\frac{ek}{d})^d / \binom{k}{n}$ . ■

DANIELY SABATO BEN-DAVID SHALEV-SHWARTZ

# Mixability is Bayes Risk Curvature Relative to Log Loss

**Tim van Erven**

CWI

Amsterdam, The Netherlands

TIM.VAN.ERVEN@CWI.NL

**Mark D. Reid**

**Robert C. Williamson**

ANU and NICTA

Canberra, Australia

MARK.REID@ANU.EDU.AU

BOB.WILLIAMSON@ANU.EDU.AU

**Editors:** Sham Kakade, Ulrike von Luxburg

## Abstract

Mixability of a loss governs the best possible performance when aggregating expert predictions with respect to that loss. The determination of the mixability constant for binary losses is straightforward but opaque. In the binary case we make this transparent and simpler by characterising mixability in terms of the second derivative of the Bayes risk of proper losses. We then extend this result to multiclass proper losses where there are few existing results. We show that mixability is governed by the Hessian of the Bayes risk, relative to the Hessian of the Bayes risk for log loss. We conclude by comparing our result to other work that bounds prediction performance in terms of the geometry of the Bayes risk. Although all calculations are for proper losses, we also show how to carry the results across to improper losses.

**Keywords:** mixability, regret bounds, probability estimation, proper losses

## 1. Introduction

Mixability is an important property of a loss function that governs the performance of an aggregating forecaster in the prediction with experts setting. The notion is due to Vovk (1990, 1995). Extensions to mixability were presented by Kalnishkan and Vyugin (2002b). The motivation for studying mixability is summarised below (this summary is based on the presentation of Kalnishkan and Vyugin (2008)<sup>1</sup>).

Let  $n \in \mathbb{N}$  and  $\mathcal{Y} = \{1, \dots, n\}$  be the outcome space. We will consider a prediction game where the loss of the learner making predictions  $v_1, v_2, \dots \in \mathcal{V}$  is measured by a loss function  $\ell: \mathcal{Y} \times \mathcal{V} \rightarrow \mathbb{R}_+$  cumulatively: for  $T \in \mathbb{N}$ ,  $\text{Loss}(T) := \sum_{t=1}^T \ell(y_t, v_t)$ , where  $y_1, y_2, \dots \in \mathcal{Y}$  are outcomes. The learner has access to predictions  $v_t^i$ ,  $t = 1, 2, \dots, i \in \{1, \dots, N\}$  generated by  $N$  experts  $\mathcal{E}_1, \dots, \mathcal{E}_N$  that attempt to predict the same sequence. The goal of the learner is to predict nearly as well as the best expert. A *merging strategy*  $\mathcal{M}: \bigcup_{t=1}^{\infty} (\mathcal{Y}^{t-1} \times (\mathcal{V}^N)^t) \rightarrow \mathcal{V}$  takes the outcomes  $y_1, \dots, y_{t-1}$  and predictions  $v_s^i$ ,  $i = 1, \dots, N$  for times  $s = 1, \dots, t$  and outputs an aggregated prediction  $v_t^{\mathcal{M}}$ , incurring loss  $\ell(y_t, v_t^{\mathcal{M}})$  when  $y_t$  is revealed. After  $T$  rounds, the loss of  $\mathcal{M}$  is  $\text{Loss}_{\mathcal{M}}(T) = \sum_{t=1}^T \ell(y_t, v_t^{\mathcal{M}})$ . The loss of expert  $\mathcal{E}_i$  is  $\text{Loss}_{\mathcal{E}_i}(T) = \sum_{t=1}^T \ell(y_t, v_t^i)$ . When  $\mathcal{M}$  is the aggregating algorithm (which can be used for all losses

---

1. Kalnishkan and Vyugin (2008) denote mixability by  $\bar{\beta} \in (0, 1)$ ; we use  $\beta = -\ln \bar{\beta} \in (0, \infty)$ .

considered in this paper) (Vovk, 1995),  $\beta$ -mixability (see Section 3 for the definition) implies for all  $t \in \mathbb{N}$ , all  $i \in \{1, \dots, N\}$ ,

$$\text{Loss}_{\mathcal{M}}(t) \leq \text{Loss}_{\mathcal{E}_i}(t) + \frac{\ln N}{\beta}. \quad (1)$$

Conversely, if for every  $\beta \in \mathbb{R}_+$  the loss function  $\ell$  is not  $\beta$ -mixable, then it is not possible to predict as well as the best expert up to an additive constant using any merging strategy.

Thus determining  $\beta_\ell$  (the largest  $\beta$  such that  $\ell$  is  $\beta$ -mixable) is equivalent to precisely bounding the prediction error of the aggregating algorithm. The mixability of several binary losses and the Brier score in the multiclass case (Vovk and Zhdanov, 2009) is known. However a general characterisation of  $\beta_\ell$  in terms of other key properties of the loss has been missing. The present paper shows how  $\beta_\ell$  depends upon the curvature of the conditional Bayes risk for  $\ell$  when  $\ell$  is a strictly proper continuously differentiable multiclass loss (see Theorem 10).

We use the following notation throughout. Let  $[n] := \{1, \dots, n\}$  and denote by  $\mathbb{R}_+$  the non-negative reals. The transpose of a vector  $x$  is  $x'$ . If  $x$  is a  $n$ -vector,  $A = \text{diag}(x)$  is the  $n \times n$  matrix with entries  $A_{i,i} = x_i$ ,  $i \in [n]$  and  $A_{i,j} = 0$  for  $i \neq j$ . We also write  $\text{diag}(x_i)_{i=1}^n := \text{diag}(x_1, \dots, x_n) := \text{diag}((x_1, \dots, x_n)')$ . The inner product of two  $n$ -vectors  $x$  and  $y$  is denoted by matrix product  $x'y$ . We sometimes write  $A \cdot B$  for the matrix product  $AB$  for clarity when required. If  $A - B$  is positive definite (resp. semidefinite), then we write  $A \succ B$  (resp.  $A \succcurlyeq B$ ). The  $n$ -simplex  $\Delta^n := \{(x_1, \dots, x_n)' \in \mathbb{R}^n : x_i \geq 0, i \in [n], \sum_{i=1}^n x_i = 1\}$ . Other notation (the Kronecker product  $\otimes$ , the derivative  $D$ , and the Hessian  $H$ ) are defined in Appendix A which also includes several matrix calculus results we use.

## 2. Proper Multiclass Losses

We consider multiclass losses for class probability estimation. A *loss function*  $\ell : \Delta^n \rightarrow \mathbb{R}_+^n$  assigns a loss vector  $\ell(q) = (\ell_1(q), \dots, \ell_n(q))$  to each distribution  $q \in \Delta^n$  where  $\ell_i(q)$  ( $= \ell(i, q)$  traditionally) is the penalty for predicting  $q$  when outcome  $i \in [n]$  occurs<sup>2</sup>. If the outcomes are distributed with probability  $p \in \Delta^n$  then the *risk* for predicting  $q$  is just the expected loss

$$L(p, q) := p' \ell(q) = \sum_{i=1}^n p_i \ell_i(q).$$

The *Bayes risk* for  $p$  is the minimal achievable risk for that outcome distribution,

$$\underline{L}(p) := \inf_{q \in \Delta^n} L(p, q).$$

We say that a loss is *proper* whenever the minimal risk is always achieved by predicting the true outcome distribution, that is,  $\underline{L}(p) = L(p, p)$  for all  $p \in \Delta^n$ . We say a proper loss is *strictly proper* if there exists no  $q \neq p$  such that  $L(p, q) = \underline{L}(p)$ . The log loss  $\ell_{\log}(p) := (-\ln(p_1), \dots, -\ln(p_n))'$  is strictly proper. Its corresponding Bayes risk is  $\underline{L}_{\log}(p) = -\sum_{i=1}^n p_i \ln(p_i)$ .

---

2. Technically, we should allow  $\ell(p) = \infty$  to allow for log loss. However, we are only concerned with the behaviour of  $\ell$  in the relative interior of  $\Delta^n$  and use  $\Delta^n$  in that sense.

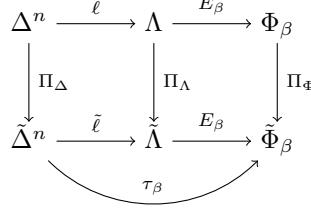


Figure 1: Mappings and spaces.

Proper losses are defined only on  $\Delta^n$  which is a  $(n - 1)$ -dimensional submanifold of  $\mathbb{R}_+^n$ . In order to define the derivatives we will need, it is necessary to project down onto  $n - 1$  dimensions. Let  $\Pi_\Delta : \Delta^n \rightarrow \tilde{\Delta}^n$  denote the projection of the  $n$ -simplex  $\Delta^n$  onto its “bottom”, denoted  $\tilde{\Delta}^n$ . That is,

$$\Pi_\Delta(p) := (p_1, \dots, p_{n-1}) =: \tilde{p} \in \tilde{\Delta}^n$$

is the projection of  $p$  onto its first  $n - 1$  coordinates. Similarly, we will project  $\ell$ 's image  $\Lambda := \ell(\Delta^n)$  using  $\Pi_\Lambda(\lambda) := (\lambda_1, \dots, \lambda_{n-1})$  for  $\lambda \in \Lambda$  with range denoted  $\tilde{\Lambda}$ . Since  $p_n = p_n(\tilde{p}) := 1 - \sum_{i=1}^{n-1} \tilde{p}_i$  we see that  $\Pi_\Delta$  is invertible. Specifically,  $\Pi_\Delta^{-1}(\tilde{p}) = (\tilde{p}_1, \dots, \tilde{p}_{n-1}, p_n(\tilde{p}))$ . Thus, any function of  $p$  can be expressed as a function of  $\tilde{p}$ . In particular, given a loss  $\ell : \Delta^n \rightarrow \mathbb{R}^n$  we can write  $\ell(\tilde{p}) = \ell(\Pi_\Delta^{-1}(\tilde{p}))$  for  $\tilde{p} \in \tilde{\Delta}^n$  and use  $\tilde{\ell}(\tilde{p}) := \Pi_\Lambda(\ell(\tilde{p}))$  to denote its projection onto its first  $n - 1$  coordinates (see Figure 1).

As it is central to our results, we assume all losses are proper and suitably continuously differentiable for the remainder of the paper. We will additionally assume strict properness whenever we require the Hessian of the Bayes risk to be invertible (see Lemma 5).

**Lemma 1** *A continuously differentiable (strictly) proper loss  $\ell$  has (strictly) concave Bayes risk  $\underline{L}$  and a risk  $L$  that satisfies the stationarity condition: for each  $p$  in the relative interior of  $\Delta^n$  we have*

$$p'D\ell(\tilde{p}) = 0_{n-1}. \quad (2)$$

*Furthermore,  $\ell$  and  $\Pi_\Lambda$  are invertible and for all  $p \in \Delta^n$ , the vector  $p$  is normal to the surface  $\Lambda = \ell(\Delta^n)$  at  $\ell(p)$ .*

**Proof** The Bayes risk  $\underline{L}(p)$  is the infimum of a set of linear functions  $p \mapsto p'\ell(q)$  and thus concave. Each linear function is tangent to  $\ell(\Delta^n)$  at a single point when  $\ell$  is strictly proper and so  $\underline{L}$  is strictly concave. Properness guarantees that for all  $p, q \in \Delta^n$  we have  $p'\ell(p) \leq p'\ell(q)$  so the function  $L_p : q \mapsto p'\ell(q)$  has a minima at  $p = q$ . Hence the function  $\tilde{L}_p : \tilde{q} \mapsto p'\ell(\tilde{q})$  has a minima at  $\tilde{q} = \tilde{p}$ . Thus  $D\tilde{L}_p(\tilde{q}) = p'D\ell(\tilde{q}) = 0_{n-1}$  at  $\tilde{q} = \tilde{p}$  and so  $p'D\ell(\tilde{p}) = 0_{n-1}$ . Since for every  $p \in \Delta^n$ ,  $p'D\ell(\tilde{p}) = 0$  we see  $p$  is orthogonal to the tangent space of  $\Lambda$  at  $\ell(\tilde{p})$  and thus normal to  $\Lambda$  at  $\ell(\tilde{p}) = \ell(p)$ . Now suppose there exist  $p, q \in \Delta^n$  such that  $\ell(p) = \ell(q)$ . Since we have just shown that  $p$  and  $q$  must both be normal to  $\Lambda$  at  $\ell(p) = \ell(q)$  and as  $\ell$  is assumed to be continuously differentiable, it must be the case the normals are co-linear, that is,  $p = \alpha q$  for some  $\alpha \in \mathbb{R}$ . However, since  $p \in \Delta^n$ ,  $1 = \sum_i p_i = \alpha \sum_i q_i = \alpha$  and thus  $p = q$ , showing  $\ell$  is invertible.

In order to establish that  $\Pi_\Lambda$  is invertible we proceed by contradiction and assume  $\ell$  is proper and there exist  $p, q \in \Delta^n$  s.t.  $\ell_i(p) = \ell_i(q)$  for  $i \in [n - 1]$  but  $\ell_n(p) \neq \ell_n(q)$ .

Without loss of generality assume  $\ell_n(p) < \ell_n(q)$  (otherwise just swap  $p$  and  $q$ ). This means that  $q'\ell(p) = \sum_{i=1}^n q_i \ell_i(p) = \sum_{i=1}^{n-1} q_i \ell_i(p) + q_n \ell_n(p) < q'\ell(q)$ . However, this contradicts properness of  $\ell$  and therefore the assumption that  $\ell_n(p) \neq \ell_n(q)$ . ■

### 3. Mixability

We use the following characterisation of mixability (as discussed by Vovk and Zhdanov (2009)) and motivate our main result by looking at the binary case. To define mixability we need the notions of a superprediction set and a parametrised exponential operator. The *superprediction set*  $S_\ell$  for a loss  $\ell : \Delta^n \rightarrow \mathbb{R}^n$  is the set of points in  $\mathbb{R}^n$  that point-wise dominate some point on the loss surface. That is,

$$S_\ell := \{\lambda \in \mathbb{R}^n : \exists q \in \Delta^n, \forall i \in [n], \ell_i(q) \leq \lambda_i\}.$$

The  $\beta$ -exponential operator is defined for all  $\lambda \in \mathbb{R}^n$  by

$$E_\beta(\lambda) := (e^{-\beta\lambda_1}, \dots, e^{-\beta\lambda_n}).$$

It is clearly invertible, with inverse  $E_\beta^{-1}(\phi) = -\beta^{-1}(\ln \phi_1, \dots, \ln \phi_n)$ . A loss  $\ell$  is  $\beta$ -mixable when the set  $\Phi_\beta := E_\beta(S_\ell)$  is convex. The *mixability constant*  $\beta_\ell$  of a loss  $\ell$  is the largest  $\beta$  such that  $\ell$  is  $\beta$ -mixable:

$$\beta_\ell := \sup\{\beta > 0 : \ell \text{ is } \beta\text{-mixable}\}.$$

Now

$$\begin{aligned} E_\beta(S_\ell) &= \{E_\beta(\lambda) : \lambda \in \mathbb{R}^n, \exists q \in \Delta^n, \forall i \in [n], \ell_i(q) \leq \lambda_i\} \\ &= \{z \in \mathbb{R}^n : \exists q \in \Delta^n, \forall i \in [n], e^{-\beta\ell_i(q)} \geq z_i\}, \end{aligned}$$

since  $x \mapsto e^{-\beta x}$  is decreasing for  $\beta > 0$ . Hence in order for  $\Phi_\beta$  to be convex the function  $f$  such that  $\text{graph}(f) = \{(e^{-\beta\ell_1(q)}, \dots, e^{-\beta\ell_n(q)}) : q \in \Delta^n\}$  needs to be *concave*.

#### 3.1. The Binary Case

For twice differentiable binary losses  $\ell$  it is known (Haussler et al., 1998) that

$$\beta_\ell = \min_{\tilde{p} \in [0,1]} \frac{\tilde{\ell}'_1(\tilde{p})\tilde{\ell}''_2(\tilde{p}) - \tilde{\ell}''_1(\tilde{p})\tilde{\ell}'_2(\tilde{p})}{\tilde{\ell}'_1(\tilde{p})\tilde{\ell}'_2(\tilde{p})(\tilde{\ell}_2(\tilde{p}) - \tilde{\ell}_1(\tilde{p}))}. \quad (3)$$

When a proper binary loss  $\ell$  is differentiable, the stationarity condition (2) implies

$$\begin{aligned} \tilde{p}\ell'_1(\tilde{p}) + (1-\tilde{p})\ell'_2(\tilde{p}) &= 0 \\ \Rightarrow \tilde{p}\ell'_1(\tilde{p}) &= (\tilde{p}-1)\ell'_2(\tilde{p}) \end{aligned} \quad (4)$$

$$\Rightarrow \frac{\ell'_1(\tilde{p})}{\tilde{p}-1} = \frac{\ell'_2(\tilde{p})}{\tilde{p}} =: w(\tilde{p}) =: w_\ell(\tilde{p}) \quad (5)$$

We have  $\underline{L}(\tilde{p}) = \tilde{p}\ell_1(\tilde{p}) + (1 - \tilde{p})\ell_2(\tilde{p})$ . Thus by differentiating both sides of (4) and substituting into  $\underline{L}''(\tilde{p})$  one obtains  $\underline{L}''(\tilde{p}) = \frac{\ell'_1(\tilde{p})}{1 - \tilde{p}} = -w(\tilde{p})$ . (See Reid and Williamson (2011)). Equation 5 implies  $\tilde{\ell}'_1(\tilde{p}) = (\tilde{p} - 1)w(\tilde{p})$ ,  $\tilde{\ell}'_2(\tilde{p}) = \tilde{p}w(\tilde{p})$  and hence  $\tilde{\ell}''_1(\tilde{p}) = w(\tilde{p}) + (\tilde{p} - 1)w'(\tilde{p})$  and  $\tilde{\ell}''_2(\tilde{p}) = w(\tilde{p}) + \tilde{p}w'(\tilde{p})$ . Substituting these expressions into (3) gives

$$\beta_\ell = \min_{\tilde{p} \in [0,1]} \frac{(\tilde{p} - 1)w(\tilde{p})[w(\tilde{p}) + \tilde{p}w'(\tilde{p})] - [w(\tilde{p}) + (\tilde{p} - 1)w'(\tilde{p})]\tilde{p}w(\tilde{p})}{(\tilde{p} - 1)w(\tilde{p})\tilde{p}w(\tilde{p})[\tilde{p}w(\tilde{p}) - (\tilde{p} - 1)w(\tilde{p})]} = \min_{\tilde{p} \in [0,1]} \frac{1}{\tilde{p}(1 - \tilde{p})w(\tilde{p})}.$$

Observing that  $\underline{L}_{\log}(p) = -p_1 \ln p_1 - p_2 \ln p_2$  we have  $\tilde{\underline{L}}_{\log}(\tilde{p}) = -\tilde{p} \ln \tilde{p} - (1 - \tilde{p}) \ln(1 - \tilde{p})$  and thus  $\tilde{\underline{L}}_{\log}''(\tilde{p}) = \frac{-1}{\tilde{p}(1 - \tilde{p})}$  and so  $w_{\log}(\tilde{p}) = \frac{1}{\tilde{p}(1 - \tilde{p})}$ . Thus

$$\boxed{\beta_\ell = \min_{\tilde{p} \in [0,1]} \frac{w_{\log}(\tilde{p})}{w_\ell(\tilde{p})} = \min_{\tilde{p} \in [0,1]} \frac{\underline{L}_{\log}''(\tilde{p})}{\underline{L}''(\tilde{p})}.} \quad (6)$$

That is, the mixability constant of binary proper losses is the minimal ratio of the weight functions for log loss and the loss in question. The rest of this paper is devoted to the generalisation of (6) to the multiclass case. That there is a relationship between Bayes risk and mixability was also pointed out (in a less explicit form) by Kalnishkan et al. (2004).

### 3.2. Mixability and the Concavity of the function $f_\beta$

Our aim is to understand mixability in terms of other intrinsic properties of the loss function. In particular, we will relate mixability of a loss to the curvature of its Bayes risk surface. In order to do so, we need to be able to compute the curvature of the  $\beta$ -exponentiated superprediction set to determine when it is convex. This is done by first defining a function  $f_\beta : \mathbb{R}^{n-1} \rightarrow \mathbb{R}$  with hypograph

$$\text{hyp}(f_\beta) := \{(\tilde{\phi}, y) \subset \mathbb{R}^n : y \leq f_\beta(\tilde{\phi})\}$$

equal to  $E_\beta(S_\ell)$  and then computing the curvature of  $f_\beta$ . Before we can define  $f_\beta$  we require certain properties of  $E_\beta$  and a mapping  $\tau_\beta : \tilde{\Delta}^n \rightarrow \mathbb{R}^{n-1}$  defined by

$$\tau_\beta(\tilde{p}) := E_\beta(\tilde{\ell}(\tilde{p})) = \left( e^{-\beta\tilde{\ell}_1(\tilde{p})}, \dots, e^{-\beta\tilde{\ell}_{n-1}(\tilde{p})} \right).$$

This takes a point  $\tilde{p}$  to a point  $\tilde{\phi}$  which is the projection of  $\phi = E_\beta(\ell(p))$  onto its first  $n - 1$  coordinates. The range of  $\tau_\beta$  is denoted  $\tilde{\Phi}_\beta$  (see Figure 1).<sup>3</sup>

**Lemma 2** *Let  $\lambda \in \Lambda$  and  $\phi := E_\beta(\lambda)$ . Then  $E_\beta^{-1}(\phi) = -\beta^{-1}(\ln \phi_1, \dots, \ln \phi_n)$  and for all  $\alpha \neq 0$ ,  $E_{\alpha\beta}(E_\beta^{-1}(\phi)) = (\phi_1^\alpha, \dots, \phi_n^\alpha)$ . The derivatives of  $E_\beta$  and its inverse satisfy  $D E_\beta(\lambda) = -\beta \text{diag}(E_\beta(\lambda))$  and  $D E_\beta^{-1}(\phi) = -\beta^{-1} [\text{diag}(\phi)]^{-1}$ . The Hessian of  $E_\beta^{-1}$  is*

$$\mathbf{H} E_\beta^{-1}(\phi) = \frac{1}{\beta} \begin{bmatrix} \text{diag}(\phi_1^{-2}, 0, \dots, 0) \\ \vdots \\ \text{diag}(0, \dots, 0, \phi_n^{-2}) \end{bmatrix}. \quad (7)$$

3. We overload  $E_\beta^{-1}$  using it as both a map  $\Lambda \rightarrow \Phi_\beta$  and from  $\tilde{\Lambda} \rightarrow \tilde{\Phi}_\beta$ . This should not cause confusion because the latter is simply a codimension 1 restriction of the former. Lemma 2 holds for  $n$  and  $n - 1$ .

When  $\beta = 1$  and  $\ell = \ell_{\log} = p \mapsto -(\ln p_1, \dots, \ln p_n)$  the map  $\tau_1$  is the identity map—that is,  $\tilde{\phi} = \tau_1(\tilde{p}) = \tilde{p}$ —and  $E_1^{-1}(\tilde{p}) = \tilde{\ell}_{\log}(\tilde{p})$  is the (projected) log loss.

**Proof** The results concerning inverses and derivatives follow immediately from the definitions. By (24) the Hessian  $\mathbf{H}E_{\beta}^{-1}(\phi) = \mathbf{D}(\mathbf{D}E_{\beta}^{-1}(\phi))$  and so

$$\mathbf{H}E_{\beta}^{-1}(\phi) = \mathbf{D}\left(\left(-\frac{1}{\beta}[\text{diag}(\phi)]^{-1}\right)'\right) = -\frac{1}{\beta}\mathbf{D}\text{diag}(\phi_i^{-1})_{i=1}^n.$$

Let  $h(\phi) = \text{diag}(\phi_i^{-1})_{i=1}^n$ . We have

$$\mathbf{D}h(\phi) := \mathbf{D}\text{vec}h(\text{vec}\phi) = \mathbf{D}\text{vec}h(\phi) = \begin{bmatrix} \text{diag}(-\phi_1^{-2}, 0, \dots, 0) \\ \vdots \\ \text{diag}(0, \dots, 0, -\phi_n^{-2}) \end{bmatrix}.$$

The result for  $\beta = 1$  and  $\ell_{\log}$  follows from  $\tau_1(\tilde{p}) = E_1(\tilde{\ell}(\tilde{p})) = (e^{-1.-\ln \tilde{p}_1}, \dots, e^{-1.-\ln \tilde{p}_{n-1}})$ . ■

**Lemma 3** *The map  $\tilde{\ell}: \tilde{\Delta}^n \rightarrow \tilde{\Lambda}$  is invertible. Also, for all  $\beta > 0$ , the mapping  $\tau_{\beta}: \tilde{\Delta}^n \rightarrow \tilde{\Phi}_{\beta}$  is invertible with inverse  $\tau_{\beta}^{-1} = \tilde{\ell}^{-1} \circ E_{\beta}^{-1}$ .*

**Proof** By “diagram chasing” in Figure 1 we see that  $\tilde{\ell}^{-1} = \Pi_{\Lambda} \circ \ell \circ \Pi_{\Delta}^{-1}$  and  $\tau_{\beta}^{-1} = \Pi_{\Delta} \circ \ell^{-1} \circ E_{\beta}^{-1} \circ \Pi_{\Phi}^{-1}$  provided all the functions on the right hand sides exist.  $\Pi_{\Delta}$  and  $\ell$  exist by definition,  $\Pi_{\Delta}^{-1}$  exists since  $p_i(\tilde{p}) = \tilde{p}_i$  for  $i \in [n-1]$  and  $p_n(\tilde{p}) = 1 - \sum_i \tilde{p}_i$ . The inverse  $\ell^{-1}$  exists by Lemma 1 and  $E_{\beta}^{-1}$  by Lemma 2. Lastly,  $\Pi_{\Phi}$  is invertible since we see  $\Pi_{\Phi} = E_{\beta} \circ \Pi_{\Lambda}^{-1} \circ \tilde{E}_{\beta}^{-1}$  and  $\tilde{E}_{\beta}^{-1}$  clearly exists due to its form and  $\Pi_{\Lambda}^{-1}$  because of Lemma 1. ■

We can now define

$$f_{\beta}: \tilde{\Phi}_{\beta} \ni \tilde{\phi} \mapsto e^{-\beta\tilde{\ell}_n(\tau_{\beta}^{-1}(\tilde{\phi}))} \in [0, \infty). \quad (8)$$

This can be thought of as the inverse of the projection of the  $\beta$ -exponentiated superprediction set  $\tilde{\Phi}_{\beta}$  onto its first  $n-1$  coordinates. That is, if  $\phi \in \tilde{\Phi}_{\beta}$  and  $\tilde{\phi} = \Pi_{\Phi_{\beta}}(\phi)$  then  $\phi = (\tilde{\phi}_1, \dots, \tilde{\phi}_{n-1}, f_{\beta}(\tilde{\phi}))$ . This function plays a central role in the remainder of this paper because it coincides with the boundary of the  $\beta$ -exponentiated superprediction set.

**Lemma 4** *Let  $\beta > 0$  and  $f_{\beta}$  be defined as in (8). Then  $\text{hyp } f_{\beta} = \Phi_{\beta}$ .*

**Proof** We have  $\phi = (\phi_1, \dots, \phi_n)' = (e^{-\beta\ell_1(\tilde{p})}, \dots, e^{-\beta\ell_n(\tilde{p})})'$ . We express  $\phi_n$  as a function of  $\tilde{\phi} = (\phi_1, \dots, \phi_{n-1})' = \tau_{\beta}(\tilde{p})$  using  $\phi_n = e^{-\beta\ell_n(\tilde{p})} = e^{-\beta\ell_n(\tau_{\beta}^{-1}(\tilde{\phi}))}$ . Hence  $\text{graph}(f_{\beta}) = \{(e^{-\beta\ell_1(p)}, \dots, e^{-\beta\ell_n(p)})': p \in \Delta^n\}$ . Since for  $\beta > 0$ ,  $E_{\beta}$  is monotone decreasing in each argument,  $\lambda_i \geq \ell_i(p)$  for all  $i \in [n]$  implies  $E_{\beta}(\lambda) \leq E_{\beta}(\ell(p))$  (coordinatewise). ■

### 3.3. Relating Concavity of $f_\beta$ to the Hessian of $\underline{L}$

The aim of this subsection is to express the Hessian of  $f_\beta$  in terms of the Bayes risk of the loss function defining  $f_\beta$ . We first note that a twice differentiable function  $f : X \rightarrow \mathbb{R}$  defined on  $X \subseteq \mathbb{R}^n$  is concave if and only if its Hessian at  $x$ ,  $Hf(x)$ , is negative semi-definite for all  $x \in X$  (Hiriart-Urruty and Lemaréchal, 1993). The argument that follows consists of repeated applications of the chain and inverse rules for Hessians to compute  $Hf_\beta$ .

We rely on some consequences of the strict properness of  $\ell$  that allow us to derive simple expressions for the Jacobian and Hessian of the projected Bayes risk  $\tilde{\underline{L}} := \underline{L} \circ \Pi_\Delta^{-1} : \Delta^n \rightarrow \mathbb{R}_+$ .

**Lemma 5** *Let  $y(\tilde{p}) := -[p_n(\tilde{p})]^{-1}\tilde{p}$ . Then  $Y(\tilde{p}) := -p_n(\tilde{p})Dy(\tilde{p}) = \left(I_{n-1} + \frac{1}{p_n(\tilde{p})}\tilde{p}\mathbb{1}'_{n-1}\right)$  is invertible for all  $\tilde{p}$ , and*

$$D\tilde{\ell}_n(\tilde{p}) = y(\tilde{p})' \cdot D\tilde{\ell}(\tilde{p}). \quad (9)$$

*The projected Bayes risk function defined by  $\tilde{\underline{L}}(\tilde{p}) := \underline{L}(\Pi_\Delta^{-1}(\tilde{p}))$  satisfies*

$$D\tilde{\underline{L}}(\tilde{p}) = \tilde{\ell}(\tilde{p})' - \tilde{\ell}_n(\tilde{p})\mathbb{1}'_{n-1} \quad (10)$$

$$H\tilde{\underline{L}}(\tilde{p}) = Y(\tilde{p})' \cdot D\tilde{\ell}(\tilde{p}). \quad (11)$$

*Furthermore, for strictly proper  $\ell$  the matrix  $H\tilde{\underline{L}}(\tilde{p})$  is negative definite and invertible for all  $\tilde{p}$  and when  $\beta = 1$  and  $\ell = \ell_{\log}$  is the log loss,*

$$H\tilde{\underline{L}}_{\log}(\tilde{p}) = -Y(\tilde{p})' [\text{diag}(\tilde{p})]^{-1}. \quad (12)$$

**Proof** The stationarity condition (Lemma 1) guarantees that  $p'D\tilde{\ell}(\tilde{p}) = 0_{n-1}$  for all  $p \in \Delta^n$ . This is equivalent to  $\tilde{p}'D\tilde{\ell}(\tilde{p}) + p_n(\tilde{p})D\tilde{\ell}_n(\tilde{p}) = 0_{n-1}$ , which can be rearranged to obtain (9).

By the product rule,

$$\begin{aligned} Dy(\tilde{p}) &= -\tilde{p}D[p_n(\tilde{p})^{-1}] - [p_n(\tilde{p})^{-1}]D\tilde{p} \\ &= \tilde{p}[p_n(\tilde{p})^{-2}]Dp_n(\tilde{p}) - [p_n(\tilde{p})^{-1}]I_{n-1} \\ &= -\tilde{p}[p_n(\tilde{p})^{-2}]\mathbb{1}'_{n-1} - [p_n(\tilde{p})^{-1}]I_{n-1} \\ &= -\frac{1}{p_n(\tilde{p})} \left[ I_{n-1} + \frac{1}{p_n(\tilde{p})}\tilde{p}\mathbb{1}'_{n-1} \right] \end{aligned}$$

since  $p_n(\tilde{p}) = 1 - \sum_{i \in [n-1]} \tilde{p}_i$  implies  $Dp_n(\tilde{p}) = -\mathbb{1}'_{n-1}$ . This establishes that  $Y(\tilde{p}) = I_{n-1} + \frac{1}{p_n(\tilde{p})}\tilde{p}\mathbb{1}'_{n-1}$ . That this matrix is invertible can be easily checked since  $(I_{n-1} - \tilde{p}\mathbb{1}'_{n-1})(I_{n-1} + \frac{1}{p_n(\tilde{p})}\tilde{p}\mathbb{1}'_{n-1}) = I_{n-1}$  by expanding and noting  $\tilde{p}\mathbb{1}'_{n-1}\tilde{p}\mathbb{1}'_{n-1} = (1 - p_n)\tilde{p}\mathbb{1}'_{n-1}$ .

The Bayes risk  $\tilde{\underline{L}}(\tilde{p}) = \tilde{p}'\tilde{\ell}(\tilde{p}) + p_n(\tilde{p})\tilde{\ell}_n(\tilde{p})$ . Taking the derivative and using the product rule ( $Dab = (Da)b + a'(Db)$ ) gives

$$\begin{aligned} D\tilde{\underline{L}}(\tilde{p}) &= D[\tilde{p}'\tilde{\ell}(\tilde{p})] + D[p_n(\tilde{p})\tilde{\ell}_n(\tilde{p})] \\ &= \tilde{\ell}(\tilde{p}) + \tilde{p}'D\tilde{\ell}(\tilde{p}) + [Dp_n(\tilde{p})]\tilde{\ell}_n(\tilde{p}) + p_n(\tilde{p})D\tilde{\ell}_n(\tilde{p}) \\ &= \tilde{\ell}(\tilde{p}) - p_n(\tilde{p})D\tilde{\ell}_n(\tilde{p}) - \tilde{\ell}_n(\tilde{p})\mathbb{1}'_{n-1} + p_n(\tilde{p})D\tilde{\ell}_n(\tilde{p}) \end{aligned}$$

by (9). Thus,  $D\tilde{\underline{L}}(\tilde{p}) = \tilde{\ell}(\tilde{p})' - \tilde{\ell}_n(\tilde{p})\mathbb{1}'_{n-1}$ , establishing (10).

Equation 11 is obtained by taking derivatives once more and using (9) again, yielding

$$\mathsf{H}\tilde{\underline{L}}(\tilde{p}) = \mathsf{D} \left( \left( \mathsf{D}\tilde{\underline{L}}(\tilde{p}) \right)' \right) = \mathsf{D}\tilde{\ell}(\tilde{p}) - \mathbb{1}_{n-1} \cdot \mathsf{D}\tilde{\ell}_n(\tilde{p}) = \left( I_{n-1} + \frac{1}{p_n} \mathbb{1}_{n-1} \tilde{p}' \right) \mathsf{D}\tilde{\ell}(\tilde{p})$$

as required. Now  $\tilde{\underline{L}}(\tilde{p}) = \underline{L}(p_1, \dots, p_{n-1}, p_n(\tilde{p})) = \underline{L}(p_1, \dots, p_{n-1}, 1 - \sum_{i=1}^{n-1} p_i) = \underline{L}(C(\tilde{p}))$  where  $C$  is affine. Since  $p \mapsto \underline{L}(p)$  is strictly concave (Lemma 1) it follows (Hiriart-Urruty and Lemaréchal, 1993) that  $\tilde{\underline{L}}$  is also strictly concave and thus  $\mathsf{H}\tilde{\underline{L}}(\tilde{p})$  is negative definite. It is invertible since we have shown  $Y(\tilde{p})$  is invertible and  $\mathsf{D}\tilde{\ell}$  is invertible by the inverse function theorem and the invertibility of  $\tilde{\ell}$  (Lemma 3).

Finally, equation 12 holds since Lemma 2 gives us  $E_1^{-1} = \tilde{\ell}_{\log}$  so (11) specialises to  $\mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{p}) = Y(\tilde{p})' \cdot \mathsf{D}\tilde{\ell}_{\log}(\tilde{p}) = Y(\tilde{p})' \cdot \mathsf{D}E_1^{-1}(\tilde{p}) = -Y(\tilde{p})' \cdot [\text{diag}(\tilde{p})]^{-1}$ , also by Lemma 2. ■

### 3.4. Completion of the Argument

Recall that our aim is to compute the Hessian of the boundary of the  $\beta$ -exponentiated superprediction set and determine when it is negative semidefinite. The boundary is described by the function  $f_\beta$  which can be written as the composition  $h_\beta \circ g_\beta$  where  $h_\beta : \mathbb{R} \rightarrow [0, \infty)$  and  $g_\beta : \tilde{\Phi}_\beta \rightarrow \mathbb{R}_+$  are defined by  $h_\beta(z) := e^{-\beta z}$  and  $g_\beta(\tilde{\phi}) := \tilde{\ell}_n(\tau_\beta^{-1}(\tilde{\phi}))$ . The Hessian of  $f_\beta$  can be expanded in terms of  $g_\beta$  using the chain rule for the Hessian (Theorem 13) as follows.

**Lemma 6** *For all  $\tilde{\phi} \in \tilde{\Phi}$ , the Hessian of  $f_\beta$  at  $\tilde{\phi}$  is*

$$\mathsf{H}f_\beta(\tilde{\phi}) = \beta e^{-\beta g_\beta(\tilde{\phi})} \Gamma_\beta(\tilde{\phi}), \quad (13)$$

where  $\Gamma_\beta(\tilde{\phi}) := \beta \mathsf{D}g_\beta(\tilde{\phi})' \cdot \mathsf{D}g_\beta(\tilde{\phi}) - \mathsf{H}g_\beta(\tilde{\phi})$ . Furthermore, for  $\beta > 0$  the negative semi-definiteness of  $\mathsf{H}f_\beta(\tilde{\phi})$  (and thus the concavity of  $f_\beta$ ) is equivalent to the negative semi-definiteness of  $\Gamma_\beta(\tilde{\phi})$ .

**Proof** Using  $f := f_\beta$  and  $g := g_\beta$  temporarily and letting  $z = g(\tilde{\phi})$ , the chain rule for  $\mathsf{H}$  gives

$$\begin{aligned} \mathsf{H}f(\tilde{\phi}) &= \left( I_1 \otimes \mathsf{D}g(\tilde{\phi})' \right) \cdot (\mathsf{H}h_\beta(z)) \cdot \mathsf{D}g(\tilde{\phi}) + (\mathsf{D}h_\beta(z) \otimes I_{n-1}) \cdot \mathsf{H}g(\tilde{\phi}) \\ &= \beta^2 e^{-\beta z} \mathsf{D}g(\tilde{\phi})' \cdot \mathsf{D}g(\tilde{\phi}) - \beta e^{-\beta z} \mathsf{H}g(\tilde{\phi}) \\ &= \beta e^{-\beta g(\tilde{\phi})} \left[ \beta \mathsf{D}g(\tilde{\phi})' \cdot \mathsf{D}g(\tilde{\phi}) - \mathsf{H}g(\tilde{\phi}) \right] \end{aligned}$$

since  $\alpha \otimes A = \alpha A$  for scalar  $\alpha$  and matrix  $A$  and  $\mathsf{D}h_\beta(z) = \mathsf{D}[\exp(-\beta z)] = -\beta e^{-\beta z}$  so  $\mathsf{H}h(z) = \beta^2 e^{-\beta z}$ . Whether  $\mathsf{H}f \preccurlyeq 0$  depends only on  $\Gamma_\beta$  since  $\beta e^{-\beta g(\tilde{\phi})}$  is positive for all  $\beta > 0$  and  $\tilde{\phi}$ . ■

**Lemma 7** For strictly proper  $\ell$  and  $\lambda := E_\beta^{-1}(\tilde{\phi})$  and  $\tilde{p} := \tilde{\ell}^{-1}(\lambda)$ ,

$$\mathsf{D}g_\beta(\tilde{\phi}) = y(\tilde{p})' A_\beta(\tilde{\phi}) \quad (14)$$

$$\mathsf{H}g_\beta(\tilde{\phi}) = -\frac{1}{p_n(\tilde{p})} A_\beta(\tilde{\phi})' \cdot \left[ \beta \operatorname{diag}(\tilde{p}) + Y(\tilde{p}) \cdot \left[ \mathsf{H}\tilde{L}(\tilde{p}) \right]^{-1} \cdot Y(\tilde{p})' \right] \cdot A_\beta(\tilde{\phi}), \quad (15)$$

where  $A_\beta(\tilde{\phi}) := \mathsf{D}E_\beta^{-1}(\tilde{\phi})$ .

**Proof** By definition,  $g_\beta(\tilde{\phi}) := \tilde{\ell}_n(\tau_\beta^{-1}(\tilde{\phi}))$ . Since  $\tau_\beta^{-1} = \tilde{\ell}^{-1} \circ E_\beta^{-1}$  we have  $g_\beta = \tilde{\ell}_n \circ \tilde{\ell}^{-1} \circ E_\beta^{-1}$ . Thus, by Lemma 5 equation (9), the inverse function theorem, and chain rule we have

$$\mathsf{D}g_\beta(\tilde{\phi}) = \mathsf{D}\tilde{\ell}_n(\tilde{p}) \cdot \mathsf{D}\tilde{\ell}^{-1}(\lambda) \cdot \mathsf{D}E_\beta^{-1}(\tilde{\phi}) = y(\tilde{p})' \mathsf{D}\tilde{\ell}(\tilde{p}) \cdot \left[ \mathsf{D}\tilde{\ell}(\tilde{p}) \right]^{-1} \cdot \left[ \mathsf{D}E_\beta^{-1}(\tilde{\phi}) \right] = y(\tilde{p})' A_\beta(\tilde{\phi})$$

yielding (14). Since  $\tilde{p} = \tau_\beta^{-1}(\tilde{\phi})$  and  $\mathsf{H}g_\beta = \mathsf{D}((\mathsf{D}g_\beta)')$  (see (24)), the chain and product rules give

$$\begin{aligned} \mathsf{H}g_\beta(\tilde{\phi}) &= \mathsf{D} \left[ \left( \mathsf{D}E_\beta^{-1}(\tilde{\phi}) \right)' \cdot y \left( \tau_\beta^{-1}(\tilde{\phi}) \right) \right] \\ &= \left( y(\tau_\beta^{-1}(\tilde{\phi}))' \otimes I_{n-1} \right) \cdot \mathsf{D} \left( \mathsf{D}E_\beta^{-1}(\tilde{\phi})' \right) + \left( I_1 \otimes (\mathsf{D}E_\beta^{-1}(\tilde{\phi}))' \right) \cdot \mathsf{D} \left( y \left( \tau_\beta^{-1}(\tilde{\phi}) \right) \right) \\ &= (y(\tilde{p})' \otimes I_{n-1}) \cdot \mathsf{H}E_\beta^{-1}(\tilde{\phi}) + \left( \mathsf{D}E_\beta^{-1}(\tilde{\phi}) \right)' \cdot \mathsf{D}y(\tilde{p}) \cdot \mathsf{D}\tau_\beta^{-1}(\tilde{\phi}) \\ &= -\frac{\beta}{p_n(\tilde{p})} A_\beta(\tilde{\phi}) \cdot \operatorname{diag}(\tilde{p}) \cdot A_\beta(\tilde{\phi}) + A_\beta(\tilde{\phi})' \cdot \mathsf{D}y(\tilde{p}) \cdot \mathsf{D}\tau_\beta^{-1}(\tilde{\phi}). \end{aligned} \quad (16)$$

The first summand above is due to (7) and the fact that

$$\begin{aligned} (y \otimes I_{n-1}) \cdot \mathsf{H}E_\beta^{-1}(\tilde{\phi}) &= \frac{1}{\beta} [y_1 I_{n-1}, \dots, y_{n-1} I_{n-1}] \cdot \begin{bmatrix} \operatorname{diag}(\phi_1^{-2}, 0, \dots, 0) \\ \vdots \\ \operatorname{diag}(0, \dots, 0, \phi_{n-1}^{-2}) \end{bmatrix} \\ &= \frac{1}{\beta} \sum_{i=1}^{n-1} y_i \cdot I_{n-1} \cdot \operatorname{diag}(0, \dots, 0, \phi_i^{-2}, 0, \dots, 0) \\ &= \frac{1}{\beta} \operatorname{diag}(y_i \phi_i^{-2})_{i=1}^{n-1} \\ &= \frac{-\beta}{p_n(\tilde{p})} A_\beta(\tilde{\phi})' \cdot \operatorname{diag}(\tilde{p}) \cdot A_\beta(\tilde{\phi}). \end{aligned}$$

The last equality holds because  $A_\beta(\tilde{\phi})' \cdot A_\beta(\tilde{\phi}) = \beta^{-2} \operatorname{diag}(\phi_i^{-2})_{i=1}^{n-1}$  by Lemma 2, the definition of  $y(\tilde{p}) = -[p_n(\tilde{p})]^{-1} \tilde{p}$ , and because all the matrices are diagonal and thus commute.

The second summand in (16) reduces by  $\mathsf{D}y(\tilde{p}) = -\frac{1}{p_n(\tilde{p})} Y(\tilde{p})$  from Lemma 5 and  $\tau_\beta = E_\beta \circ \tilde{\ell}$ :

$$\mathsf{D}\tau_\beta^{-1}(\tilde{\phi}) = \left[ \mathsf{D}E_\beta(\lambda) \cdot \mathsf{D}\tilde{\ell}(\tilde{p}) \right]^{-1} = \left[ \mathsf{D}E_\beta(\lambda) \cdot (Y(\tilde{p})')^{-1} \cdot \mathsf{H}\tilde{L}(\tilde{p}) \right]^{-1} = \left[ \mathsf{H}\tilde{L}(\tilde{p}) \right]^{-1} \cdot Y(\tilde{p})' \cdot \mathsf{D}E_\beta^{-1}(\lambda).$$

Substituting these into (16) gives

$$\mathsf{H}g_\beta(\tilde{\phi}) = -\frac{\beta}{p_n(\tilde{p})} A_\beta(\tilde{\phi}) \cdot \text{diag}(\tilde{p}) \cdot A_\beta(\tilde{\phi}) - \frac{1}{p_n(\tilde{p})} A_\beta(\tilde{\phi})' \cdot Y(\tilde{p}) \cdot [\mathsf{H}\tilde{L}(\tilde{p})]^{-1} \cdot Y(\tilde{p})' \cdot A_\beta(\tilde{\phi}),$$

which can be factored into the required result.  $\blacksquare$

We can now use the last two lemmata to express the function  $\Gamma_\beta$  in terms of the Hessian of the Bayes risk functions for the specified loss  $\ell$  and the log loss.

**Lemma 8** *The matrix-valued function  $\Gamma_\beta$  satisfies, for all  $\tilde{\phi} \in \tilde{\Phi}$  and  $\tilde{p} = \tau_\beta^{-1}(\tilde{\phi})$ ,*

$$\Gamma_\beta(\tilde{\phi}) = \frac{1}{p_n} A_\beta(\tilde{\phi})' \cdot Y(\tilde{p}) \left[ [\mathsf{H}\tilde{L}(\tilde{p})]^{-1} - \beta [\mathsf{H}\tilde{L}_{\log}(\tilde{p})]^{-1} \right] \cdot Y(\tilde{p})' \cdot A_\beta(\tilde{\phi}), \quad (17)$$

and, for each  $\tilde{\phi}$ , is negative semi-definite if and only if  $R(\beta, \ell, \tilde{p}) := [\mathsf{H}\tilde{L}(\tilde{p})]^{-1} - \beta [\mathsf{H}\tilde{L}_{\log}(\tilde{p})]^{-1}$  is negative semi-definite.

**Proof** Substituting the values of  $\mathsf{D}g_\beta$  and  $\mathsf{H}g_\beta$  from Lemma 7 into the definition of  $\Gamma_\beta$  from Lemma 6 and then using Lemma 2 and the definition of  $y(\tilde{p})$ , we obtain

$$\begin{aligned} \Gamma_\beta(\tilde{\phi}) &= \beta A_\beta(\tilde{\phi})' \cdot y(\tilde{p}) \cdot y(\tilde{p})' \cdot A_\beta(\tilde{\phi}) + \frac{1}{p_n(\tilde{p})} A_\beta(\tilde{\phi})' \cdot \left[ \beta \text{diag}(\tilde{p}) + Y(\tilde{p}) \cdot [\mathsf{H}\tilde{L}(\tilde{p})]^{-1} \cdot Y(\tilde{p})' \right] \cdot A_\beta(\tilde{\phi}) \\ &= \frac{1}{p_n} A_\beta(\tilde{\phi})' \cdot \left[ \beta \frac{1}{p_n} \tilde{p} \cdot \tilde{p}' + \beta \text{diag}(\tilde{p}) + Y(\tilde{p}) \cdot [\mathsf{H}\tilde{L}(\tilde{p})]^{-1} \cdot Y(\tilde{p})' \right] \cdot A_\beta(\tilde{\phi}). \end{aligned} \quad (18)$$

Using Lemma 5 we then see that

$$\begin{aligned} -Y(\tilde{p}) \cdot [\mathsf{H}\tilde{L}_{\log}(\tilde{p})]^{-1} \cdot Y(\tilde{p})' &= -Y(\tilde{p}) \cdot [-Y(\tilde{p})' \text{diag}(\tilde{p})^{-1}]^{-1} \cdot Y(\tilde{p})' \\ &= Y(\tilde{p}) \cdot \text{diag}(\tilde{p}) \cdot (Y(\tilde{p})')^{-1} \cdot Y(\tilde{p})' \\ &= (I_{n-1} + \frac{1}{p_n} \mathbb{1}_{n-1} \tilde{p}') \cdot \text{diag}(\tilde{p}) \\ &= \text{diag}(\tilde{p}) + \frac{1}{p_n} \tilde{p} \cdot \tilde{p}'. \end{aligned}$$

Substituting this for the appropriate terms in (18) gives

$$\Gamma_\beta(\tilde{\phi}) = \frac{1}{p_n} A_\beta(\tilde{\phi})' \cdot \left[ Y(\tilde{p}) \cdot [\mathsf{H}\tilde{L}(\tilde{p})]^{-1} \cdot Y(\tilde{p})' - \beta Y(\tilde{p}) \cdot [\mathsf{H}\tilde{L}_{\log}(\tilde{p})]^{-1} \cdot Y(\tilde{p})' \right] \cdot A_\beta(\tilde{\phi})$$

which equals (17).

Since  $\Gamma_\beta = [p_n]^{-1} B R B'$  where  $B = A_\beta(\tilde{\phi})' Y(\tilde{p})$  and  $R = R(\beta, \ell, \tilde{p})$  the definition of negative semi-definiteness and the positivity of  $p_n$  means we need to show that  $\forall x : x' \Gamma_\beta x \leq 0 \iff \forall y : y' R y \leq 0$ . It suffices to show that  $B$  is invertible, since we can let  $y = Bx$  to establish the equivalence. The matrix  $A_\beta(\tilde{\phi})$  is invertible since, by definition,  $A_\beta(\tilde{\phi}) = \mathsf{D}E_\beta^{-1}(\tilde{\phi}) = -\beta^{-1}[\text{diag}(\tilde{\phi})]^{-1}$  by Lemma 2 and so has matrix inverse  $-\beta \text{diag}(\tilde{\phi})$ .

The matrix  $Y(\tilde{p})$  is invertible by Lemma 7. Thus,  $B$  is invertible because it is the product of two invertible matrices.  $\blacksquare$

The above arguments result in a characterisation of the concavity of the function  $f_\beta$  (via its Hessian)—and hence the convexity of the  $\beta$ -exponentiated superprediction set—in terms of the Hessian of the Bayes risk function of the loss  $\ell$  and the log loss  $\ell_{\log}$ . As in the binary case (cf. (6)), this means we are now able to specify the mixability constant  $\beta_\ell$  in terms of the curvature  $\mathbf{H}\tilde{L}$  of the Bayes risk for  $\ell$  relative to the curvature  $\mathbf{H}\tilde{L}_{\log}$  of the Bayes risk for log loss.

**Lemma 9** *The mixability constant  $\beta_\ell$  of a twice differentiable strictly proper loss  $\ell$  is*

$$\beta_\ell = \sup \left\{ \beta > 0 : \forall \tilde{p} \in \tilde{\Delta}^n, \beta \mathbf{H}\tilde{L}(\tilde{p}) \succcurlyeq \mathbf{H}\tilde{L}_{\log}(\tilde{p}) \right\}, \quad (19)$$

where  $\tilde{L}(\tilde{p}) := L(p)$  is the Bayes risk of  $\ell$  and  $\tilde{L}_{\log}$  is the Bayes risk for the log loss.

**Proof** By Lemma 6 and Lemma 8 we know  $\mathbf{H}f_\beta(\tilde{p}) \preccurlyeq 0 \iff R(\beta, \ell, \tilde{p}) \preccurlyeq 0$ . By Lemma 5,  $\mathbf{H}\tilde{L}(\tilde{p}) \prec 0$  and  $\mathbf{H}\tilde{L}_{\log}(\tilde{p}) \prec 0$  for all  $\tilde{p}$  and so we can use the fact that for positive definite matrices  $A$  and  $B$  we have  $A \succcurlyeq B \iff B^{-1} \succcurlyeq A^{-1}$  (Horn and Johnson, 1985, Corollary 7.7.4). This means  $R(\beta, \ell, \tilde{p}) \preccurlyeq 0 \iff \mathbf{H}\tilde{L}(\tilde{p})^{-1} \preccurlyeq \beta \mathbf{H}\tilde{L}_{\log}(\tilde{p})^{-1} \iff \beta^{-1} \mathbf{H}\tilde{L}_{\log}(\tilde{p}) \preccurlyeq \mathbf{H}\tilde{L}(\tilde{p}) \iff \beta \mathbf{H}\tilde{L}(\tilde{p}) \succcurlyeq \mathbf{H}\tilde{L}_{\log}(\tilde{p})$ . Therefore  $f_\beta$  is concave at  $\tilde{p}$  if and only if  $\beta \mathbf{H}\tilde{L}(\tilde{p}) \succcurlyeq \mathbf{H}\tilde{L}_{\log}(\tilde{p})$ . The mixability constant  $\beta_\ell$  is defined in Section 3 to be the largest  $\beta > 0$  such that the  $\beta$ -exponentiated superprediction set  $E_\beta(S_\ell)$  is convex. This is equivalent to the function  $f_\beta$  being concave at all  $\tilde{p}$ . Thus, we have shown  $\beta_\ell = \sup\{\beta > 0 : \forall \tilde{p} \in \tilde{\Delta}^n, \beta \mathbf{H}\tilde{L}(\tilde{p}) \succcurlyeq \mathbf{H}\tilde{L}_{\log}(\tilde{p})\}$  as required.  $\blacksquare$

The mixability constant can also be expressed in terms of the maximal eigenvalue of the “ratio” of the Hessian matrices for the Bayes risk for log loss and the loss in question. In the following,  $\lambda_i(A)$  will denote the  $i$ th largest (possibly repeated) eigenvalue of the  $n \times n$  symmetric matrix  $A$ . That is,  $\lambda_{\min}(A) := \lambda_1(A) \leq \lambda_2(A) \leq \dots \leq \lambda_n =: \lambda_{\max}(A)$  where each  $\lambda_i(A)$  satisfies  $|A - \lambda_i(A)I| = 0$ .

**Theorem 10** *For any twice differentiable strictly proper loss  $\ell$ , the mixability constant is*

$$\boxed{\beta_\ell = \min_{\tilde{p} \in \tilde{\Delta}^n} \lambda_{\max} \left( (\mathbf{H}\tilde{L}(\tilde{p}))^{-1} \cdot \mathbf{H}\tilde{L}_{\log}(\tilde{p}) \right)}. \quad (20)$$

Equation 20 reduces to (6) when  $n = 2$  since the maximum eigenvalue of a  $1 \times 1$  matrix is simply its single entry.

**Proof** We define  $C_\beta(\tilde{p}) := \beta \mathbf{H}\tilde{L}(\tilde{p}) - \mathbf{H}\tilde{L}_{\log}(\tilde{p})$  and  $\rho(\tilde{p}) := \mathbf{H}\tilde{L}(\tilde{p})^{-1} \cdot \mathbf{H}\tilde{L}_{\log}(\tilde{p})$  and for any fixed  $\tilde{p}$ , we first show that zero is an eigenvalue of  $C_\beta(\tilde{p})$  if and only if  $\beta$  is an eigenvalue of  $\rho(\tilde{p})$ . This can be seen since  $\mathbf{H}\tilde{L}(\tilde{p})$  is invertible (Lemma 5) so

$$\begin{aligned} |C_\beta(\tilde{p}) - 0I| = 0 &\iff |\beta \mathbf{H}\tilde{L}(\tilde{p}) - \mathbf{H}\tilde{L}_{\log}(\tilde{p})| = 0 \iff |\mathbf{H}\tilde{L}(\tilde{p})^{-1}| |\beta \mathbf{H}\tilde{L}(\tilde{p}) - \mathbf{H}\tilde{L}_{\log}(\tilde{p})| = 0 \\ &\iff \left| \mathbf{H}\tilde{L}(\tilde{p})^{-1} \cdot [\beta \mathbf{H}\tilde{L}(\tilde{p}) - \mathbf{H}\tilde{L}_{\log}(\tilde{p})] \right| = 0 \iff |\beta I - \mathbf{H}\tilde{L}(\tilde{p})^{-1} \cdot \mathbf{H}\tilde{L}_{\log}(\tilde{p})| = 0. \end{aligned}$$

Since a symmetric matrix is p.s.d. if and only if all its eigenvalues are non-negative it must be the case that if  $\lambda_{\min}(C_\beta(\tilde{p})) \geq 0$  then  $C_\beta(\tilde{p}) \succcurlyeq 0$  since every other eigenvalue is bigger than the minimum one. Conversely, if  $C_\beta(\tilde{p}) \not\succcurlyeq 0$  then at least one eigenvalue must be negative, thus the smallest eigenvalue must be negative. Thus,  $\lambda_{\min}(C_\beta(\tilde{p})) \geq 0 \iff C_\beta(\tilde{p}) \succcurlyeq 0$ . Now define  $\beta(\tilde{p}) := \sup\{\beta > 0 : C_\beta(\tilde{p}) \succcurlyeq 0\} = \sup\{\beta > 0 : \lambda_{\min}(C_\beta(\tilde{p})) \geq 0\}$ . We show that for each  $\tilde{p}$  the function  $\beta \mapsto \lambda_{\min}(C_\beta(\tilde{p}))$  is continuous and only has a single root. First, continuity is because the entries of  $C_\beta(\tilde{p})$  are continuous in  $\beta$  for each  $\tilde{p}$  and eigenvalues are continuous functions of their matrix's entries (Horn and Johnson, 1985, Appendix D). Second, as a function of its matrix arguments, the minimum eigenvalue  $\lambda_{\min}$  is known to be concave (Magnus and Neudecker, 1999, §11.6). Thus, for any fixed  $\tilde{p}$ , its restriction to the convex set of matrices  $\{C_\beta(\tilde{p}) : \beta > 0\}$  is also concave in its entries and so in  $\beta$ . Since  $C_0(\tilde{p}) = -\mathsf{H}\tilde{L}_{\log}(\tilde{p})$  is positive definite for every  $\tilde{p}$  (Lemma 5) we have  $\lambda_{\min}(C_0(\tilde{p})) > 0$  and so, by the concavity of the map  $\beta \mapsto \lambda_{\min}(C_\beta(\tilde{p}))$ , there can be only one  $\beta > 0$  for which  $\lambda_{\min}(C_\beta(\tilde{p})) = 0$  and by continuity it must be largest non-negative one, that is,  $\beta(\tilde{p})$ .

Thus  $\beta(\tilde{p}) = \sup\{\beta > 0 : \lambda_{\min}(C_\beta(\tilde{p})) = 0\} = \sup\{\beta > 0 : \beta \text{ is an eigenvalue of } \rho(\tilde{p})\} = \lambda_{\max}(\rho(\tilde{p}))$ . Now let  $\beta^* := \min_{\tilde{p}} \beta(\tilde{p}) = \min_{\tilde{p}} \lambda_{\max}(\rho(\tilde{p}))$  and let  $\tilde{p}^*$  be a minimiser so that  $\beta^* = \beta(\tilde{p}^*)$ . We now claim that  $C_{\beta^*}(\tilde{p}) \succcurlyeq 0$  for all  $\tilde{p}$  since if there was some  $\tilde{q} \in \tilde{\Delta}^n$  such that  $C_{\beta^*}(\tilde{q}) \not\succcurlyeq 0$  we would have  $\beta(\tilde{q}) < \beta^*$  since  $\beta \mapsto \lambda_{\min}(C_\beta(\tilde{q}))$  only has a single root—a contradiction. Thus, since we have shown  $\beta^*$  is the largest  $\beta$  such that  $C_{\beta^*}(\tilde{p}) \succcurlyeq 0$  it must be  $\beta_\ell$ , by Lemma 9, as required. ■

#### 4. Discussion

In combination with the existing results on mixability, our result bounds the performance of certain predictors in terms of the Hessian of the Bayes risk  $\mathsf{H}\underline{L}$  which depends on the choice of loss function. This implies a generalisation of the main result of Kalnishkan and Vyugin (2002a) which shows there can be no “predictive complexity” when the curvature of  $f_\beta$  vanishes (in the binary case). This means there can not exist a mixability constant  $\beta_\ell$  of the form (1) in such a situation. This is apparent from (20) since  $\beta_\ell$  is not defined when  $\mathsf{H}\tilde{L}(\tilde{p})$  is singular (which occurs when  $\mathsf{H}f_\beta$  vanishes).

One can use Lemma 9 to confirm that the mixability constant for the Brier score is one, in accord with the calculation of Vovk and Zhdanov (2009). (See Appendix B for the proof.)

The main result is stated for proper losses. However it turns out that this is not really a limitation<sup>4</sup>. Suppose  $\ell_{\text{imp}}: [n] \times \mathcal{V} \rightarrow [0, +\infty]$  is an *improper* loss (i.e. not proper). Let  $L_{\text{imp}}: \Delta^n \times \mathcal{V} \rightarrow [0, +\infty]$  and  $\underline{L}_{\text{imp}}: \Delta^n \rightarrow [0, +\infty]$  denote the corresponding conditional risk and conditional Bayes risk respectively. Let  $\psi_{\text{imp}}: \Delta^n \rightarrow \mathcal{V}$  be a *reference link* (cf. Reid and Williamson (2010))—that is, a (possibly non-unique) function satisfying

$$L_{\text{imp}}(p, \psi_{\text{imp}}(p)) = \underline{L}_{\text{imp}}(p).$$

---

4. We thank a referee for pointing this out by referring us to Chernov et al. (2010).

This function can be seen as one which ‘‘calibrates’’  $\ell_{\text{imp}}$  by returning  $\psi_{\text{imp}}(p)$ , the best possible prediction under labels distributed by  $p$ . Let

$$\ell(y, q) := \ell_{\text{imp}}(y, \psi_{\text{imp}}(q)), \quad y \in [n], q \in \Delta^n \quad (21)$$

and thus

$$L(p, q) = L_{\text{imp}}(p, \psi_{\text{imp}}(q)), \quad p, q \in \Delta^n.$$

We claim that  $\ell$  is proper. It suffices to show that  $p \in \arg \min_{q \in \Delta^n} L(p, q)$  which we demonstrate by contradiction. Thus suppose that for arbitrary  $p \in \Delta^n$ , there exists  $p^* \neq p$  such that

$$\begin{aligned} & L(p, p^*) < L(p, p) \\ \Leftrightarrow & L_{\text{imp}}(p, \psi_{\text{imp}}(p^*)) < L_{\text{imp}}(p, \psi_{\text{imp}}(p)) = \underline{L}_{\text{imp}}(p) = \min_{v \in \mathcal{V}} L_{\text{imp}}(p, v) \end{aligned}$$

which is indeed a contradiction. Thus  $\ell$  defined by (21) is proper. Observe too that  $\underline{L}_{\text{imp}}(p) = L_{\text{imp}}(p, \psi_{\text{imp}}(p)) = L(p, p) = \underline{L}(p)$ . Thus the method of identifying the conditional Bayes risk of an improper loss with that of a proper loss (confer (Grünwald and Dawid, 2004, §3.4) and Chernov et al. (2010)) is equivalent to the above use of the reference link.

We now briefly relate our result to recent work by Abernethy et al. (2009). They formulate the problem slightly differently. They do not restrict themselves to proper losses and so the predictions are not restricted to the simplex. This means it is not necessary to go to a submanifold in order for derivatives to be well defined. (It may well be that one can avoid the explicit projection down to  $\tilde{\Delta}^n$  using the intrinsic methods of differential geometry (Thorpe, 1979); we have been unable as yet to prove our result using that machinery.)

Abernethy et al. (2009) have developed their own bounds on cumulative loss in terms of the  $\alpha$ -flatness (defined below) of  $\underline{L}$ . They show that  $\alpha$ -flatness is implied by strong convexity of the loss  $\ell$ . The duality between the loss surface and Bayes risk that they established through the use of support functions can also be seen in Lemma 5 in the relationship between the Hessian of  $\tilde{\underline{L}}$  and the derivative of  $\tilde{\ell}$ . Although it is obscured somewhat due to our use of functions of  $\tilde{p}$ , this relationship is due to the properness of  $\ell$  guaranteeing that  $\ell^{-1}$  is the (homogeneously extended) Gauss map for the surface  $\tilde{\underline{L}}$ . Below we point out the relationship between  $\alpha$ -flatness and the positive definiteness of  $\underline{H}\underline{L}$  (we stress that in our work we used  $\underline{H}\tilde{\underline{L}}$ ). The connection below suggests that the  $\alpha$ -flatness condition is stronger than necessary.

A convex function  $f: \mathcal{X} \rightarrow \mathbb{R}$  is said to be  $\alpha$ -flat if for all  $x, x_0 \in \mathcal{X}$ ,

$$f(x) - f(x_0) \leq Df(x_0) \cdot (x - x_0) + \alpha \|x - x_0\|^2. \quad (22)$$

A concave function  $g$  is  $\alpha$ -flat if the convex function  $-g$  is  $\alpha$ -flat.

**Theorem 11** *For  $\alpha > 0$ ,  $f$  is  $\alpha$ -flat if and only if  $f - \alpha\|\cdot\|^2$  is concave.*

**Proof** Hiriart-Urruty and Lemaréchal (1993, page 183) show a function  $h$  is convex if and only if

$$h(x) \geq h(x_0) + Dh(x_0) \cdot (x - x_0), \quad \forall x, x_0.$$

A function  $h$  is concave if and only if  $-h$  is convex. Thus  $h$  is concave if and only

$$h(x) \leq h(x_0) + Dh(x_0) \cdot (x - x_0), \quad \forall x, x_0.$$

Let  $h(x) = f(x) - \alpha\|x\|^2$ . The concavity of  $h$  is equivalent to the following holding for all  $x, x_0$ :

$$\begin{aligned} & f(x) - \alpha\|x\|^2 \leq f(x_0) - \alpha\|x_0\|^2 + (Df(x_0) - 2\alpha x_0) \cdot (x - x_0) \\ \Leftrightarrow & f(x) - \alpha\|x\|^2 \leq f(x_0) - \alpha\|x_0\|^2 + Df(x_0) \cdot (x - x_0) - 2\alpha x_0 \cdot (x - x_0) \\ \Leftrightarrow & f(x) \leq f(x_0) - \alpha\|x_0\|^2 + Df(x_0) \cdot (x - x_0) + \alpha\|x\|^2 - 2\alpha x_0 \cdot x + 2\alpha\|x_0\|^2 \\ \Leftrightarrow & f(x) \leq f(x_0) + Df(x_0) \cdot (x - x_0) + \alpha\|x - x_0\|^2 \\ \Leftrightarrow & (22). \end{aligned}$$

■

Thus  $f$  is  $\alpha$ -flat if and only if  $H(f - \alpha\|\cdot\|^2)$  is negative semidefinite, which is equivalent to  $Hf - 2\alpha I \preceq 0 \iff Hf \preceq 2\alpha I$ . Hence requiring  $-\underline{L}$  is  $\alpha$ -flat is a constraint on the curvature of  $\underline{L}$  relative to a flat surface:  $\underline{L}$  is  $\alpha$ -flat iff  $H\underline{L} \succeq -2\alpha I$ . However our main result shows that the mixability constant (which is the best possible constant one can have in a bound such as (1)) is governed by the curvature of  $\tilde{\underline{L}}$  normalised by the curvature of  $\tilde{\underline{L}}_{\log}$ . The necessity of comparison with log loss is not that surprising in light of the observations regarding mixability by Grünwald (2007, §17.9).

## 5. Conclusion

We have characterised the mixability constant for strictly proper multiclass losses (and shown how the result also applies to improper losses). The result shows in a precise and intuitive way the effect of the choice of loss function on the performance of an aggregating forecaster and the special role played by Log-loss in such settings.

## Acknowledgments

This work was supported by the Australian Research Council and NICTA through backing Australia's ability. Some of the work was done while all the authors were visiting Microsoft Research, Cambridge and some was done while Tim van Erven was visiting ANU and NICTA. It was also supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

## References

- Jacob Abernethy, Alekh Agarwal, Peter L. Bartlett, and Alexander Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.
- Nicolò Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

- Alexey Chernov, Yuri Kalnishkan, Fedor Zhdanov, and Vladimir Vovk. Supermartingales in prediction with expert advice. *Theoretical Computer Science*, 411:2647–2669, 2010.
- Paul K. Fackler. Notes on matrix calculus. North Carolina State University, 2005.
- Wendell H. Fleming. *Functions of Several Variables*. Springer, 1977.
- Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.
- Peter D. Grünwald and A. Phillip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.
- David Haussler, Jyrki Kivinen, and Manfred K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998.
- Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms: Part I: Fundamentals*. Springer, Berlin, 1993.
- Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, 1985.
- Yuri Kalnishkan and Michael V. Vyugin. On the absence of predictive complexity for some games. In *Proceedings of the 13th International Conference on Algorithmic Learning Theory*, volume 2533 of *Lecture Notes in Artificial Intelligence*, pages 164–172. Springer-Verlag, 2002a.
- Yuri Kalnishkan and Michael V. Vyugin. Mixability and the existence of weak complexities. In *The 15th Annual Conference on Computational Learning Theory (COLT 2002)*, volume 2375 of *Lecture Notes in Artificial Intelligence*, pages 105–120. Springer-Verlag, 2002b.
- Yuri Kalnishkan and Michael V. Vyugin. The weak aggregating algorithm and weak mixability. *Journal of Computer and System Sciences*, 74:1228–1244, 2008.
- Yuri Kalnishkan, Volodya Vovk, and Michael V. Vyugin. Loss functions, complexities, and the Legendre transformation. *Theoretical Computer Science*, 313:195–207, 2004.
- Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics (revised edition)*. John Wiley & Sons, Ltd., 1999.
- Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.
- Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, March 2011.
- John A. Thorpe. *Elementary Topics in Differential Geometry*. Springer, 1979.
- Volodya Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT)*, pages 371–383, 1990.

Volodya Vovk. A game of prediction with expert advice. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 51–60. ACM, 1995.

Volodya Vovk and Fedor Zhdanov. Prediction with expert advice for the Brier game. *Journal of Machine Learning Research*, 10:2445–2471, 2009.

## Appendix A. Matrix Calculus

We adopt notation from Magnus and Neudecker (1999):  $I_n$  is the  $n \times n$  identity matrix,  $A'$  is the transpose of  $A$ , the  $n$ -vector  $\mathbb{1}_n := (1, \dots, 1)'$ , and  $0_{n \times m}$  denotes the zero matrix with  $n$  rows and  $m$  columns. The unit  $n$ -vector  $e_i^n := (0, \dots, 0, 1, 0, \dots, 0)'$  has a 1 in the  $i$ th coordinate and zeroes elsewhere. If  $A = [a_{ij}]$  is an  $n \times m$  matrix,  $\text{vec } A$  is the vector of columns of  $A$  stacked on top of each other. The *Kronecker product* of an  $m \times n$  matrix  $A$  with a  $p \times q$  matrix  $B$  is the  $mp \times nq$  matrix

$$A \otimes B := \begin{pmatrix} A_{1,1}B & \cdots & A_{1,n}B \\ \vdots & \ddots & \vdots \\ A_{m,1}B & \cdots & A_{m,n}B \end{pmatrix}.$$

We use the following properties of Kronecker products (see Chapter 2 of Magnus and Neudecker (1999)):  $(A \otimes B)(C \otimes D) = (AC \otimes BD)$  for all appropriately sized  $A, B, C, D$  and  $(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1})$  for invertible  $A$  and  $B$ .

If  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is differentiable at  $c$  then the *partial derivative* of  $f_i$  w.r.t. the  $j$ th coordinate at  $c$  is denoted  $D_j f_i(c)$  and is often<sup>5</sup> also written as  $[\partial f_i / \partial x_j]_{x=c}$ . The  $m \times n$  matrix of partial derivatives of  $f$  is the *Jacobian* of  $f$  and denoted

$$(Df(c))_{i,j} := D_j f_i(c) \quad \text{for } i \in [m], j \in [n].$$

The *inverse function theorem* relates the Jacobians of a function and its inverse (cf. Fleming (1977, §4.5)):

**Theorem 12** *Let  $S \subset \mathbb{R}^n$  be an open set and  $g : S \rightarrow \mathbb{R}^n$  be a  $C^q$  function with  $q \geq 1$  (i.e., continuous with at least one continuous derivative). If  $Dg(s) \neq 0$  then: there exists an open set  $S_0$  such that  $s \in S_0$  and the restriction of  $g$  to  $S_0$  is invertible;  $g(S_0)$  is open;  $f$ , the inverse of the restriction of  $g$  to  $S_0$ , is  $C^q$ ; and  $Df(t) = [Dg(s)]^{-1}$  for  $t = g(s)$  and  $s \in S_0$ .*

If  $F$  is a matrix valued function  $DF(X) := Df(\text{vec } X)$  where  $f(X) = \text{vec } F(X)$ .

We will require the product rule for matrix valued functions (Fackler, 2005): Suppose  $f : \mathbb{R}^n \rightarrow \mathbb{R}^{m \times p}$ ,  $g : \mathbb{R}^n \rightarrow \mathbb{R}^{p \times q}$  so that  $(f \times g) : \mathbb{R}^n \rightarrow \mathbb{R}^{m \times q}$ . Then

$$D(f \times g)(x) = (g(x)' \otimes I_m) \cdot Df(x) + (I_q \otimes f(x)) \cdot Dg(x). \quad (23)$$

The *Hessian* at  $x \in X \subseteq \mathbb{R}^n$  of a real-valued function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is the  $n \times n$  real, symmetric matrix of second derivatives at  $x$

$$(Hf(x))_{j,k} := D_{k,j} f(x) = \frac{\partial^2 f}{\partial x_k \partial x_j}.$$

---

5. See Chapter 9 of Magnus and Neudecker (1999) for why the  $\partial/\partial x$  notation is a poor one for multivariate differential calculus despite its popularity.

Note that the derivative  $D_{k,j}$  is in row  $j$ , column  $k$ . It is easy to establish that the Jacobian of the transpose of the Jacobian of  $f$  is the Hessian of  $f$ . That is,

$$\mathbf{H}f(x) = \mathbf{D}((\mathbf{D}f(x))') \quad (24)$$

(cf. Chapter 10 of (Magnus and Neudecker, 1999)). If  $f : X \rightarrow \mathbb{R}^m$  for  $X \subseteq \mathbb{R}^n$  is a vector valued function then the Hessian of  $f$  at  $x \in X$  is the  $mn \times n$  matrix that consists of the Hessians of the functions  $f_i$  stacked vertically:

$$\mathbf{H}f(x) := \begin{pmatrix} \mathbf{H}f_1(x) \\ \vdots \\ \mathbf{H}f_m(x) \end{pmatrix}.$$

The following theorem regarding the chain rule for Hessian matrices can be found in (Magnus and Neudecker, 1999, pg. 110).

**Theorem 13** *Let  $S$  be a subset of  $\mathbb{R}^n$ , and  $f : S \rightarrow \mathbb{R}^m$  be twice differentiable at a point  $c$  in the interior of  $S$ . Let  $T$  be a subset of  $\mathbb{R}^m$  containing  $f(S)$ , and  $g : T \rightarrow \mathbb{R}^p$  be twice differentiable at the interior point  $b = f(c)$ . Then the function  $h(x) := g(f(x))$  is twice differentiable at  $c$  and*

$$\mathbf{H}h(c) = (I_p \otimes \mathbf{D}f(c))' \cdot (\mathbf{H}g(b)) \cdot \mathbf{D}f(c) + (\mathbf{D}g(b) \otimes I_n) \cdot \mathbf{H}f(c).$$

Applying the chain rule to functions that are inverses of each other gives the following corollary.

**Corollary 14** *Suppose  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is invertible with inverse  $g := f^{-1}$ . If  $b = f(c)$  then*

$$\mathbf{H}f^{-1}(b) = - (G \otimes G') \mathbf{H}f(c) G$$

where  $G := [\mathbf{D}f(c)]^{-1} = \mathbf{D}g(b)$ .

**Proof** Since  $f \circ g = \text{id}$  and  $\mathbf{H}[\text{id}] = 0_{n^2 \times n}$  Theorem 13 implies that for  $c$  in the interior of the domain of  $f$  and  $b = f(c)$

$$\mathbf{H}(g \circ f)(c) = (I_n \otimes \mathbf{D}f(c))' \cdot \mathbf{H}g(b) \cdot \mathbf{D}f(c) + (\mathbf{D}g(b) \otimes I_n) \cdot \mathbf{H}f(c) = 0_{n^2 \times n}.$$

Solving this for  $\mathbf{H}g(b)$  gives

$$\mathbf{H}g(b) = - [(I_n \otimes \mathbf{D}f(c))']^{-1} (\mathbf{D}g(b) \otimes I_n) \cdot \mathbf{H}f(c) \cdot [\mathbf{D}f(c)]^{-1}.$$

Since  $(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1})$  and  $(A')^{-1} = (A^{-1})'$  we have  $[(I \otimes B)']^{-1} = [(I \otimes B)^{-1}]' = (I^{-1} \otimes B^{-1})' = (I \otimes B^{-1})'$  so the first term in the above product simplifies to  $-[(I_n \otimes \mathbf{D}f(c)^{-1})']$ . The inverse function theorem implies  $\mathbf{D}g(b) = [\mathbf{D}f(c)]^{-1} =: G$  and so

$$\begin{aligned} \mathbf{H}g(b) &= -(I_n \otimes G)' \cdot (G \otimes I_n) \cdot \mathbf{H}f(c) \cdot G \\ &= -(G \otimes G') \cdot \mathbf{H}f(c) \cdot G \end{aligned}$$

as required, since  $(A \otimes B)(C \otimes D) = (AC \otimes BD)$ . ■

## Appendix B. Mixability of the Brier Score

The  $n$ -class Brier score is<sup>6</sup>

$$\ell_{\text{Brier}}(y, \hat{p}) = \sum_{i=1}^n (\llbracket y_i = 1 \rrbracket - \hat{p}_i)^2,$$

where  $y \in \{0, 1\}^n$  and  $\hat{p} \in \Delta^n$ . Thus

$$L_{\text{Brier}}(p, \hat{p}) = \sum_{i=1}^n \mathbb{E}_{Y \sim p} (\llbracket Y_i = 1 \rrbracket - \hat{p}_i)^2 = \sum_{i=1}^n (p_i - 2p_i\hat{p}_i + \hat{p}_i^2).$$

Hence  $\underline{L}_{\text{Brier}}(p) = L_{\text{Brier}}(p, p) = \sum_{i=1}^n (p_i - 2p_i p_i + p_i^2) = 1 - \sum_{i=1}^n p_i^2$  since  $\sum_{i=1}^n p_i = 1$ , and  $\tilde{L}_{\text{Brier}}(\tilde{p}) = 1 - \sum_{i=1}^{n-1} p_i^2 - \left(1 - \sum_{i=1}^{n-1} p_i\right)^2$ .

As first proved by Vovk and Zhdanov (2009), the Brier score is mixable with mixability constant 1. We will reprove this result using the following restatement of Lemma 9:

**Lemma 15** *Let  $\ell$  be a twice differentiable, strictly proper loss, with Bayes risk  $\tilde{\underline{L}}(\tilde{p}) := \underline{L}(p)$ . Let  $\tilde{\underline{L}}_{\log}(\tilde{p}) := \underline{L}_{\log}(p)$  be the Bayes risk for the log loss. Then the following statements are equivalent:*

- (i.)  $\ell$  is  $\beta$ -mixable;
- (ii.)  $\beta\underline{L}(p) - \underline{L}_{\log}(p)$  is convex;
- (iii.)  $\beta\tilde{\underline{L}}(\tilde{p}) - \tilde{\underline{L}}_{\log}(\tilde{p})$  is convex.

**Proof** Equivalence of (i) and (iii) follows from Lemma 9 upon observing that  $\beta\tilde{\underline{L}}(\tilde{p}) - \tilde{\underline{L}}_{\log}(\tilde{p})$  is convex if and only if  $\beta H\tilde{\underline{L}}(\tilde{p}) \succcurlyeq H\tilde{\underline{L}}_{\log}(\tilde{p})$  (Hiriart-Urruty and Lemaréchal, 1993). Equivalence of (ii) and (iii) follows by linearity of the map  $p_n(\tilde{p}) = 1 - \sum_{i=1}^{n-1} \tilde{p}_i$ . ■

**Theorem 16** *The Brier score is mixable, with mixability constant  $\beta_{\text{Brier}} = 1$ .*

**Proof** It can be verified by basic calculus that  $\ell_{\text{Brier}}$  is twice differentiable. To see that it is strictly proper, note that for  $\hat{p} \neq p$  the inequality  $L_{\text{Brier}}(p, \hat{p}) > \underline{L}_{\text{Brier}}(p)$  is equivalent to

$$\sum_{i=1}^n (p_i^2 - 2p_i\hat{p}_i + \hat{p}_i^2) > 0 \quad \text{or} \quad \sum_{i=1}^n (p_i - \hat{p}_i)^2 > 0,$$

and the latter inequality is true because  $p_i \neq \hat{p}_i$  for at least one  $i$  by assumption. Hence the conditions of Lemma 15 are satisfied.

We will first prove that  $\beta_{\text{Brier}} \leq 1$  by showing that convexity of  $\beta\tilde{\underline{L}}_{\text{Brier}}(\tilde{p}) - \tilde{\underline{L}}_{\log}(\tilde{p})$  implies  $\beta \leq 1$ . If  $\beta\tilde{\underline{L}}_{\text{Brier}}(\tilde{p}) - \tilde{\underline{L}}_{\log}(\tilde{p})$  is convex, then it is convex as a function of  $p_1$  when

---

6. This is the definition used by Vovk and Zhdanov (2009). Cesa-Bianchi and Lugosi (2006) use a different definition (for the binary case) which differs by a constant. Their definition results in  $\tilde{\underline{L}}(\tilde{p}) = \tilde{p}(1-\tilde{p})$  and thus  $\tilde{\underline{L}}''(\tilde{p}) = -2$ . If  $n = 2$ , then  $\tilde{\underline{L}}_{\text{Brier}}$  as defined above leads to  $\tilde{\underline{L}}''_{\text{Brier}}(\tilde{p}) = H\tilde{\underline{L}}_{\text{Brier}}(\tilde{p}) = -2(1+1) = -4$ .

all other elements of  $\tilde{p}$  are kept fixed. Consequently, the second derivative with respect to  $p_1$  must be nonnegative:

$$0 \leq \frac{\partial^2}{\partial p_1^2} \left( \beta \tilde{L}_{\text{Brier}}(\tilde{p}) - \tilde{L}_{\log}(\tilde{p}) \right) = \frac{1}{p_1} + \frac{1}{p_n} - 4\beta.$$

By evaluating at  $p_1 = p_n = 1/2$ , it follows that  $\beta \leq 1$ .

It remains to show that  $\beta \underline{L}_{\text{Brier}}(p) \geq \underline{L}_{\log}(p)$ . By Lemma 15 it is sufficient to show that, for  $\beta \leq 1$ ,  $\beta \underline{L}_{\text{Brier}}(p) - \underline{L}_{\log}(p)$  is convex. We proceed by induction. For  $n = 1$ , the required convexity holds trivially. Suppose the lemma holds for  $n - 1$ , and let  $f_n(p_1, \dots, p_n) = \beta \underline{L}_{\text{Brier}}(p) - \underline{L}_{\log}(p)$  for all  $n$ . Then for  $n \geq 2$

$$f_n(p_1, \dots, p_n) = f_{n-1}(p_1 + p_2, p_3, \dots, p_n) + g(p_1, p_2),$$

where  $g(p_1, p_2) = -\beta p_1^2 - \beta p_2^2 + \beta(p_1 + p_2)^2 + p_1 \ln p_1 + p_2 \ln p_2 - (p_1 + p_2) \ln(p_1 + p_2)$ . As  $f_{n-1}$  is convex by inductive assumption and the sum of two convex functions is convex, it is therefore sufficient to show that  $g(p_1, p_2)$  is convex or, equivalently, that its Hessian is positive semi-definite. Abbreviating  $q = p_1 + p_2$ , we have that

$$\mathsf{H}g(p_1, p_2) = \begin{pmatrix} 1/p_1 - 1/q & 2\beta - 1/q \\ 2\beta - 1/q & 1/p_2 - 1/q \end{pmatrix}.$$

A  $2 \times 2$  matrix is positive semi-definite if its trace and determinant are both non-negative, which is easily verified in the present case:  $\text{Tr}(\mathsf{H}g(p_1, p_2)) = 1/p_1 + 1/p_2 - 2/q \geq 0$  and  $|\mathsf{H}g(p_1, p_2)| = (1/p_1 - 1/q)(1/p_2 - 1/q) - (2\beta - 1/q)^2$ , which is non-negative if

$$\begin{aligned} \frac{1}{p_1 p_2} - \frac{1}{p_1 q} - \frac{1}{p_2 q} &\geq 4\beta^2 - \frac{4\beta}{q} \\ 0 &\geq 4\beta^2 q - 4\beta \\ \beta q &\leq 1. \end{aligned}$$

As  $q = p_1 + p_2 \leq 1$ , this inequality holds for  $\beta \leq 1$ , which shows that  $g(p_1, p_2)$  is convex and thereby completes the proof.  $\blacksquare$

VAN ERVEN, REID AND WILLIAMSON

# Distribution-Independent Evolvability of Linear Threshold Functions

**Vitaly Feldman**

*IBM Almaden Research Center*

VITALY@POST.HARVARD.EDU

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

Valiant’s model of evolvability models the evolutionary process of acquiring useful functionality as a restricted form of learning from random examples (Valiant, 2009). Linear threshold functions and their various subclasses, such as conjunctions and decision lists, play a fundamental role in learning theory and hence their evolvability has been the primary focus of research on Valiant’s framework. One of the main open problems regarding the model is whether conjunctions are evolvable distribution-independently (Feldman and Valiant, 2008). We show that the answer is negative. Our proof is based on a new combinatorial parameter of a concept class that lower-bounds the complexity of learning from correlations.

We contrast the lower bound with a proof that linear threshold functions having a non-negligible margin on the data points are evolvable distribution-independently via a simple mutation algorithm. Our algorithm relies on a non-linear loss function being used to select the hypotheses instead of 0-1 loss in Valiant’s original definition. The proof of evolvability requires that the loss function satisfies several mild conditions that are, for example, satisfied by the quadratic loss function studied in several other works (Michael, 2007; Feldman, 2009b; Valiant, 2011). An important property of our evolution algorithm is monotonicity, that is the algorithm guarantees evolvability without any decreases in performance. Previously, monotone evolvability was only shown for conjunctions with quadratic loss (Feldman, 2009b) or when the distribution on the domain is severely restricted (Michael, 2007; Feldman, 2009b; Kanade et al., 2010).

**Keywords:** Evolvability, statistical query dimension, conjunction, halfspace, linear threshold function, quadratic loss

## 1. Introduction

Evolution is the source of the spectacularly complex organisms and behavior that we see around us. Yet we know very little about the computational mechanisms that can lead to such complexity while respecting the constraints of the Darwinian evolutionary process and using a plausible amount of resources. Recently Valiant (2009) proposed that an appropriate framework for understanding the power of evolution to produce complex behavior is that of computational learning theory since both evolution and learning involve processes that adapt their behavior on the basis of experience. Accordingly, in his model, evolvability of a certain useful functionality is cast as a problem of learning the desired functionality through a process in which, at each step, the most “fit” candidate function is chosen from a small pool of mutations of the current candidate. Limits on the number of steps and the amount

of computation performed at each step are imposed to make this process naturally plausible. A class of functions  $C$  is considered evolvable if there exists a single representation scheme  $R$  and a mutation algorithm  $M$  on  $R$  that, when guided by such selection, guarantees convergence to the desired function for every function in  $C$ . Here the requirements closely follow those of the celebrated PAC learning model (Valiant, 1984). In fact, every evolution algorithm (here and below in the sense defined in Valiant’s model) can be simulated by an algorithm that is given random examples of the desired function. In addition, many properties of learning algorithms such as distribution-independence, weakness and attribute-efficiency apply equally to evolvability.

### 1.1. Prior Work

The constrained way in which evolution algorithms have to converge to the target function makes finding such algorithms a substantially more involved task than designing PAC learning algorithms. Initially, only the evolvability of monotone conjunctions of Boolean variables, and only when the distribution over the domain is uniform, was demonstrated (if not specified otherwise, the domain is  $\{0, 1\}^n$ ) (Valiant, 2009). Subsequently this result was simplified (Diogenes and Turán, 2009) and strengthened to general conjunctions (Jacobson, 2007; Kanade et al., 2010). Later Michael (2007) described an algorithm for evolving decision lists over the uniform distribution that used a larger space of hypotheses and a different performance metric over hypotheses (specifically, quadratic loss). In our earlier work we showed that evolvability is, at least within polynomial limits, equivalent to learning by a natural restriction of well-studied statistical queries (SQ)(Kearns, 1998), referred to as *correlational statistical queries* (CSQ) (Feldman, 2008). This result gives distribution-specific algorithms for any SQ learnable class of functions. By characterizing weak distribution-independent evolvability and using communication-complexity-based lower bounds (Sherstov, 2007; Buhrman et al., 2007), we also proved that general linear threshold functions (also referred to as halfspaces) and even decision lists are not evolvable distribution-independently.

In another work (Feldman, 2009a) we examined the relative power of a number of variants of the model discussed in Valiant’s and other works (Valiant, 2009). Among them we considered a generalization of the model to real-valued hypotheses for which one needs to specify the loss function used to measure the loss in performance at every point. We demonstrated that a number of variants of the model are all equivalent to learning by CSQs and hence to the original model (Feldman, 2009a). The only two properties which we found to influence the power of the model are the choice of the loss function (with the original 0/1 loss being equivalent to evolving with the linear loss) and monotonicity, or requirement that the performance of hypotheses does not decrease in the course of evolution. Valiant’s original selection rule allows small decreases in performance<sup>1</sup>. This somewhat unnatural property has been exploited in all the results showing equivalence to learning by CSQs<sup>2</sup> and hence evolution algorithms obtained through such general transformations are non-monotone. In a recent work Kanade et al. (2010) show that the equivalence to learning

---

1. In this context we refer to empirical performance rather than true expected performance.  
2. The decreases in performance can be avoided if the evolution algorithm starts in a certain fixed state, i.e. is initialized.

Distribution	Concept class	Loss	Monot.	References
Uniform	monotone conjunctions	Boolean	yes	(Valiant, 2009; Kanade et al., 2010)
Uniform	conjunctions	Boolean	yes	(Jacobson, 2007; Kanade et al., 2010)
any spherically symmetric/product normal	homogeneous LTFs	Boolean	yes	(Kanade et al., 2010)
All	Single points	Boolean	yes	(Feldman, 2009a)
Any family $\mathcal{D}$	any CSQ learnable over $\mathcal{D}$	Boolean	no	(Feldman, 2008)
Fixed $D$	any SQ learnable over $D$	Boolean	no	(Feldman, 2008)
Uniform	decision lists	quadratic	yes	(Michael, 2007)
All	conjunctions	quadratic	yes	(Feldman, 2009b)
Any family $\mathcal{D}$	any SQ learnable over $\mathcal{D}$	quadratic	no	(Feldman, 2009a)
Fixed $D$	any SQ learnable over $D$	quadratic	yes	(Feldman, 2009b)

Table 1: Positive results on evolvability. For the distribution entry “All” refers to distribution-independent evolvability.  $\mathcal{D}$  refers to any fixed set of distribution (including “All”). All results for Boolean loss also apply to all other loss functions.

by CSQs still holds if the total allowed decrease in performance is bounded by any non-negligible value chosen in advance (they refer to such algorithms as *quasi-monotone*). The first general transformation that yields monotone algorithms was given in our subsequent work (Feldman, 2009b) where we showed that every concept class SQ learnable over a fixed distribution  $D$  is evolvable monotonically over  $D$  when using quadratic loss. By exploiting some of the techniques of the general transformation, we also showed that conjunctions are evolvable distribution-independently when using quadratic loss (Feldman, 2009b). We summarize these results and several other known evolution algorithms in Table 1.

## 1.2. Our Results

As can be seen from Table 1, evolvability of even the most basic concept classes is still only partially understood. Most notably, prior to this work it was unknown whether conjunctions are evolvable distribution-independently with Boolean loss (even without requiring monotonicity) and this question was posed by Valiant and the author as an open problem at COLT 2008 (Feldman and Valiant, 2008). In our first result (Section 3) we show that the answer is negative. Specifically, we prove that for any  $k = \omega(1)$ , monotone conjunctions of at most  $k$  variables are not evolvable distribution-independently to any accuracy  $\epsilon = o(1)$ . Our technique is based on a new combinatorial parameter of a concept class that, roughly, measures the maximum number of correlational query functions required to distin-

guish every target function-distribution pair from a fixed function-distribution pair. This general approach is based on our recent characterization of strong SQ learnability (Feldman, 2009b). For a given size of conjunction  $k$ , we then come up with a construction of a set of conjunction-distribution pairs  $\{(t_S, D_S) \mid |S| = k\}$  that cannot be distinguished from a constant function over the uniform distribution using a polynomial number of queries. The distribution  $D_S$  is designed in such a way that it hides all Fourier coefficients of the conjunction  $t_S$  up to degree  $k/3$ . Simple facts from Fourier analysis of Boolean functions then imply that distinguishing between a superpolynomial number of such conjunction-distribution pairs is impossible using a polynomial number of queries.

We interpret this negative result as highlighting significant limitations of evolvability based on the Boolean feedback only. We note that many functions in biological evolution are not Boolean. For example, for most genes the amount of gene expression (that is the amount of protein produced) can vary in a certain range continuously (up to, of course, the granularity of a single molecule). Therefore it is natural to assume that when evolving the optimal regulation of gene expression (described by a Boolean function), intermediate amounts of the protein will be produced. The intermediate values are likely to cause intermediate values of loss relative to the optimal 0 or 1 value. It is therefore important to understand evolvability with other loss functions. Toward this goal in Section 4 we show that linear threshold functions are evolvable monotonically and distribution-independently for quadratic loss function and all other loss functions satisfying a set of mild conditions. We refer to loss functions that satisfy the required conditions as *well-behaved*. The amount of resources required by our algorithm depends quadratically on  $1/\gamma$  where  $\gamma$  is the margin of the target halfspace on the data points. Therefore, like the famous Perceptron and Winnow algorithms (Rosenblatt, 1958; Littlestone, 1987), it is efficient only when the margin is non-negligible or lower bounded by the inverse of a polynomial in  $n$ . In the Support Vector Machine (SVM) literature this condition is usually referred to as having a *large* margin. Further, the representation used by our evolution algorithm is similar to linear thresholds and the mutation algorithm is fairly simple and natural. The only operations it requires are adding the function  $\alpha \cdot x_i$  to the current function for a real  $\alpha$  and bounding the value of the function to be in  $[-1, 1]$ .

A very popular and powerful approach to learning when data points are not linearly separable is to embed the data points in a different (often higher dimensional) Euclidean space where the examples become linearly separable and then use a halfspace learning algorithm such as Perceptron or SVM to produce a classifier. Such approach also works in the context of evolvability and implies monotone evolvability of any concept class that can be efficiently embedded into large-margin halfspaces over some Euclidean space (efficiency of the embedding also bounds the dimension of the space). Therefore our second result approaches some of the most important and strongest results for PAC learning while also being a natural algorithm in Valiant's framework of evolvability.

We note that a similar mutation algorithm was used in our result for conjunctions (Feldman, 2009b). However our analysis here is new and differs conceptually from the analysis for conjunctions which cannot be extended to halfspaces. It also gives substantially stronger bounds. For example, it improves the dependence of the improvement in each step on  $\epsilon$  from  $\epsilon^6$  to  $\epsilon^2$ . The key to this result for the quadratic loss function is a simple proof that for every distribution  $D$ , halfspace  $f$  and any real-valued function  $\phi$  with the range in

$[-1, 1]$ , there exists a variable  $x_i$  that is correlated with the gradient of the loss function at point  $\phi$ . The absolute value of the correlation is lower-bounded by the inverse of a polynomial in  $n, 1/\epsilon$  and  $1/\gamma$  and therefore is sufficient to imply that a small step in the direction of  $x_i$  (or  $-x_i$ ) will reduce the loss.

A recent work by P. Valiant (2011) examines the extension of the model of evolvability to real-valued target functions. His results paint a picture quite similar to what we know about the evolvability of Boolean functions. In particular, his simple algorithm for evolving linear functions when using the quadratic loss can be seen as the counterpart of our algorithm for halfspaces.

## 2. Preliminaries

For a positive integer  $\ell$ , let  $[\ell]$  denote the set  $\{1, 2, \dots, \ell\}$  and for  $i \leq \ell$  let  $[i..]\ell$  denote the set  $\{i, i+1, \dots, \ell\}$ . We denote the domain of our learning problems by  $X$ . As usual it is parameterized by an (implicit) dimension  $n$ . A *concept class* over  $X$  is a set of  $\{-1, 1\}$ -valued functions over  $X$  referred to as *concepts*. Let  $\mathcal{F}_1^\infty$  denote the set of all functions from  $X$  to  $[-1, 1]$  (that is all the functions with  $L_\infty$  norm bounded by 1). It will be convenient to view a distribution  $D$  over  $X$  as defining the product  $\langle \phi, \psi \rangle_D = \mathbf{E}_{x \sim D}[\phi(x) \cdot \psi(x)]$  over the space of real-valued functions on  $X$ . It is easy to see that this is simply a non-negatively weighted version of the standard dot product over  $\mathbb{R}^X$  and hence is a positive semi-inner product over  $\mathbb{R}^X$ . The corresponding norm is defined as  $\|\phi\|_D = \sqrt{\mathbf{E}_D[\phi^2(x)]} = \sqrt{\langle \phi, \phi \rangle_D}$ .

Let  $B_n = \{x \mid \|x_i\| \leq 1\}$  be the ball of radius 1 in  $\mathbb{R}^n$ ,  $X$  be a subset of  $B_n$ , and  $f = \text{sign}(\sum w_i x_i - \theta)$  be a linear threshold function (halfspace). We define the margin  $\gamma$  of  $f$  on  $X$  as  $\gamma = \inf_{x \in X} \{|\sum w_i x_i - \theta|\}$ . For convenience we use  $x_0$  to refer to the constant function 1.

### 2.1. PAC Learning

The models we consider are based on the well-known PAC learning model introduced by Valiant (1984). Let  $C$  be a concept class over  $X$ . In the basic PAC model a learning algorithm is given examples of an unknown function  $f$  from  $C$  on points randomly chosen from some unknown distribution  $D$  over  $X$  and should produce a hypothesis  $h$  that approximates  $f$ . Formally, an *example oracle*  $\text{EX}(f, D)$  is an oracle that upon being invoked returns an example  $\langle x, f(x) \rangle$ , where  $x$  is chosen randomly with respect to  $D$ , independently of any previous examples.

An algorithm is said to PAC learn  $C$  in time  $t$  if for every  $\epsilon > 0$ ,  $f \in C$ , and distribution  $D$  over  $X$ , the algorithm given  $\epsilon$  and access to  $\text{EX}(f, D)$  outputs, in time  $t$  and with probability at least  $2/3$ , a hypothesis  $h$  that is evaluable in time  $t$  and satisfies  $\mathbf{Pr}_D[f(x) \neq h(x)] \leq \epsilon$ . We say that an algorithm *efficiently* learns  $C$  when  $t$  is upper bounded by a polynomial in  $n, 1/\epsilon$ .

The basic PAC model is also referred to as *distribution-independent* learning to distinguish it from *distribution-specific* PAC learning in which the learning algorithm is required to learn only with respect to a single distribution  $D$  known in advance. More generally, following Kearns et al. (1994), one can analogously define the learnability of a set of distribution-function pairs over the same domain  $X$ . Namely, a set of distribution-function pairs  $\mathcal{Z}$  is

PAC learnable if there exists a learning algorithm that learns  $f$  over  $D$  (as in the definition above) for every  $(D, f) \in \mathcal{Z}$ .

A *weak* learning algorithm (Kearns and Valiant, 1994) is a learning algorithm that produces a hypothesis whose disagreement with the target concept is noticeably less than  $1/2$  (and not necessarily any  $\epsilon > 0$ ). More precisely, a weak learning algorithm produces a Boolean hypothesis  $h$  such that  $\mathbf{Pr}_D[f(x) \neq h(x)] \leq 1/2 - 1/p(n)$  for some fixed polynomial  $p$ .

## 2.2. The Statistical Query Learning Model

In the *statistical query model* of Kearns (1998) the learning algorithm is given access to  $\text{STAT}(f, D)$  – a *statistical query oracle* for target concept  $f$  with respect to distribution  $D$  instead of  $\text{EX}(f, D)$ . A query to this oracle is a function  $\psi : X \times \{-1, 1\} \rightarrow \{-1, 1\}$ . The oracle may respond to the query with any value  $v$  satisfying  $|\mathbf{E}_D[\psi(x, f(x))] - v| \leq \tau$  where  $\tau \in [0, 1]$  is a real number called the *tolerance* of the query. An algorithm  $\mathcal{A}$  is said to learn  $C$  in time  $t$  from statistical queries of tolerance  $\tau$  if  $\mathcal{A}$  PAC learns  $C$  using  $\text{STAT}(f, D)$  in place of the example oracle. In addition, each query  $\psi$  made by  $\mathcal{A}$  has tolerance  $\tau$  and can be evaluated in time  $t$ .

The algorithm is said to (efficiently) SQ learn  $C$  if  $t$  is polynomial in  $n$  and  $1/\epsilon$ , and  $\tau$  is lower-bounded by the inverse of a polynomial in  $n$  and  $1/\epsilon$ .

A *correlational* statistical query is a statistical query for a correlation of a function over  $X$  with the target (Bshouty and Feldman, 2002). Namely the query function  $\psi(x, \ell) \equiv \phi(x) \cdot \ell$  for a function  $\phi \in \mathcal{F}_1^\infty$ . A concept class is said to be CSQ learnable if it is learnable by a SQ algorithm that uses only CSQ queries.

## 2.3. Evolvability

We start by presenting a brief overview of the model. For a detailed description and intuition behind the various choices made in model the reader is referred to (Valiant, 2009; Feldman, 2009a). The goal of the model is to specify how organisms can acquire complex mechanisms via a resource-efficient process based on random mutations and guided by performance-based selection. The mechanisms are described in terms of the multi argument functions they implement. The performance of such a mechanism is measured by evaluating the agreement of the mechanism with some “ideal” behavior function. The value of the “ideal” function on some input describes the most beneficial behavior for the condition represented by the input. The evaluation of the agreement with the “ideal” function is derived by evaluating the function on a moderate number of inputs drawn from a probability distribution over the conditions that arise. These evaluations correspond to the experiences of one or more organisms that embody the mechanism.

Random variation is modeled by the existence of an explicit algorithm that acts on some fixed representation of mechanisms and for each representation of a mechanism produces representations of mutated versions of the mechanism. The model requires that the mutation algorithm be efficiently implementable. Selection is modeled by an explicit rule that determines the probabilities with which each of the mutations of a mechanism will be chosen to “survive” based on the performance of all the mutations of the mechanism and the probabilities with which each of the mutations is produced by the mutation algorithm.

As can be seen from the above description, a performance landscape (given by a specific “ideal” function and a distribution over the domain), a mutation algorithm, and a selection rule jointly determine how each step of an evolutionary process is performed. A class of functions  $C$  is considered evolvable if there exist a representation of mechanisms  $R$  and a mutation algorithm  $M$  such that for every “ideal” function  $f \in C$ , a sequence of evolutionary steps starting from any representation in  $R$  and performed according to the description above “converges” in a polynomial number of steps to  $f$ . This process is essentially PAC learning of  $C$  with the selection rule (rather than explicit examples) providing the only target-specific feedback. We now define the model formally using the notation from (Feldman, 2009a).

## 2.4. Definition of Evolvability

The description of an evolution algorithm  $\mathcal{A}$  consists of the definition of the representation class  $R$  of possibly randomized hypotheses in  $\mathcal{F}_1^\infty$  and the description of polynomial time mutation algorithm that for every  $r \in R$  and  $\epsilon > 0$  outputs a random mutation of  $r$ .

**Definition 1** A evolution algorithm  $\mathcal{A}$  is defined by a pair  $(R, M)$  where

- $R$  is a representation class of functions over  $X$  with range in  $[-1, 1]$ .
- $M$  is a randomized polynomial time algorithm that, given  $r \in R$  and  $\epsilon$  as input, outputs a representation  $r_1 \in R$  with probability  $\mathbf{Pr}_{\mathcal{A}}(r, r_1)$ . The set of representations that can be output by  $M(r, \epsilon)$  is referred to as the neighborhood of  $r$  for  $\epsilon$  and denoted by  $\text{Neigh}_{\mathcal{A}}(r, \epsilon)$ .

A loss function  $L$  on a set of values  $Y$  is a non-negative mapping  $L : Y \times Y \rightarrow \mathbb{R}^+$ .  $L(y, y')$  measures the “distance” between the desired value  $y$  and the predicted value  $y'$ . In the context of learning Boolean functions using hypotheses with values in  $[-1, 1]$  we only consider functions  $L : \{-1, 1\} \times \{-1, 1\} \rightarrow \mathbb{R}^+$ . Valiant’s original model only considers Boolean hypotheses and hence only the disagreement loss (or 0-1 loss) which is equal to  $L_\Delta(y, y') = y \cdot y'$ . It was shown in our earlier work (Feldman, 2009a) that such loss is equivalent to the linear loss  $L_1(y, y') = |y' - y|$  over hypotheses with the range in  $[-1, 1]$ . The other loss function we use here is the quadratic loss  $L_Q(y, y') = (y' - y)^2$  function. For a function  $\phi \in \mathcal{F}_1^\infty$  its performance relative to loss function  $L$ , distribution  $D$  over the domain and target function  $f$  is defined as

$$L\text{Perf}_f(\phi, D) = 1 - 2 \cdot \mathbf{E}_D[L(f(x), \phi(x))] / L(-1, 1) .$$

For an integer  $s$ , functions  $\phi, f \in \mathcal{F}_1^\infty$  over  $X$ , distribution  $D$  over  $X$  and loss function  $L$ , the empirical fitness  $L\text{Perf}_f(\phi, D, s)$  of  $\phi$  is a random variable that equals  $1 - \frac{2}{s} \frac{1}{L(-1, 1)} \sum_{i \in [s]} L(f(z_i), \phi(z_i))$  for  $z_1, z_2, \dots, z_s \in X$  chosen randomly and independently according to  $D$ .

A number of natural ways of modeling selection were discussed in prior work (Valiant, 2009; Feldman, 2009a). For concreteness here we describe the selection rule used in Valiant’s main definition in a slightly generalized version from (Feldman, 2009a). In selection rule  $\text{SelNB}[L, t, p, s]$   $p$  candidate mutations are sampled using the mutation algorithm. Then beneficial and neutral mutations are defined on the basis of their empirical fitness  $L\text{Perf}$

in  $s$  experiments (or examples) using tolerance  $t$ . If some beneficial mutations are available one is chosen randomly according to their relative frequencies in the candidate pool. If none is available then one of the neutral mutations is output randomly according to their relative frequencies. If neither neutral or beneficial mutations are available,  $\perp$  is output to mean that no mutation “survived”.

**Definition 2** For a loss function  $L$ , tolerance  $t$ , candidate pool size  $p$ , sample size  $s$ , selection rule  $\text{SelNB}[L, t, p, s]$  is an algorithm that for any function  $f$ , distribution  $D$ , mutation algorithm  $\mathcal{A} = (R, M)$ , a representation  $r \in R$ , accuracy  $\epsilon$ ,  $\text{SelNB}[L, t, p, s](f, D, \mathcal{A}, r)$  outputs a random variable that takes a value  $r_1$  determined as follows. First run  $M(r, \epsilon)$   $p$  times and let  $Z$  be the set of representations obtained. For  $r' \in Z$ , let  $\Pr_Z(r')$  be the relative frequency with which  $r'$  was generated among the  $p$  observed representations. For each  $r' \in Z \cup \{r\}$ , compute an empirical value of fitness  $v(r') = L\text{Perf}_f(r', D, s)$ . Let  $\text{Bene}(Z) = \{r' \mid v(r') \geq v(r) + t\}$  and  $\text{Neut}(Z) = \{r' \mid |v(r') - v(r)| < t\}$ . Then

- (i) if  $\text{Bene}(Z) \neq \emptyset$  then output  $r_1 \in \text{Bene}$  with probability  $\Pr_Z(r_1) / \sum_{r' \in \text{Bene}(Z)} \Pr_Z(r')$ ;
- (ii) if  $\text{Bene}(Z) = \emptyset$  and  $\text{Neut}(Z) \neq \emptyset$  then output  $r_1 \in \text{Neut}(Z)$  with probability  $\Pr_Z(r_1) / \sum_{r' \in \text{Neut}(Z)} \Pr_Z(r')$ .
- (iii) If  $\text{Neut}(Z) \cup \text{Bene}(Z) = \emptyset$  then output  $\perp$ .

A concept class  $C$  is said to be evolvable by an evolution algorithm  $\mathcal{A}$  guided by a selection rule  $\text{Sel}$  over distribution  $D$  if for every target concept  $f \in C$ , mutation steps as defined by  $\mathcal{A}$  and guided by  $\text{Sel}$  will converge to  $f$ . For simplicity here we only consider the selection rule  $\text{SelNB}$ .

**Definition 3** For concept class  $C$  over  $X$ , distribution  $D$ , mutation algorithm  $\mathcal{A}$ , loss function  $L$  we say that the class  $C$  is evolvable over  $D$  by  $\mathcal{A}$  with  $L$  if there exist polynomials  $1/t(n, 1/\epsilon)$ ,  $s(n, 1/\epsilon)$ ,  $p(n, 1/\epsilon)$  and  $g(n, 1/\epsilon)$  such that for every  $n$ ,  $f \in C$ ,  $\epsilon > 0$ , and every  $r_0 \in R$ , with probability at least  $1 - \epsilon$ , a sequence  $r_0, r_1, r_2, \dots$ , where  $r_i \leftarrow \text{SelNB}[L, t, p, s](f, D, \mathcal{A}, r_{i-1})$  will have  $L\text{Perf}_f(r_{g(n, 1/\epsilon)}, D) > 1 - \epsilon$ .

As in PAC learning, we say that a concept class  $C$  is evolvable if it is evolvable over all distributions by a single evolution algorithm (we emphasize this by saying *distribution-independently* evolvable). Similarly, we say that a class of distribution-function pairs  $\mathcal{Z}$  is evolvable if the evolution algorithm is successful for all pairs  $(D, f) \in \mathcal{Z}$ .

We say that an evolution algorithm  $\mathcal{A}$  evolves  $C$  over  $D$  *monotonically* if with probability at least  $1 - \epsilon$ , for every  $i \leq g(n, 1/\epsilon)$ ,  $L\text{Perf}_f(r_i, D) \geq L\text{Perf}_f(r_0, D)$ , where  $g(n, 1/\epsilon)$  and  $r_0, r_1, r_2, \dots$  are defined as above. Note that since the evolution algorithm can be started in any representation, this is equivalent to requiring that with probability at least  $1 - \epsilon$ ,  $L\text{Perf}_f(r_{i+1}, D) \geq L\text{Perf}_f(r_i, D)$  for every  $i$ .

### 3. Lower Bounds on Distribution-Independent CSQ Learnability

In this section we demonstrate that conjunctions are not evolvable with Boolean loss (or the equivalent linear loss). We obtain this result by exploiting the equivalence of evolvability

with Boolean loss and efficient CSQ learnability. Our technique is based on a combinatorial parameter of a concept class  $C$ , referred to CSQD that lower bounds the complexity of distribution-independent CSQ learning of  $C$ . This parameter can be seen as a generalization of the approximation-based strong statistical query dimension given in our earlier work (Feldman, 2009b) to the distribution-independent setting.

**Definition 4** For a concept class  $C$ , and  $\epsilon, \tau > 0$  we define  $\text{CSQD}(C, \epsilon, \tau)$  as the smallest number  $d$  for which it holds that for every distribution  $D$  and function  $\psi \in \mathcal{F}_1^\infty$ , there exists a set of  $d$  functions  $G_\psi \subset \mathcal{F}_1^\infty$  and a Boolean function  $h_\psi$  such that for every  $f \in C$  and distribution  $D'$ , at least one of the following conditions holds:

1. there exists  $g \in G_\psi$  such that  $|\langle f, g \rangle_{D'} - \langle \psi, g \rangle_D| \geq \tau$  or
2.  $\Pr_{D'}[f(x) \neq h_\psi(x)] \leq \epsilon$ .

We now give a simple proof that  $\text{CSQD}(C, \epsilon, \tau)$  lower bounds the number of correlational statistical queries of tolerance  $\tau$  required to learn  $C$  distribution-independently to accuracy  $\epsilon$ . Our proof is based on the proof of the analogous result for the strong SQ dimension (Feldman, 2009b).

**Theorem 5** If  $C$  is learnable by a deterministic CSQ algorithm that uses  $q(n, 1/\epsilon)$  queries of tolerance  $\tau(n, 1/\epsilon)$  then  $\text{CSQD}(C, \epsilon, \tau(n, 1/\epsilon)) \leq q(n, 1/\epsilon)$ .

**Proof** Let  $\mathcal{A}$  be the assumed CSQ learning algorithm for  $C$ . Let  $\psi \in \mathcal{F}_1^\infty$  be any function and  $D$  be any distribution. The set  $G_\psi$  and function  $h_\psi$  are constructed as follows. Simulate algorithm  $\mathcal{A}$  and for every correlational query  $(\phi_i \cdot \ell, \tau)$  add  $\phi_i$  to  $G_\psi$  and respond with  $\langle \psi, \phi_i \rangle_D = \mathbf{E}_D[\phi_i(x) \cdot \psi(x)]$  to the query. Continue the simulation until  $\mathcal{A}$  outputs a hypothesis. Let  $h_\psi$  be the hypothesis output by  $\mathcal{A}$ .

First, by the definition of  $G_\psi$ ,  $|G_\psi| \leq q(n, 1/\epsilon)$ . Now, let  $f$  be any function in  $C$  and  $D'$  be a distribution. If there does not exist  $g \in G_\psi$  such that  $|\langle f, g \rangle_{D'} - \langle \psi, g \rangle_D| \geq \tau$  (the first condition) then for every correlational query function  $\phi_i \in G_\psi$ ,  $\langle \psi, \phi_i \rangle_D$  is within  $\tau$  of  $\langle f, \phi_i \rangle_{D'}$ . Therefore the answers provided by our simulator are valid for the execution of  $\mathcal{A}$  when the target function is  $f$  and the distribution is  $D'$ . That is they could have been returned by  $\text{STAT}(f, D')$  with tolerance  $\tau$ . Therefore, by the definition of  $\mathcal{A}$ , the hypothesis  $h_\psi$  satisfies  $\Pr_{D'}[f(x) \neq h_\psi(x)] \leq \epsilon$  (the second condition). ■

### 3.1. Conjunctions are not CSQ Learnable Distribution-Independently

We now demonstrate that for a carefully constructed set of distributions and conjunctions, no polynomial-size approximating set satisfying the conditions of Definition 4 exists. Let  $U$  be the uniform distribution over  $X = \{0, 1\}^n$ . For a set  $S \subseteq [n]$  we denote by  $t_S(x)$  a conjunction of the variables with indices in  $S$  and by  $\chi_S(x)$  the parity function of the variables with indices in  $S$ . A well-known fact about the Fourier representation of conjunctions (e.g. Jackson, 1997) is that

$$t_S(x) = -1 + 2^{-|S|+1} \sum_{I \subseteq S} \chi_I(x) .$$

To obtain the desired lower bound we note that any pair  $(D, g)$  where  $g$  is a real-valued function over  $\{0, 1\}^n$  and  $D$  is a distribution can be viewed as a real-valued function  $g'(x) = g(x)D(x)/U(x) = 2^n \cdot g(x)D(x)$ . Here and below  $D(x)$  refers to the probability density function of  $D$ . By definition, for every  $x$ ,  $g(x)D(x) = g'(x)U(x)$  and therefore for any real-valued function  $h$ ,  $\langle h, g \rangle_D = \langle h, g' \rangle_U$ . This simple transformation allows us to view distribution-function pairs as functions over the uniform distribution and vice versa.

The basis of our constructions are functions whose Fourier transform equals to the Fourier transform of  $t_S(x)$  but with all the Fourier coefficients for non-empty sets of size at most  $k/3$  removed (the Fourier coefficient of the empty set is simply the constant term). We claim that these functions can be seen as conjunctions over a close-to-uniform distribution.

**Lemma 6** *Let  $k \geq 9$  be an integer divisible by 3 and let  $S \subset [n]$  be any set of size  $k$ . There exists a function  $\theta_S(x)$  and distribution  $D_S$  such that for every point  $x$ ,  $D_S(x)t_S(x) = U(x)\theta_S(x)$  and in addition*

1.  $\theta_S(x) = \alpha \left( -1 + 2^{-|S|+1} + 2^{-|S|+1} \cdot \sum_{I \subseteq S, |I| > k/3} \chi_I(x) \right)$  for a constant  $\alpha \in [2/3, 2]$ .
2. for every  $x$ ,  $D_S(x)/U(x) \in [1/3, 3]$ .

**Proof** Let  $\phi_S(x) = -1 + 2^{-k+1} + 2^{-k+1} \cdot \sum_{I \subseteq S, |I| > k/3} \chi_I(x)$ , in other words  $t_S$  with all the parities for subsets of size  $i \in [k/3]$  removed. Note that the total number of parities that were removed from  $t_S(x)$  is  $\sum_{i \in [k/3]} \binom{k}{i}$ . Note that  $2^{-k} \sum_{i \in [k/3]} \binom{k}{i}$  upper bounds the probability that in  $k$  flips of a fair coin at most  $k/3$  coins will come out as heads. As is well-known, this probability is a monotone decreasing function of  $k$  and for  $k = 9$  is less than  $1/4$ . This implies that for  $k \geq 9$ ,  $\sum_{i \in [k/3]} \binom{k}{i} < 2^{k-2}$ . Therefore for every  $x$ ,

$$|t_S(x) - \phi_S(x)| \leq 2^{-k+1} \left| \sum_{I \subseteq S, |I| \in [k/3]} \chi_I(x) \right| < 2^{-k+1} \cdot 2^{k-2} = 1/2 .$$

This implies that for every  $x$ ,  $\text{sign}(\phi_S(x)) = \text{sign}(t_S(x))$  and  $L_1(\phi_S) = \mathbf{E}_U[|\phi_S(x)|] \in [1/2, 3/2]$ . Now let  $D_S(x) = U(x) \cdot |\phi_S(x)| / L_1(\phi_S)$  and  $\theta_S(x) = \phi_S(x) / L_1(\phi_S)$ . This definition implies that  $\sum_{x \in X} D_S(x) = \mathbf{E}_U[|\phi_S(x)|] / L_1(\phi_S) = 1$ . Hence  $D_S(x)$  is a valid probability density function over  $\{0, 1\}^n$ . Further,  $D_S(x)t_S(x) = U(x)\theta_S(x)$ . In other words the conjunction  $t_S$  over the distribution  $D_S$  can be viewed as the function  $\theta_S(x)$  over the uniform distribution. Finally, note that  $\alpha = 1/L_1(\phi_S) \in [2/3, 2]$  and  $D_S(x)/U(x) = |\phi_S(x)| / L_1(\phi_S) \in [1/3, 3]$ .  $\blacksquare$

We now establish that the number of monotone conjunctions of  $k$  variables such any two conjunctions share at most  $k/3$  variables is large.

**Lemma 7** *For any integer  $k \in [9..n/2]$  divisible by 3, there exists a set  $\mathcal{S}_k \subseteq 2^{[n]}$ , such that*

- for every  $S \in \mathcal{S}_k$ ,  $|S| = k$ ;
- for every distinct  $S, T \in \mathcal{S}_k$ ,  $|S \cap T| \leq k/3$ ;

- $|\mathcal{S}_k| \geq (n/(8k))^{k/3} + 1$ .

**Proof** There are  $\binom{n}{k}$  different size- $k$  subsets of  $[n]$  and each subset of size  $k$  shares more than  $k/3$  elements with at most  $\binom{k}{k/3} \binom{n-k/3}{2k/3}$  other subsets of size  $k$ . Hence by greedily constructing  $\mathcal{S}_k$  we will obtain at least

$$\frac{\binom{n}{k}}{\binom{k}{k/3} \binom{n-k/3}{2k/3}} = \frac{1}{\binom{k}{k/3}} \cdot \frac{n! \cdot (2k/3)!}{(n-k/3)! \cdot k!} > \frac{1}{2^k} \cdot \frac{n \cdot (n-1) \cdots (n-k/3+1)}{k \cdot (k-1) \cdots (2k/3+1)} \geq 2^{-k} \left(\frac{n}{k}\right)^{k/3} = \left(\frac{n}{8k}\right)^{k/3}$$

subsets. ■

We are now ready to show that conjunctions of superconstant size are not CSQ learnable to subconstant accuracy. Let  $C_k$  denote the concept class of conjunctions of size at most  $k$ .

**Theorem 8** *If  $C_k$  is CSQ learnable to accuracy  $\epsilon \leq 2^{-k}/6$  by a deterministic algorithm that uses  $q$  queries of tolerance  $\tau$  then  $q/\tau^2 \geq (\frac{n}{8k})^{k/3}/16$ .*

**Proof** We apply Theorem 5 to the assumed CSQ algorithm for  $C_k$  and obtain that  $\text{CSQD}(C_k, \epsilon, \tau) \leq q$ . Let  $\psi(x) \equiv \alpha(-1 + 2^{-k+1})$  and  $D$  be the uniform distribution. By the Definition 4, there exists a set  $G$  of  $q$  functions and a Boolean function  $h$  such that for every  $f \in C_k$  and distribution  $D'$  at least one of the following conditions holds:

1.  $\mathbf{Pr}_{D'}[f(x) \neq h(x)] \leq \epsilon$  or
2. there exists  $g \in G$  such that  $|\langle f, g \rangle_{D'} - \langle \psi, g \rangle_U| \geq \tau$ .

Let  $\mathcal{S}_k$  be the set given by Lemma 7 and  $S \in \mathcal{S}_k$ . We apply these conditions to  $f = t_S$  and distribution  $D_S$  defined in Lemma 6 to obtain that  $\mathbf{Pr}_{D_S}[t_S(x) \neq h(x)] \leq \epsilon$  or there exists  $g \in G$  such that  $|\langle t_S, g \rangle_{D_S} - \langle \psi, g \rangle_U| \geq \tau$ . We first consider the implications of the first condition. By our assumption  $\epsilon \leq 2^{-k}/6$ . For any two subsets  $S, T \in \mathcal{S}_k$ ,  $\mathbf{Pr}_U[t_S \neq t_T] > 2^{-k}$ . This implies that if  $\mathbf{Pr}_{D_S}[t_S \neq h] \leq \epsilon$  then

$$\mathbf{Pr}_U[t_S \neq h] = \sum_{t_S(x) \neq h(x)} U(x) \stackrel{(*)}{\leq} \sum_{t_S(x) \neq h(x)} 3 \cdot D_S(x) = 3 \cdot \mathbf{Pr}_{D_S}[t_S \neq h] \leq 3\epsilon \leq 2^{-k}/2 , \quad (1)$$

where (\*) is implied by property 2 in Lemma 6. Further,  $\mathbf{Pr}_U[t_T \neq h] \geq \mathbf{Pr}_U[t_S \neq t_T] - \mathbf{Pr}_U[t_S \neq h] > 2^{-k}/2$  and hence, by the same argument as equation (1),

$$\mathbf{Pr}_{D_T}[t_T \neq h] \geq \mathbf{Pr}_U[t_T \neq h]/3 > (2^{-k}/2)/3 = 2^{-k}/6 \geq \epsilon .$$

In other words,  $h$  can be  $\epsilon$ -close to at most one conjunction  $t_S$  for  $S \in \mathcal{S}_k$ .

Now consider a subset  $S$  for which the second condition holds. By the definition of  $D_S$ ,  $\langle t_S, g \rangle_{D_S} = \langle \theta_S, g \rangle_U$  and therefore the second condition is equivalent to

$$|\langle \theta_S - \psi, g \rangle_U| \geq \tau .$$

We observe that  $\theta_S - \psi = \alpha 2^{-k+1} \cdot \sum_{I \subseteq S, |I|>k/3} \chi_I$  and hence

$$\tau \leq \left| \langle \alpha 2^{-k+1} \cdot \sum_{I \subseteq S, |I| > k/3} \chi_I, g \rangle_U \right| \leq \alpha 2^{-k+1} \cdot \sum_{I \subseteq S, |I| > k/3} |\langle \chi_I, g \rangle_U|. \quad (2)$$

Equation (2) implies that there exists  $I_S \subseteq S$  such that  $|I_S| > k/3$  and

$$|\langle \chi_{I_S}, g \rangle_U| \geq \tau \cdot 2^{k-1}/(\alpha 2^k) \geq \tau/(2 \cdot \alpha) \geq \tau/4.$$

We now need two crucial observations:

1. For distinct  $S, T \in \mathcal{S}_k$ ,  $I_S \neq I_T$ . This is true since  $I_S$  is a subset of size at least  $k/3 + 1$  of  $S$  and  $S$  shares at most  $k/3$  elements with  $T$  (of which  $I_T$  is a subset).
2. For any function  $g \in \mathcal{F}_1^\infty$ , there exist at most  $16/\tau^2$  sets  $I$  such that  $|\langle \chi_I, g \rangle_U| \geq \tau/4$ . This is true since  $\langle \chi_I, g \rangle_U$  is simply the Fourier coefficient of  $g$  with index  $I$  denoted by  $\hat{g}(I)$ . Parseval's identity states that  $\sum_{I \subseteq [n]} \hat{g}(I)^2 = \|g\|_U^2 \leq 1$  and therefore no more than  $16/\tau^2$  Fourier coefficients of  $g$  can be larger than  $\tau/4$ .

Combining these two observations gives that the number of subsets of  $\mathcal{S}_k$  for which the second condition holds is at most  $16 \cdot q/\tau^2$ . By combining this with the fact that the first condition can hold for at most one set in  $\mathcal{S}_k$  we obtain that  $16 \cdot q/\tau^2 \geq |\mathcal{S}_k| - 1 \geq (\frac{n}{8k})^{k/3}$ . ■

**Remark 9** *Theorem 8 also applies to CSQ learning by randomized algorithms since a randomized algorithm for the set of conjunction-distribution pairs we consider can be converted to a non-uniform deterministic algorithm via a standard transformation (Bshouty and Feldman, 2002).*

**Corollary 10** *For any  $k = \omega(1)$  and  $\epsilon = o(1)$ ,  $C_k$  is not evolvable to accuracy  $\epsilon$ .*

Interestingly, conjunctions are known to be weakly CSQ learnable distribution-independently. Therefore Corollary 10 also implies that traditional boosting algorithms (Schapire, 1990; Freund, 1995) cannot be adapted to CSQ learning (and hence evolvability).

#### 4. Evolvability of Halfspaces with Non-Linear Loss Functions

In this section we demonstrate that halfspaces are evolvable distribution-independently for a wide class of loss functions using a polynomial in  $n$ ,  $1/\epsilon$  and  $1/\gamma$  amount of resources. Here  $\gamma$  is the margin of the target halfspace on the domain  $X$  of the learning problem. For example, if we set  $X = \{-1/\sqrt{n}, 1/\sqrt{n}\}^n$  (or the Boolean hypercube scaled to fit in the unit ball  $B_n$ ) then all functions that can be represented by a halfspace with integer weights upper-bounded in absolute value by  $m$ , will have the margin of at least  $1/(nm)$ . Consequently, our result implies that such functions are evolvable distribution-independently over the Boolean hypercube for any  $m$  upper-bounded by a polynomial in  $n$ . This class of functions includes conjunctions, disjunctions, decision lists of length  $O(\log n)$  and majority functions. The mutation algorithm we use is very simple and natural for evolving halfspaces. The only

operations it requires are adding  $\alpha \cdot x_i$  to the current function for a real  $\alpha$  and “clipping” the values of the function outside of  $[-1, 1]$ .

A more general way to describe this result is to take the domain to be  $B_n$  and define the margin relative to the support of the target distribution. Specifically, let  $\text{HS}_\gamma$  denote the set of distribution-function pairs over  $B_n$  such that  $(D, f) \in \text{HS}_\gamma$ , if (and only if)  $f$  can be represented by a halfspace with margin  $\gamma$  on the support of distribution  $D$  (for brevity we use “margin on  $D$ ” to refer to the margin on the support of  $D$ ). For  $X \subseteq B_n$ , we denote by  $\text{HS}_\gamma(X)$  the set of all functions that can be represented by a halfspace with margin  $\gamma$  on  $X$ .

Our proof of evolvability relies on the lemma which proves that for every current hypothesis  $\phi \in \mathcal{F}_1^\infty$ , there exists an efficiently computable and small neighborhood  $N(\phi)$  of  $\phi$  such that for every target halfspace  $f$  with margin  $\gamma$  on distribution  $D$ , if the fitness of  $\phi$  is not  $\epsilon$ -close to the optimum then there exist  $\phi' \in N$  whose fitness is observably higher than the fitness of  $\phi$ . Following Kanade et al. (2010), we refer to such function  $N$  as *strictly beneficial* neighborhood function. Strictly beneficial neighborhood function immediately implies monotone evolvability from any starting function (Feldman, 2009b). To see this observe that for a mutation algorithm that produces a random member of the strictly beneficial neighborhood, every step of the evolution algorithm will increase performance by an inverse-polynomial amount until it reaches  $1 - \epsilon$ . Further, as was observed by Kanade et al. (2010), it also implies evolvability when the target function is allowed to change gradually, or *drift*.

We first show the existence of a strictly beneficial neighborhood function for halfspaces with the quadratic loss function and then examine the conditions on the loss function that allow a similar argument to go through. For  $a \in \mathbb{R}$ , define the “clipping” function

$$P_1(a) \triangleq \begin{cases} a & |a| \leq 1 \\ \text{sign}(a) & \text{otherwise.} \end{cases}$$

**Theorem 11** *For  $\phi(x) \in \mathcal{F}_1^\infty$ , let*

$$N_\alpha(\phi) = \{P_1(\phi + \alpha' \cdot x_i) \mid i \in [0..n], |\alpha'| = \alpha\} \cup \{\phi\}.$$

*For every halfspace  $f$  with margin  $\gamma$  on distribution  $D$  over  $B_n$  and every  $\epsilon > 0$ , there exists  $\phi' \in N_\alpha(\phi)$  for which*

$$\|f - \phi'\|_D^2 \leq \max \{\|f - \phi\|_D^2 - \alpha^2, \epsilon\} ,$$

*where  $\alpha = \frac{\epsilon\gamma}{3\sqrt{n}}$ .*

**Proof** Let  $f = \text{sign}(\sum_{i \in [n]} w_i x_i - \theta)$  be the representation of  $f$  that has margin  $\gamma$  on  $D$ . Note that we can assume that  $|\theta| \leq 1$  since for every point  $x \in B_n$ ,  $\sum_{i \in [n]} w_i x_i \leq \|w\|^2 \cdot \|x\|^2 \leq 1$ . The claim holds if  $\|f - \phi\|_D^2 \leq \epsilon$ . We can therefore assume that  $\|f - \phi\|_D^2 > \epsilon$ . In particular, since for every  $a \in [-2, 2]$ ,  $|a| \geq a^2/2$  we obtain that  $\mathbf{E}_D[|f - \phi|] \geq \epsilon/2$ .

For every  $x$  in the support of  $D$ ,  $f(x) - \phi(x)$  has the same sign as  $f(x)$  and therefore also the same sign as  $\sum_{i \in [n]} w_i x_i - \theta$ . Therefore,

$$\mathbf{E}_D \left[ (f - \phi) \left( \sum_{i \in [n]} w_i x_i - \theta \right) \right] \geq \gamma \mathbf{E}_D [|f - \phi|] \geq \epsilon \gamma / 2. \quad (3)$$

At the same time, using the Cauchy-Schwartz inequality we can obtain

$$\begin{aligned} \mathbf{E}_D \left[ (f - \phi) \left( \sum_{i \in [n]} w_i x_i - \theta \right) \right] &\leq \sum_{i \in [n]} |w_i| \cdot |\mathbf{E}_D[(f - \phi)x_i]| + |\theta| \cdot |\mathbf{E}_D[(f - \phi)]| \\ &\leq \sqrt{\theta^2 + \sum_{i \in [n]} w_i^2} \cdot \sqrt{\sum_{i \in [0..n]} \mathbf{E}_D[(f - \phi)x_i]^2} \\ &\leq \sqrt{2} \sqrt{\sum_{i \in [0..n]} \mathbf{E}_D[(f - \phi)x_i]^2}. \end{aligned} \quad (4)$$

By combining equations (3) and (4) we obtain that

$$\sum_{i \in [0..n]} \mathbf{E}_D[(f - \phi)x_i]^2 \geq (\epsilon \gamma)^2 / 8.$$

From here we can conclude that there exists  $j \in [0..n]$  such that

$$|\mathbf{E}_D[(f - \phi)x_j]| \geq \epsilon \gamma / \sqrt{8(n+1)} \geq \epsilon \gamma / (3\sqrt{n}). \quad (5)$$

Now we claim that a step in the direction of  $x_j$  from  $\phi$  will decrease the distance (in  $\|\cdot\|_D$  norm) to  $f$ . Formally,

**Lemma 12** *Let  $\alpha' = \alpha \cdot \text{sign}(\mathbf{E}_D[(f - \phi)x_j])$ , where  $\alpha = \frac{\epsilon \gamma}{3\sqrt{n}}$  (as defined in the statement of the theorem). Then*

$$\|f - (\phi + \alpha' \cdot x_j)\|_D^2 \leq \|f - \phi\|_D^2 - \alpha'^2.$$

### Proof

$$\|f - (\phi + \alpha' \cdot x_j)\|_D^2 = \|f - \phi\|_D^2 + \alpha'^2 \|x_j\|_D^2 - 2\langle f - \phi, \alpha' \cdot x_j \rangle_D.$$

To obtain the claim it remains to observe that  $\|x_j\|_D^2 \leq 1$  and that

$$2\langle f - \phi, \alpha' \cdot x_j \rangle_D = 2\alpha' \mathbf{E}_D[(f - \phi)x_j] \geq 2\alpha'^2 = 2\alpha^2.$$

■

Now let  $\phi' = P_1(\phi + \alpha' \cdot x_j)$ . If for a point  $x$ ,  $\phi'(x) = \phi(x) + \alpha' \cdot x_j$  then clearly  $f(x) - \phi'(x) = f(x) - (\phi(x) + \alpha' \cdot x_j)$ . Otherwise, if  $|\phi(x) + \alpha' \cdot x_j| > 1$  then  $\phi'(x) = \text{sign}(\phi(x) + \alpha' \cdot x_j)$  and for any value  $f(x) \in \{-1, 1\}$ ,  $|f(x) - \phi'(x)| \leq |f(x) - (\phi(x) + \alpha' \cdot x_j)|$ . This implies that

$$\|(f - \phi')\|_D^2 \leq \|f - (\phi + \alpha' \cdot x_j)\|_D^2 \leq \|f - \phi\|_D^2 - \alpha^2.$$

By definition,  $\phi' \in N_\alpha(\phi)$  and hence we obtain the claimed result.  $\blacksquare$

We now demonstrate that a similar result can be obtained under several mild conditions on the loss function. In essence, we require that the loss function can be well approximated by a linear function with a slope that is not too close to 0. Formally,

**Definition 13** For positive constants  $a, A$  and  $B$  we say that a loss function  $L : \{-1, 1\} \times [-1, 1] \rightarrow \mathbb{R}^+$  is well-behaved with bounds  $a, A, B$  if

1.  $L(-1, -1) = L(1, 1) = 0$ ;
2.  $L(1, -1) = L(-1, 1) = 2$ ;
3. for  $\ell \in \{-1, 1\}$ ,  $L(\ell, z)$  is twice differentiable in  $[-1, 1]$  (the differentiation is always in the second variable);
4. for  $\ell \in \{-1, 1\}$ ,  $L'(\ell, \ell) = 0$  and  $-\ell \cdot L'(\ell, \ell(1-z)) \geq A \cdot L(\ell, \ell(1-z))^a$ ;
5. for  $\ell \in \{-1, 1\}$ , for every  $z \in [-1, 1]$ ,  $L''(\ell, z) \leq B$ .

We remark that condition (2) is for convenience only and can be achieved by scaling any loss function satisfying the other conditions. Condition (4) ensures that the loss function is monotone (that is for all  $y, y' \in [-1, 1]$ , if  $y \leq y'$  then  $L(-1, y) \leq L(-1, y')$  and  $L(1, y') \leq L(1, y)$ ) and that it has a non-negligible slope whenever the loss itself is non-negligible. Condition (5) ensures that the linear approximation to  $L$  dominates the remainder term in the Taylor series. A simple example of a well-behaved loss function is  $L(y, z) = |y - z|^c / 2^{c-1}$  for any constant  $c \geq 2$ . It is also easy to see that any convex combination of well-behaved loss functions is well-behaved. Note that the linear loss function is not twice differentiable on  $[-1, 1]$  and also does not satisfy condition (4). Hence our result for linear threshold functions does not contradict the lower bound for conjunctions. We now prove a generalization of Theorem 11 to well-behaved loss functions.

**Theorem 14** Let  $L$  be a well-behaved loss function with bounds  $a, A$  and  $B$ . For  $\phi(x) \in \mathcal{F}_1^\infty$ , let

$$N_\alpha(\phi) = \{P_1(\phi + \alpha' \cdot x_i) \mid i \in [0..n], |\alpha'| = \alpha\} \cup \{\phi\}.$$

For every halfspace  $f$  with margin  $\gamma$  on distribution  $D$  over  $B_n$  and every  $\epsilon > 0$ , there exists  $\phi' \in N_\alpha(\phi)$  for which

$$\mathbf{E}_D[L(f, \phi')] \leq \max\{\mathbf{E}_D[L(f, \phi)] - \alpha^2 \cdot B/2, \epsilon\},$$

where  $\alpha = A \cdot \gamma \cdot \epsilon^{a+1} / (B \cdot 2^{a+3} \sqrt{n})$ .

**Proof** As before, we can assume that  $\mathbf{E}_D[L(f, \phi)] > \epsilon$ . In particular,  $\mathbf{Pr}_D[L(f, \phi) \geq \epsilon/2] \geq \epsilon/4$ . Then, by property (4) of well-behaved loss-functions,  $\mathbf{Pr}_D[|L'(f, \phi)| \geq A(\epsilon/2)^a] \geq \epsilon/4$ . This implies that

$$\mathbf{E}_D[|L'(f, \phi)|] \geq \epsilon/4 \cdot A \cdot (\epsilon/2)^a = A \cdot \epsilon^{a+1} / 2^{a+2}.$$

By monotonicity of  $L$  (or property (4)), for every  $x$  in the support of  $D$ ,  $-L'(f(x), \phi(x))$  has the same sign as  $f(x)$  and therefore also the same sign as  $\sum_{i \in [n]} w_i x_i - \theta$ . This gives

$$\mathbf{E}_D \left[ -L'(f, \phi) \left( \sum_{i \in [n]} w_i x_i - \theta \right) \right] \geq \gamma \mathbf{E}_D [|L'(f, \phi)|] \geq A \cdot \gamma \epsilon^{a+1} / 2^{a+2}. \quad (6)$$

In addition, as in equation (4), we have

$$\mathbf{E}_D \left[ -L'(f, \phi) \left( \sum_{i \in [n]} w_i x_i - \theta \right) \right] \leq \sqrt{2} \sqrt{\sum_{i \in [0..n]} \mathbf{E}_D [L'(f, \phi) \cdot x_i]^2}. \quad (7)$$

By combining equations (6) and (7) we obtain that

$$\sum_{i \in [0..n]} \mathbf{E}_D [L'(f, \phi) x_i]^2 \geq A^2 \cdot \gamma^2 \epsilon^{2a+2} / 2^{2a+5}.$$

From here we can conclude that there exists  $j \in [0..n]$  such that

$$|\mathbf{E}_D [L'(f, \phi) x_j]| \geq A \cdot \gamma \cdot \epsilon^{a+1} / (2^{a+2} \sqrt{2n+2}) \geq A \cdot \gamma \cdot \epsilon^{a+1} / (2^{a+3} \sqrt{n}). \quad (8)$$

We denote the right side of this inequality by  $\rho$ .

To finish the proof we prove an analogue of Lemma 12 saying that a step in the direction of  $x_j$  from  $\phi$  will decrease the loss. Formally,

**Lemma 15** *Let  $\alpha' = -\alpha \cdot \text{sign}(\mathbf{E}_D [L'(f, \phi) x_j])$ , and  $\phi' = P_1(\phi + \alpha' \cdot x_j)$ , where  $\alpha = \rho/B$  (as defined in the statement of the theorem). Then*

$$\mathbf{E}_D [L(f, \phi')] \leq \mathbf{E}_D [L(f, \phi)] - \alpha^2 \cdot B/2.$$

**Proof** Let  $x \in X$  be any point. Assume that  $f(x) = -1$ . For convenience we extend the loss function  $L(-1, z)$  to values  $z \in [-2, -1]$  by setting  $L(-1, z) = L(-1, -2 - z)$  (that is by making the loss symmetric around  $-1$ ). By the properties of the loss function,  $L(-1, -1) = 0$ ,  $L'(-1, -1) = 0$  and for  $z \in [-2, -1]$ ,  $L''(-1, z) = L''(-1, -2 - z)$ . This implies that the extended  $L$  is twice differentiable in  $[-2, 1]$  and  $L''(-1, z) \leq B$  for every  $z \in [-2, 1]$ . We first assume that  $\phi + \alpha' \cdot x_j \in [-2, 1]$ .  $L$  is twice differentiable and therefore Taylor's theorem gives

$$L(-1, \phi(x) + \alpha' \cdot x_j) - L(-1, \phi(x)) = \alpha' \cdot x_j \cdot L'(-1, \phi(x)) + (\alpha' \cdot x_j)^2 \cdot L''(-1, \zeta)/2,$$

where  $\zeta \in [\phi(x), \phi(x) + \alpha' \cdot x_j] \subseteq [-2, 1]$ . Also note that in this case,  $L(-1, \phi(x) + \alpha' \cdot x_j) \geq L(-1, \phi'(x))$ . This means that

$$L(-1, \phi'(x)) - L(-1, \phi(x)) \leq \alpha' \cdot x_j \cdot L'(-1, \phi(x)) + \alpha^2 \cdot B/2, \quad (9)$$

Now if  $\phi + \alpha' \cdot x_j > 1$  then  $\phi'(x) = 1$  and  $\alpha' \cdot x_j > 1 - \phi(x) > 0$ . Then  $\alpha' \cdot x_j \cdot L'(-1, \phi(x)) \geq (1 - \phi(x)) \cdot L'(-1, \phi(x))$  (as  $L'(-1, \phi(x)) > 0$ ). Hence,

$$\begin{aligned} L(-1, \phi'(x)) - L(-1, \phi(x)) &= (1 - \phi(x)) \cdot x_j \cdot L'(-1, \phi(x)) + ((1 - \phi(x)) \cdot x_j)^2 \cdot L''(-1, \zeta)/2 \\ &\leq \alpha' \cdot x_j \cdot L'(-1, \phi(x)) + \alpha^2 \cdot B/2, \end{aligned} \quad (10)$$

where  $\zeta \in [\phi(x), 1]$ . By treating the case when  $f(x) = 1$  symmetrically and combining equations (9) and (10) we will obtain that for every  $x$ ,

$$L(f(x), \phi'(x)) - L(f(x), \phi(x)) \leq \alpha' \cdot x_j \cdot L'(f(x), \phi(x)) + \alpha^2 \cdot B/2.$$

This immediately implies that

$$\begin{aligned} \mathbf{E}_D[L(f(x), \phi'(x))] - \mathbf{E}_D[L(f(x), \phi(x))] &\leq \alpha' \mathbf{E}_D[x_j \cdot L'(f(x), \phi(x))] + \alpha^2 \cdot B/2 \\ &\leq -\alpha\rho + \alpha^2 \cdot B/2 = \alpha^2 \cdot B/2. \end{aligned}$$

■

To finish the proof of Th. 14 we observe that  $\phi'(x) \in N_\alpha(\phi)$ .

■

As we have mentioned, a simple corollary of Theorem 14 is distribution-independent evolvability of large margin halfspaces with any well-behaved loss function.

**Theorem 16** *For every well-behaved loss function  $L$  and  $\gamma \geq 1/q(n)$  for some polynomial  $q(\cdot)$ ,  $\text{HS}_\gamma$  over  $B_n$  is monotonically evolvable with  $L$ .*

We make two remarks regarding these theorems.

**Remark 17** *In both Theorems 11 and 14 it is not necessary to know the exact value of  $\alpha$  to create a strictly beneficial neighborhood. It is easy to see from the analysis that the bound holds for every  $\alpha_0 < \max_{j \in [0..n]} \{|\mathbf{E}_D[L'(f, \phi)x_j]|\}$ . Therefore by including in the neighborhood steps for all values of  $\alpha_0 = 2^{-t}$  for  $t \in [n]$ , the neighborhood will include a function with at least  $1/4$  of the improvement that can be achieved when a bound on  $\alpha$  is known in advance.*

**Remark 18** *Theorem 14 does not require the loss function to be the same for all  $x$  as long as for every point  $x$ , the loss-function  $L_x$  is well-behaved with the same bounds  $a, A, B$ . Similarly the loss function does not need to stay the same between generations and can change arbitrarily as long as it is well-behaved with the same bounds  $a, A, B$ .*

A number of popular machine learning algorithms work by embedding the data points in a different Euclidean space (most commonly by using a kernel) and then applying a learning algorithm for halfspaces, such as SVM. This method is also used in a number of theoretical algorithms such as the DNF learning algorithm based on the polynomial threshold function representation of Klivans and Servedio (2004). As expected, this technique can be easily translated to the evolvability framework and then used together with our result. Formally, let  $C$  and  $C'$  be concept classes over the domains  $X$  and  $X'$ , respectively. The concept  $C$  over  $X$  is said to be embeddable as  $C'$  over  $X'$  if there exists a function  $\Phi : X \rightarrow X'$  such that for every  $f \in C$ , there exists  $g \in C'$  such that for every  $x \in X$ ,  $g(\Phi(x)) = f(x)$ . We also say that the embedding is efficient if  $\Phi(x)$  is computable efficiently, that is in time polynomial in the dimension of  $x$  (or description length in general). Embeddability of concept classes into large-margin halfspaces has been studied in a number of works

initiated by Forster (2002) and Forster et al. (2003) (see Linial et al., 2007; Sherstov, 2007; Linial and Shraibman, 2009, for some recent results). The inverse of the optimal margin is referred to as the *margin complexity* of a concept class (Linial et al., 2007). Besides its importance to machine learning, it has several connections to fundamental quantities in communication complexity (Goldmann et al., 1992; Sherstov, 2007; Linial and Shraibman, 2009). We cannot invoke this measure directly to upper-bound the complexity of using our evolution algorithm since margin complexity disregards the computational complexity of the embedding function. But given an efficient embedding function the application of our evolution algorithm becomes straightforward.

**Corollary 19** *Let  $C$  be a concept class over domain  $X$ ,  $X' \subseteq B_n$  and  $\gamma > 1/q(n)$  for some polynomial  $q(\cdot)$ . If there exists an efficiently computable embedding of  $C$  over  $X$  to  $\text{HS}_\gamma(X')$  over  $X'$ , then  $C$  is evolvable monotonically with any well-behaved loss function.*

## 5. Conclusions and Open Problems

Our lower bound in Section 3 provides strong hardness results for learning that is limited to observing the accuracy (or alternatively, Boolean loss performance) of hypotheses. In particular, it implies that evolvability in Valiant’s original model is severely limited unless the distribution over the domain is strongly restricted. An interesting direction for further work would be to find a complete characterization of distribution independent CSQ learnability (as was recently achieved for SQ learnability (Simon, 2007; Feldman, 2009b; Szörényi, 2009)). Potentially simpler questions left open in this work are whether conjunction are CSQ learnable to constant accuracy (say  $1/4$ ) and whether the lower bound for conjunctions can be strengthened from a quasi-polynomial to an exponential number of queries.

At the same time results of Section 4 demonstrate that the limitation of Boolean loss can be overcome by using a real-valued hypotheses with a non-linear loss function. The evolution algorithm we described is based on the first mutation algorithm that is simple, general and robust enough to be a plausible candidate for biological evolution. It would be interesting to know if similar result can be proved for other SQ learnable concept classes (e.g. general linear threshold functions) and whether the result can be extended to more general loss functions.

## References

- N. Bshouty and V. Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2:359–395, 2002. ISSN 1533-7928.
- H. Buhrman, N. Vereshchagin, and R. de Wolf. On computation and communication with small bias. In *Proceedings of IEEE Conference on Computational Complexity*, pages 24–32, 2007.
- D. Diachnos and G. Turán. On evolvability: The swapping algorithm, product distributions, and covariance. In *Proceedings of Stochastic Algorithms: Foundations and Applications (SAGA)*, pages 74–88, 2009.

- V. Feldman. Evolvability from learning algorithms. In *Proceedings of STOC*, pages 619–628, 2008.
- V. Feldman. Robustness of evolvability. In *Proceedings of COLT*, pages 277–292, 2009a.
- V. Feldman. A complete characterization of statistical query learning with applications to evolvability. In *Proceedings of FOCS*, pages 375–384, 2009b.
- V. Feldman and L. G. Valiant. The learning power of evolution. In *Proceedings of COLT*, pages 513–514, 2008.
- J. Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *Journal of Computer and System Sciences*, 65(4):612–625, 2002.
- J. Forster, N. Schmitt, H.U. Simon, and T. Suttorp. Estimating the optimal margins of embeddings in euclidean half spaces. *Machine Learning*, 51(3):263–281, 2003.
- Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, 1995.
- M. Goldmann, J. Håstad, and A. Razborov. Majority gates vs. general weighted threshold gates. *Computational Complexity*, 2:277–300, 1992.
- J. Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *Journal of Computer and System Sciences*, 55:414–440, 1997.
- B. Jacobson, 2007. Personal communication with L. Valiant.
- V. Kanade, L. G. Valiant, and J. Wortman Vaughan. Evolution with drifting targets. In *Proceedings of COLT*, pages 155–167, 2010.
- M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.
- M. Kearns and L. Valiant. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM*, 41(1):67–95, 1994.
- M. Kearns, R. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.
- A. Klivans and R. Servedio. Learning DNF in time  $2^{\tilde{O}(n^{1/3})}$ . *Journal of Computer and System Sciences*, 68(2):303–318, 2004.
- N. Linial and A. Shraibman. Learning complexity vs communication complexity. *Combinatorics, Probability & Computing*, 18(1-2):227–245, 2009.
- N. Linial, S. Mendelson, G. Schechtman, and A. Shraibman. Complexity measures of sign matrices. *Combinatorica*, 27(4):439–463, 2007.
- N. Littlestone. Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1987.

- L. Michael. Evolving decision lists. Manuscript, 2007.
- F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958.
- R. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.
- A. A. Sherstov. Halfspace matrices. In *Proceedings of Conference on Computational Complexity*, pages 83–95, 2007.
- H. Simon. A characterization of strong learnability in the statistical query model. In *Proceedings of Symposium on Theoretical Aspects of Computer Science*, pages 393–404, 2007.
- B. Szörényi. Characterizing statistical query learning:simplified notions and proofs. In *ALT*, pages 186–200, 2009.
- L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
- L. G. Valiant. Evolvability. *Journal of the ACM*, 56(1):3.1–3.21, 2009. Earlier version in ECCC, 2006.
- P. Valiant. Distribution free evolvability of real functions over all convex loss functions. Unpublished manuscript, 2011.

# Lower Bounds and Hardness Amplification for Learning Shallow Monotone Formulas

**Vitaly Feldman**

*IBM Almaden Research Center*

VITALY@POST.HARVARD.EDU

**Homin K. Lee**

*University of Texas at Austin*

HOMIN@CS.UTEXAS.EDU

**Rocco A. Servedio**

*Columbia University*

ROCCO@CS.COLUMBIA.EDU

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

Much work has been done on learning various classes of “simple” monotone functions under the uniform distribution. In this paper we give the first unconditional lower bounds for learning problems of this sort by showing that polynomial-time algorithms cannot learn shallow monotone Boolean formulas under the uniform distribution in the well-studied Statistical Query (SQ) model.

We introduce a new approach to understanding the learnability of “simple” monotone functions that is based on a recent characterization of Strong SQ learnability by Simon (2007). Using the characterization we first show that depth-3 monotone formulas of size  $n^{o(1)}$  cannot be learned by any polynomial-time SQ algorithm to accuracy  $1 - 1/(\log n)^{\Omega(1)}$ . We then build on this result to show that depth-4 monotone formulas of size  $n^{o(1)}$  cannot be learned even to a certain  $\frac{1}{2} + o(1)$  accuracy in polynomial time. This improved hardness is achieved using a general technique that we introduce for amplifying the hardness of “mildly hard” learning problems in either the PAC or SQ framework. This hardness amplification for learning builds on the ideas in the work of O’Donnell (2004) on hardness amplification for approximating functions using small circuits, and is applicable to a number of other contexts.

Finally, we demonstrate that our approach can also be used to reduce the well-known open problem of learning juntas to learning of depth-3 monotone formulas.

**Keywords:** statistical query learning, Boolean formulas, statistical query dimension, hardness of learning

## 1. Introduction

**Motivation.** Over the past several decades much work in computational learning theory has focused on developing efficient algorithms for learning monotone Boolean functions under the uniform distribution, (see e.g., Amano and Maruoka, 2002; Blum et al., 1998; Bshouty and Tamon, 2006; Hancock and Mansour, 1991; Jackson et al., 2008; Kearns et al., 1994; O’Donnell and Servedio, 2007; O’Donnell and Wimmer, 2009; Sakai and Maruoka, 2000; Servedio, 2004) and other works. An intriguing question, which has driven much of this research and remains open, is whether there is an efficient algorithm to learn *monotone DNF formulas* under the uniform distribution. Such an algorithm  $A$  would have the following

performance guarantee: for any target function  $f : \{0, 1\}^n \rightarrow \{0, 1\}$  that is a monotone DNF formula with  $\text{poly}(n)$  terms, given access to independent uniform random examples  $(x, f(x))$ , algorithm  $A$  would run in  $\text{poly}(n, 1/\epsilon)$  time and with high probability output a hypothesis  $h$  that disagrees with  $f$  on at most an  $\epsilon$  fraction of inputs from  $\{0, 1\}^n$ .

Several partial positive results toward learning monotone DNF have been obtained: for constant  $\epsilon$ , algorithms are known that can learn  $2^{\sqrt{\log n}}$ -term monotone DNF (Servedio, 2004) and  $\text{poly}(n)$ -size monotone decision trees (O'Donnell and Servedio, 2007) in  $\text{poly}(n)$  time. Partial negative results have also been given: Dachman-Soled et al. (2008) has shown that (under a strong cryptographic hardness assumption) for a sufficiently large absolute constant  $d$  there is no  $\text{poly}(n)$ -time algorithm that can learn depth- $d$ , size- $n^{o(1)}$  Boolean formulas that compute monotone functions to a certain accuracy  $\frac{1}{2} + o(1)$ . However no hardness results that apply to monotone formulas of small constant depth are known.

In this work we give *unconditional* lower bounds showing that simple monotone functions – computed by monotone Boolean formulas of depth 3 or 4 and size  $n^{o(1)}$  – cannot be learned under the uniform distribution in polynomial time. Of course these results are not in the PAC model of learning from random examples (since unconditional lower bounds in this model would prove  $P \neq NP!$ ); our primary lower bounds are for the closely-related and well-studied *Statistical Query* learning model, which we describe briefly below.

**Statistical Query learning.** Kearns (1998) introduced the *statistical query (SQ)* learning model as a natural variant of the usual PAC learning model. In the SQ model, instead of having access to independent random examples  $(x, f(x))$  drawn from distribution  $\mathcal{D}$ , the learner is only allowed to obtain statistical properties of examples. Formally, it has access to a *statistical query oracle*  $SQ_{f,\mathcal{D}}$ . The oracle  $SQ_{f,\mathcal{D}}$  takes as input a *query function*  $g : X \times \{-1, +1\} \rightarrow \{-1, +1\}$  and a *tolerance parameter*  $\tau \in [0, 1]$  and outputs a value  $v$  such that:

$$|v - \mathbf{E}_{\mathcal{D}}[g(x, f(x))]| \leq \tau.$$

The learner's goal – to output a hypothesis  $h$  such that  $\Pr_{x \sim \mathcal{D}}[h(x) \neq f(x)] \leq \epsilon$  – is the same as in PAC learning. A  $\text{poly}(n, 1/\epsilon)$ -time SQ algorithm is only allowed to make queries in which  $g$  can be computed by a  $\text{poly}(n, 1/\epsilon)$ -size circuit and  $\tau$  is at most a fixed  $1/\text{poly}(n, 1/\epsilon)$  (and of course the algorithm must run for at most  $\text{poly}(n, 1/\epsilon)$  time steps).

The SQ model is an important and well-studied learning model which has received much research attention in the 15 years since it was introduced. One reason for this intense interest is that any concept class that is efficiently learnable from statistical queries is also efficiently PAC learnable in the presence of random classification noise at any noise rate bounded away from  $\frac{1}{2}$  (Kearns, 1998). In fact, since the introduction of the SQ-model virtually all known noise-tolerant learning algorithms have been obtained from (or rephrased as) SQ algorithms<sup>1</sup> (Kearns, 1998; Bylander, 1994, 1998; Dunagan and Vempala, 2004).

Even more importantly, and quite surprisingly, all known techniques in PAC learning with the exception of Gaussian elimination either fit easily into the SQ model or have SQ analogues. Thus the study of SQ learning is now an integral part of the study of noise-tolerant learning and of PAC learning in general. In addition, interest in the SQ

---

1. One prominent exception is the work of Blum et al. (2003), which gives an algorithm for learning parities which is tolerant to random noise, although in a weaker sense than the algorithms derived from statistical queries.

learning model has been stimulated by recently discovered close connections to privacy-preserving learning (Blum et al., 2005; Kasiviswanathan et al., 2008), evolvability (Valiant, 2009; Feldman, 2008, 2010) and communication complexity (Sherstov, 2008).

An important property of the SQ model that we rely on in this work is that it is possible to prove *unconditional* information-theoretic lower bounds on learning a class of functions in the SQ model (we will say much more about this below). Such lower bounds give a very strong indication that the class is unlikely to be efficiently PAC learnable and even less likely to be PAC learnable in the presence of random classification noise. In particular, it rules out almost all known approaches to the learning problem, including the algorithms that rely solely on estimates of Fourier coefficients (which is the primary technique for learning over the uniform distribution (Servedio, 2004; O'Donnell and Servedio, 2007)).

**Background on hardness results for SQ learning.** In his paper introducing the SQ model, Kearns (1998) already showed that the class of all parity functions cannot be SQ-learned in polynomial time under the uniform distribution. Soon after this Blum et al. (1994) characterized the weak learnability of every function class  $\mathcal{F}$  in the SQ model in terms of the *statistical query dimension* of  $\mathcal{F}$ ; roughly speaking, this is the largest number of functions from  $\mathcal{F}$  that are pairwise nearly orthogonal to each other (we give a precise definition in Section 2). The results of Blum et al. (1994) imply that if a class  $\mathcal{F}$  has SQ-Dimension  $n^{\omega(1)}$ , then no SQ algorithm can even weakly learn  $\mathcal{F}$  to any accuracy  $\frac{1}{2} + \frac{1}{\text{poly}(n)}$  in  $\text{poly}(n)$  time. This bound was already used to give SQ hardness results for weak learning classes such as DNF and decision trees in the work of Blum et al. (1994), and more recently for weak-learning intersections of halfspaces by Klivans and Sherstov (2006). However, it is well known that the entire class of all monotone Boolean functions over  $\{0, 1\}^n$  can be weakly learned to accuracy  $\frac{1}{2} + \frac{1}{\text{poly}(n)}$  in  $\text{poly}(n)$  time (an algorithm that achieves optimal accuracy  $\frac{1}{2} + \frac{\Theta(\log n)}{\sqrt{n}}$  was recently given by O'Donnell and Wimmer (2009)), and indeed the class of all monotone functions can easily be shown to have SQ-dimension  $O(n)$ . Thus the notion of SQ-dimension alone is not enough to yield SQ lower bounds on learning monotone functions.

Much more recently, Simon (2007) introduced a combinatorial parameter of a function class  $\mathcal{F}$  called its *strong Statistical Query dimension*, and showed that this parameter at error rate  $\epsilon$  characterizes the information-theoretic *strong* learnability of  $\mathcal{F}$  to accuracy  $1 - \epsilon$ . (We give a precise definition of the strong SQ-dimension in Section 2.) We use this characterization, which was later strengthened and simplified by Feldman (2010) (and independently by Szörényi, 2009) to obtain the lower bounds, which we now describe. (Throughout the following description of our results, the underlying distribution is always taken to be uniform over  $\{0, 1\}^n$ .)

**Our Results: Unconditional Hardness of Learning Simple Monotone Functions.** We give the first strong SQ-dimension lower bound for a class of “simple” monotone functions. More precisely, as our first main result, we show that the class of size- $n^{o(1)}$ , depth-3 monotone formulas has strong SQ-dimension  $n^{\omega(1)}$  at a certain error rate  $1/(\log n)^{\Theta(1)}$ . By the results of Simon and Feldman, this implies that such formulas cannot be efficiently learned to accuracy  $1 - 1/(\log n)^{\Theta(1)}$  by any polynomial-time SQ learning algorithm. Roughly speaking, our proof works by constructing a class of slice functions of “well-separated” parities over  $\text{polylog}(n)$  variables. We show that this class of functions

has the combinatorial properties required to satisfy the strong SQ-dimension criterion, and that every function in the class can be computed by a small monotone formula of depth 3.

In addition to this result, we show that a variant of the basic idea of our construction can be used to reduce PAC learning of  $\log n$ -juntas (or functions that depend on at most  $\log n$  variables) over  $k = \log^2 n / \log \log n$  variables to learning of depth-3 monotone functions. Learning of juntas is an important open problem in learning theory (Blum, 2003) for which the best known algorithm achieves only a polynomial factor speed-up over the trivial brute force algorithm (Mossel et al., 2007). Note that there are  $n^{\Omega(\log \log n)}$  juntas in the above class of functions and therefore our reduction implies that strong learning of monotone depth-3 formulas of  $\text{poly}(n)$ -size (even over only  $k$  variables) would require a major breakthrough on the problem of learning juntas over the uniform distribution.

These results are the first lower bounds for learning monotone depth-3 formulas, and leave only the question of learning monotone DNF formulas (which are of course depth-2 rather than depth-3 monotone formulas) open. However, these results only say that monotone depth-3 formulas cannot be learned to a rather high  $(1 - o(1))$  accuracy. Thus a natural goal is to obtain stronger hardness results which show that simple monotone functions are hard to SQ learn even to coarse accuracy – ideally to some accuracy level  $\frac{1}{2} + o(1)$  only slightly better than random guessing. Of course, we might expect that to achieve this we must use somewhat more complicated functions than depth-3 formulas, and this does turn out to be the case – but only a bit more complicated, as we describe below.

We introduce a general method of amplifying the hardness of a class of functions that are “mildly hard to learn” (i.e., hard to learn to high accuracy), to obtain a class of functions that are “very hard to learn” (i.e., hard to learn to accuracy even slightly better than random guessing). We show that our method, which builds on O’Donnell’s beautiful hardness amplification for approximating Boolean functions using small circuits (O’Donnell, 2004), can be applied both within the uniform-distribution PAC model (Th. 12) and within the uniform-distribution Statistical Query model (Th. 14). The latter is of course our main interest in this paper, but we believe that the result is of independent interest and therefore present both versions.

We note that while our hardness amplification follows the general approach of O’Donnell, the learning setting is quite different from approximation of a fixed function by a non-uniform circuit and hence new technical ideas are required to successfully translate the approach (especially in the SQ case). We defer the discussion of the proof technique and technical contributions to Section 4.

Using this hardness amplification for SQ learning together with our first main result, we obtain our second main result: we show that the class of size- $n^{o(1)}$ , depth-4 monotone formulas cannot be SQ-learned even to  $\frac{1}{2} + 2^{-(\log n)^\gamma}$  accuracy in  $\text{poly}(n)$  time for any  $\gamma < 1/2$ . We are able to increase the depth by only one (from 3 to 4) by a careful construction of the combining function in our hardness amplification framework; we use a depth-2 combining function due to Talagrand and the complement of the “tribes” function which have useful extremal noise stability properties as shown by O’Donnell (2004); Mossel and O’Donnell (2003).

The primary motivating question of learnability of monotone DNF formulas over the uniform distribution is not resolved in this work. At the same time our results suggest that this long-standing open problem can be tackled by bounding the strong SQ-dimension of

monotone DNF formulas (our own efforts to derive a non-trivial bound were not successful thus far).

**Relation to previous work.** To the best of our knowledge even the “mild hardness” result that we prove for depth-3 monotone formulas is the first unconditional negative result known for learning a class of polynomial-time computable monotone functions in the uniform-distribution SQ model. We note that the strong  $\frac{1}{2} + o(1)$  hardness that we establish for depth-4 monotone formulas is provably near-optimal, since as mentioned earlier the class of *all* monotone functions over  $\{0, 1\}^n$  can be learned to accuracy  $\frac{1}{2} + \frac{\Theta(\log n)}{\sqrt{n}}$  in polynomial time (O’Donnell and Wimmer, 2009).

While the recent work by Dachman-Soled et al. (2008) also gave negative results for learning constant-depth monotone formulas, those results are different from ours in significant ways. Dachman-Soled et al. (2008) used a strong cryptographic hardness assumption – that Blum integers are  $2^{n^\epsilon}$ -hard to factor on average for some fixed  $\epsilon > 0$  – to show that for some sufficiently large absolute constant  $d$ , the class of monotone functions computed by size- $n^{o(1)}$ , depth- $d$  formulas cannot be PAC learned, under the uniform distribution, to a certain accuracy  $\frac{1}{2} + o(1)$ . In contrast, our main hardness result applies to the more restricted classes of size- $n^{o(1)}$ , depth-3 and 4 *monotone* formulas, and gives *unconditional* hardness for polynomial-time algorithms in the Statistical Query model.

Our reduction from learning juntas can be thought of as giving a hardness result based on a relatively strong computational assumption and hence has the same flavor as the result by (Dachman-Soled et al., 2008). In addition to better depth, our reduction is substantially simpler and more direct than the reduction from factoring Blum integers.

Finally, we remark here that our hardness amplification method for PAC and SQ learning may be viewed as a significant strengthening and generalization of some earlier results. Boneh and Lipton (1993) described a form of uniform-distribution hardness amplification for PAC learning based on the XOR lemma; our PAC hardness amplification generalizes their result and extends to SQ learning. More recently, Dachman-Soled et al. (2008) used elements of O’Donnell’s technique to amplify information-theoretic hardness of learning. Specifically, the “mildly hard” class of functions  $\mathcal{F}$  used by Dachman-Soled et al. (2008) consists of all functions of the form  $\text{slice}(f)$ , where  $f$  may be any Boolean function and  $\text{slice}(f)$  is the function which agrees with Majority everywhere except the middle layer of the Boolean hypercube. An easy argument shows that  $\mathcal{F}$  is a class of monotone functions that is hard to learn to accuracy  $1 - \Theta(1)/\sqrt{n}$ . Using the fact that a random function in  $\mathcal{F}$  is trivial to predict off of the middle layer and is totally random on the middle layer, expected bias analysis from (O’Donnell, 2004) is used by Dachman-Soled et al. (2008) to derive information-theoretic hardness of learning the combined function class  $\mathcal{F}^g$ . In contrast, the hardness amplification results of this paper amplify computational hardness, do not assume any particular structure of the base class  $\mathcal{F}$  (only that it is “mildly hard to learn”) and, importantly, apply to learning in the SQ model.

**Organization.** Section 2 gives background on Statistical Query learning, the SQ-dimension, and the strong SQ-dimension. In Section 3 we describe our class of depth-3 monotone formulas and show that it is “mildly” hard to learn in the SQ model by giving a superpolynomial lower bound on its strong SQ-dimension. We give the reduction from learning juntas in Section 3.3. Section 4 presents our general hardness amplification results for the uniform-

distribution PAC model and the uniform-distribution Statistical Query learning model. We apply our hardness amplification technique from Section 4 to obtain our second main result, strong SQ hardness for depth-4 formulas, in Section 5.

## 2. The Statistical Query Model, SQ-Dimension, and Strong SQ-Dimension

Recall the definition of Statistical Query learning from Section 1. Blum et al. (1994) introduced the notion of the *SQ-dimension of function class  $\mathcal{F}$  under distribution  $\mathcal{D}$* , and showed that it characterizes the weak-learnability of  $\mathcal{F}$  under  $\mathcal{D}$  in the SQ model. Bshouty and Feldman (2002) and Yang (2005) later generalized and sharpened the result of Blum et al. (1994). We will use Yang’s version here extended to sets of arbitrary real-valued functions. We write “ $\langle f, g \rangle_{\mathcal{D}}$ ” to denote  $\mathbf{E}_{x \sim \mathcal{D}}[f(x)g(x)]$  and “ $\|f\|_{\mathcal{D}}$ ” to denote  $(\langle f, f \rangle_{\mathcal{D}})^{1/2}$ .

**Definition 1** *Given a set  $\mathcal{C}$  of real-valued functions, the SQ-dimension of  $\mathcal{C}$  with respect to  $\mathcal{D}$  (written  $SQ\text{-DIM}(\mathcal{C}, \mathcal{D})$ ) is the largest number  $d$  such that  $\exists \{f_1, \dots, f_d\} \subseteq \mathcal{C}$  with the property that  $\forall i \neq j$ ,*

$$|\langle f_i, f_j \rangle_{\mathcal{D}}| \leq \frac{1}{d}. \quad (1)$$

When  $\mathcal{D}$  is the uniform distribution we simply write  $SQ\text{-DIM}(\mathcal{C})$ . We refer to the LHS of Equation (1) as the *correlation* between  $f_i$  and  $f_j$  under  $\mathcal{D}$ .

Intuitively, this condition says that  $\mathcal{C}$  contains  $d$  “nearly-uncorrelated” functions. It is easy to see that if  $\mathcal{C}$  is a concept class with  $SQ\text{-DIM}(\mathcal{C}, \mathcal{D}) = d$  then  $\mathcal{C}$  can be weakly learned with respect to  $\mathcal{D}$  to accuracy  $\frac{1}{2} + \frac{\Theta(1)}{d}$  using  $d$  Statistical Queries with tolerance  $\frac{\Theta(1)}{d}$ ; simply ask for the correlation between the unknown target function  $f$  and each function in the set  $\{f_1, \dots, f_d\}$ . Since the set is maximal, the target function must have correlation at least  $1/d$  with at least one of the functions.

Blum et al. showed that the other direction is true as well; if  $\mathcal{C}$  is efficiently weakly learnable, then  $\mathcal{C}$  must have small SQ-dimension.

**Theorem 2 (Blum et al., 1994, Th. 12)** *Given a concept class  $\mathcal{C}$  and a distribution  $\mathcal{D}$ , let  $SQ\text{-DIM}(\mathcal{C}, \mathcal{D}) = d$ . Then if the tolerance  $\tau$  of each query is always at least  $1/d^{1/3}$ , at least  $\frac{1}{2}d^{1/3} - 1$  queries are required to learn  $\mathcal{C}$  with advantage  $1/d^3$ .*

As an example, the class  $PAR_n$  of all parity functions over  $n$  variables has  $SQ\text{-DIM}(PAR_n) = 2^n$ , and thus any SQ algorithm for learning parities over the uniform distribution  $\mathcal{U}$  to accuracy  $\frac{1}{2} + \frac{1}{2^{O(n)}}$  requires exponential time.

### 2.1. The Strong SQ-Dimension

The statistical query dimension only characterizes the weak SQ-learnability of a class and is not sufficient to characterize its strong SQ-learnability. The first characterization of strong SQ learning was given by Simon (2007), but for our application a subsequent accuracy-preserving characterization by Feldman will be more convenient to use (Feldman, 2010).

Let  $\mathcal{F}_1^\infty$  denote the set of all functions from  $\{0, 1\}^n \rightarrow [-1, 1]$ , i.e., all functions with  $L_\infty$ -norm bounded by 1. For a Boolean function  $f$ , we define  $B_{\mathcal{D}}(f, \epsilon)$  to be  $\{g : \{0, 1\}^n \rightarrow$

$\{-1, 1\} : \Pr_{\mathcal{D}}[g \neq f] \leq \epsilon\}$ , i.e., the  $\epsilon$ -ball around  $f$ . The sign function is defined as  $\text{sign}(z) = 1$  for  $z \geq 0$ ,  $\text{sign}(z) = -1$  for  $z < 0$ . Finally, for a set of real-valued functions  $\mathcal{C}$ , let  $\mathcal{C} - g = \{f - g : f \in \mathcal{C}\}$ .

**Definition 3** Given a concept class  $\mathcal{C}$  and  $\epsilon > 0$ , the strong SQ-dimension of  $\mathcal{C}$  with respect to  $\mathcal{D}$  is defined to be:

$$\text{SQ-SDIM}(\mathcal{C}, \mathcal{D}, \epsilon) = \sup_{g \in \mathcal{F}_1^\infty} \text{SQ-DIM}((\mathcal{C} \setminus B_{\mathcal{D}}(\text{sign}(g), \epsilon)) - g, \mathcal{D}).$$

Just as for the weak SQ-dimension, the strong SQ-dimension completely characterizes the strong SQ-learnability of a concept class.

**Theorem 4 (Feldman, 2010)** Let  $\mathcal{C}$  be a concept class over  $\{0, 1\}^n$ ,  $\mathcal{D}$  be a probability distribution over  $\{0, 1\}^n$  and  $\epsilon > 0$ . If there exists a polynomial  $p(\cdot, \cdot)$  such that  $\mathcal{C}$  is SQ learnable over  $\mathcal{D}$  to accuracy  $\epsilon$  from  $p(n, 1/\epsilon)$  queries of tolerance  $1/p(n, 1/\epsilon)$  then  $\text{SQ-SDIM}(\mathcal{C}, \mathcal{D}, \epsilon + 1/p(n, 1/\epsilon)) \leq p'(n, 1/\epsilon)$  for some polynomial  $p'(\cdot, \cdot)$ . Further, if  $\text{SQ-SDIM}(\mathcal{C}, \mathcal{D}, \epsilon) \leq q(n, 1/\epsilon)$  for some polynomial  $q(\cdot, \cdot)$  then  $\mathcal{C}$  is SQ learnable over  $\mathcal{D}$  to accuracy  $\epsilon$  from  $q'(n, 1/\epsilon)$  queries of tolerance  $1/q'(n, 1/\epsilon)$  for some polynomial  $q'(\cdot, \cdot)$ .

Armed with Definition 3 and Theorem 4, we can show that a concept class  $\mathcal{C}$  is not polynomial-time learnable to high accuracy by choosing a suitable  $\epsilon = \Omega(1/\text{poly}(n))$  and a suitable function  $g \in \mathcal{F}_1^\infty$  and proving that  $\text{SQ-DIM}((\mathcal{C} \setminus B_{\mathcal{U}}(\text{sign}(g), 2\epsilon)) - g) = n^{\omega(1)}$  (we can assume without loss of generality that  $\epsilon$  upper bounds the tolerance of an SQ algorithm). We do just this, for a class of depth-3 monotone formulas, in the next section.

### 3. Lower Bounds for Depth-3 Monotone Formulas

In this section we describe our lower bound for SQ learning of depth-3 monotone formulas and the reduction from learning juntas.

#### 3.1. Strong SQ lower bound

We start by showing a family of monotone functions that cannot be strong SQ-learned in polynomial time under the uniform distribution. The high-level idea is that we embed a family of non-monotone functions with high SQ-dimension – a family of parity functions – into the middle level of the  $k$ -dimensional Boolean cube, and thus obtain a class of monotone functions with high strong SQ-dimension.

A  $k$ -variable slice function for  $f$ , where  $f$  is a real-valued function over  $\{0, 1\}^k$ , is denoted  $\text{slice}_f$ . For  $x \in \{0, 1\}^k$  the value of  $\text{slice}_f(x)$  is 1 if  $x$  has more than  $\lceil k/2 \rceil$  ones,  $-1$  if  $x$  has fewer than  $\lceil k/2 \rceil$  ones, and  $f(x)$  if  $x$  has exactly  $\lceil k/2 \rceil$  ones. The functions we consider will only be defined over the first  $k$  out of  $n$  variables. Throughout the rest of this section, without loss of generality, we will always assume that  $k$  is even.

**Theorem 5** Let  $\mathcal{P}$  be the class of  $2^{k-1}$  parity functions  $\chi: \{0, 1\}^k \rightarrow \{-1, +1\}$  over an odd number of the first  $k$  variables. Let  $\mathcal{M}$  be the class of corresponding  $k$ -variable slice functions  $\text{slice}_\chi$  for  $\chi \in \mathcal{P}$ . Let  $k = \log^{2-\beta}(n)$  for  $\beta$  any absolute constant in  $(0, 1)$ . Then for every  $\epsilon = o(1/\sqrt{k})$ , we have  $\text{SQ-SDIM}(\mathcal{M}, \epsilon) = n^{\Theta(\log^{1-\beta} n)}$ , and every function in  $\mathcal{M}$  is balanced.

**Proof** We first show that every function  $\text{slice}_\chi \in \mathcal{M}$  is balanced, i.e. outputs  $+1$  and  $-1$  with equal probability. As  $k$  is even, the number of inputs with greater than  $k/2$  ones is the same as the number of inputs with fewer than  $k/2$  ones. As for the middle layer, given an input with exactly  $k/2$  ones on which  $\chi$  outputs  $+1$ , flipping all the bits gives another point with exactly  $k/2$  ones on which  $\chi$  outputs  $-1$  (as  $\chi$  is a parity on an odd number of bits). Thus every  $\text{slice}_\chi \in \mathcal{M}$  is balanced on the middle layer and thus is balanced overall.

Let  $g = \text{slice}_0$ , where  $\mathbf{0}$  is the constant 0 function. We will show that  $\text{SQ-DIM}(\mathcal{M} \setminus B_{\mathcal{U}}(\text{sign}(g), 2\epsilon) - g) = n^{\omega(1)}$ . By Stirling's approximation, the middle layer of the  $k$ -dimensional hypercube is a  $\lambda_k = \binom{k}{k/2}/2^k = \Theta(1/\sqrt{k})$  fraction of the  $2^k$  points. Thus for  $\epsilon = o(1/\sqrt{k})$  we have that  $\mathcal{M}$  is disjoint from  $B_{\mathcal{U}}(\text{sign}(g), 2\epsilon) = \emptyset$  (since  $\text{sign}(g)$  equals  $+1$  everywhere on the middle layer and every function in  $\mathcal{M}$  is balanced on the middle layer), and it is enough to lower-bound  $\text{SQ-DIM}(\mathcal{M} - g)$  in order to lower-bound the strong SQ-dimension of  $\mathcal{M}$ .

The functions in  $\mathcal{M} - g$  have a nice structure as they output 0 everywhere except the middle layer of  $\{0, 1\}^k$ , where they output  $\pm 1$ . Thus, the correlation between any two functions in  $\mathcal{M} - g$  depends only on the values on the middle slice. Let  $\chi_A, \chi_B \in \mathcal{P}$  be the parity functions over the sets of variables  $A, B \subseteq [k]$ . Recalling Equation (1),

$$|\langle \text{slice}_{\chi_A} - g, \text{slice}_{\chi_B} - g \rangle_{\mathcal{U}}| = |\mathbf{E}_{\mathcal{U}}[\mathbf{1}_{|x|=k/2} \cdot \chi_A \cdot \chi_B]| = \mathbf{E}_{\mathcal{U}}[\mathbf{1}_{|x|=k/2} \cdot \chi_{A \oplus B}] = \widehat{\mathbf{1}_{|x|=k/2}}(A \oplus B)$$

where  $A \oplus B$  denotes the symmetric difference between the sets  $A$  and  $B$ ,  $\mathbf{1}_{|x|=k/2}$  is the indicator function of the middle slice, and  $\widehat{h}(A)$  is the Fourier coefficient of  $h$  with index  $\chi_A$ . In other words, the correlation between  $(\text{slice}_{\chi_A} - g)$  and  $(\text{slice}_{\chi_B} - g)$  is exactly the Fourier coefficient of  $\mathbf{1}_{|x|=k/2}$  with index  $A \oplus B$ . Let  $s = |A \oplus B|$ . By symmetry, all  $\binom{k}{s}$  of the degree- $s$  Fourier coefficients of  $\mathbf{1}_{|x|=k/2}$  are the same, and since by Parseval's identity the squares of all the Fourier coefficients sum to  $\mathbf{E}_{\mathcal{U}}[\mathbf{1}_{|x|=k/2}^2] = \lambda_k$ , we have  $|\langle \text{slice}_{\chi_A} - g, \text{slice}_{\chi_B} - g \rangle_{\mathcal{U}}| \leq \sqrt{\lambda_k / \binom{k}{s}} \leq \binom{k}{s}^{-1/2}$ .

It remains to identify a large collection of these slice functions such that the pairwise correlations are small. This can be done easily by picking any  $\chi_A \in \mathcal{P}$ , removing all  $\chi_B \in \mathcal{P}$  such that  $|A \oplus B| \notin [k/3, 2k/3]$ , and repeating this process. Since each removal step removes at most a  $\frac{1}{2^{\Theta(k)}}$  fraction of all  $2^{k-1}$  elements of  $\mathcal{P}$ , in this fashion we can construct a set  $S$  of size  $2^{\Theta(k)}$ . Every pair of parities in  $S$  has symmetric difference  $s$  for some  $s \in [k/3, 2k/3]$ , and for such an  $s$  we have  $\binom{k}{s} = 2^{\Theta(k)}$ . Thus the set  $\{\text{slice}_\chi - g\}_{\chi \in S}$  is a collection of  $2^{\Theta(k)} = n^{\Theta(\log^{1-\beta} n)}$  functions in  $\mathcal{M} - g$  whose pairwise correlations are each at most  $1/2^{\Theta(k)} = 1/n^{\Theta(\log^{1-\beta} n)}$ , and thus the SQ-dimension of  $\mathcal{M} - g$  is at least  $n^{\Theta(\log^{1-\beta} n)}$ . ■

### 3.2. The depth-3 construction

It remains to show that every function in  $\mathcal{M}$  has a depth-3 monotone formula.

**Theorem 6** *Let  $\chi$  be any parity function over some subset of the variables  $x_1, \dots, x_k$  where  $k = \log^{2-\beta}(n)$  for  $\beta$  any absolute constant in  $(0, 1)$ . Then the  $k$ -variable slice function  $\text{slice}_\chi$  is computed by an  $n^{o(1)}$ -size, depth-3 monotone formula.*

**Proof** Let  $\text{Th}_j^k$  be the  $k$ -variable threshold function that outputs TRUE if at least  $j$  of the  $k$  inputs are set to 1, and FALSE otherwise. The threshold function  $\text{Th}_j^k$  can be computed by a monotone formula of size  $n^{o(1)}$  and depth 3 using the construction of Klawe et al. (1984).

Let  $\chi$  be a parity function on  $j$  out of the first  $k$  variables. For  $x \in \{0, 1\}^k$  let  $x^1$  refer to the  $j$  variables of  $\chi$  and  $x^2$  refer to the remaining  $k - j$  variables. We claim that

$$\text{slice}_\chi(x) = \bigvee_{\text{odd } i < j} [\text{Th}_i^j(x^1) \wedge \text{Th}_{k/2-i}^{k-j}(x^2)].$$

To see this, note that if an input  $x$  has fewer than  $k/2$  ones, then there can be no  $i$  such that  $\text{Th}_i^j(x^1)$  and  $\text{Th}_{k/2-i}^{k-j}(x^2)$  both hold, so this function outputs FALSE as it should. If  $x$  has more than  $k/2$  ones, some  $\ell$  of them are in  $x^1$ , and at least  $k/2 - \ell + 1$  of them are in  $x^2$ . If  $\ell$  is odd then  $i = \ell$  makes the OR output TRUE, and if  $\ell$  is even then  $i = (\ell - 1)$  makes the OR output TRUE. Finally, if  $x$  has exactly  $k/2$  ones, and an odd number of them are in  $x^1$ , the formula is satisfied; if an even number of them are in  $x^1$ , the formula is not satisfied.

Each  $\text{Th}_i^j$  and  $\text{Th}_{k/2-i}^{k-j}$  can be computed by a  $n^{o(1)}$ -size, depth-3 monotone formula with an OR on top (Klawe et al., 1984). Using the distributive law we can convert  $\text{Th}_i^j(x^1) \wedge \text{Th}_{k/2-i}^{k-j}(x^2)$  to also be a  $n^{o(1)}$ -size, depth-3 monotone formula with an OR on top. This OR can be collapsed with the top  $\lceil j/2 \rceil$ -wise OR, yielding a  $n^{o(1)}$ -size, depth-3 monotone formula for  $\text{slice}_\chi$ .  $\blacksquare$

We have thus established:

**Theorem 7** *For some  $\epsilon = 1/(\log n)^{\Theta(1)}$ , the class of  $n^{o(1)}$ -size, depth-3 monotone formulas has Strong SQ-Dimension  $n^{\omega(1)}$ .*

As an immediate corollary, by Theorem 4 we get:

**Corollary 8** *The class of  $n^{o(1)}$ -size, depth-3 monotone formulas is not SQ-learnable to some accuracy  $1 - 1/(\log n)^{\Theta(1)}$  in  $\text{poly}(n)$  time.*

### 3.3. Reduction from learning juntas

We now show that ideas from the proof of Theorem 6 can also be used to reduce learning of other non-monotone function classes to learning of shallow monotone formulas. Namely, we give the following reduction from learning of juntas over  $k = \log^2 n / \log \log n$  variables to learning of depth-3 monotone formulas. The  $i$ th variable of a Boolean function  $f$  is said to be *relevant* if there exist inputs  $x$  and  $y$  in  $\{0, 1\}^n$  that differ only on the  $i$ th coordinate such that  $f(x) \neq f(y)$ . A  $j$ -junta is a function that has at most  $j$  relevant variables.

**Theorem 9** *Let  $A$  be a uniform distribution PAC learning algorithm that learns the class of  $\text{poly}(n)$ -size depth-3 monotone formulas to accuracy  $\epsilon$  in time polynomial in  $n$  and  $1/\epsilon$ . Then there exists a uniform-distribution PAC learning algorithm  $C$  that exactly learns the class of  $\log(n)$ -juntas where the relevant variables are chosen from the first  $k = \log^2(n) / \log \log(n)$  variables in time polynomial in  $n$ .*

The proof of Theorem 9 appears in the full version.

We note that Theorem 9 is incomparable to Corollary 8. It is easy to translate Theorem 9 into the SQ model. As a result we would obtain a superpolynomial lower bound for strong SQ learning monotone depth-3 formulas. This is true since a junta can be a parity function and there are at least  $\binom{\log^2 n / \log \log n}{\log n} = n^{\Omega(\log \log n)}$  different parities in the above class of juntas. However both the lower bound and the accuracy parameter in Corollary 8 are substantially better. The better accuracy parameter is required for hardness amplification using a single additional level.

In the next section we introduce hardness amplification machinery that will enable us to extend the SQ learning hardness result to accuracy  $\frac{1}{2} + o(1)$  (for depth-4 formulas).

#### 4. Hardness Amplification for Uniform Distribution Learning

O’Donnell (2004) developed a general technique for hardness amplification. His approach, which may be viewed as a generalization of Yao’s XOR lemma, gives a bound on the hardness of  $g \otimes f = g(f(x_1), \dots, f(x_k))$  where  $f$  is a “mildly” hard function and  $g$  is an arbitrary  $k$ -bit combining function.

At a high-level O’Donnell’s proof has three components. The first component shows the existence of a circuit weakly approximating  $f$  over any  $\delta$ -fraction of the domain whenever there exists a circuit for  $g \otimes f$  that outperforms the expected bias of  $g$  (see definition below). The second part of O’Donnell’s proof uses Impagliazzo’s hard-core lemma (Impagliazzo, 1995) to obtain a  $\delta$ -approximating circuit for  $f$  given circuits that weakly approximate  $f$  over any  $\delta$ -fraction of the domain. The third component is the construction of combining functions that have the desired expected bias.

The first of the two primary obstacles in translating the result to the learning framework is that the first component uses non-uniform advice that depends on  $f$ . This advice is, in general, not available to the learning algorithm<sup>2</sup>. A substantial effort was devoted to obtaining (more) uniform versions of O’Donnell’s result, most notably by Trevisan (Trevisan, 2003, 2005). Both of his reductions are uniform and do not use access to  $f$  but neither is sufficient for our purposes. The first reduction only amplifies to accuracy  $3/4 + o(1)$  (Trevisan, 2003) and the second reduction uses a specific combining function of non-constant circuit depth (Trevisan, 2005). At the same time a learning algorithm has a form of access to  $f$  (either random examples or statistical queries) and hence hardness amplification need not be independent of  $f$  (or “black-box”). Indeed, it is not hard to show that Trevisan’s simpler and more uniform version of the first component (Trevisan, 2003) can be simulated using random examples of  $f$  in place of non-uniform advice (Trevisan, 2010). However, it is unclear if this approach can be used with access only to statistical queries. To solve this problem we show a uniform and general version of the first component, namely an algorithm that given a circuit for  $g \otimes f$  produces a short list of circuits that, with significant probability, contains a circuit weakly approximating  $f$  over any  $\delta$ -fraction of the domain chosen in advance (Lem. 11). The algorithm does not use access to  $f$  but can use statistical queries to find the weakly approximating circuit among the candidate circuits.

---

2. To avoid a potential source of confusion we remark that while our SQ learning lower bounds are information-theoretic and hence allow obtaining an SQ learning algorithm that uses non-uniform advice, such advice cannot depend on  $f$ .

The second obstacle is the fact that in order to obtain a circuit for  $g \otimes f$  a learning algorithm needs to simulate a statistical query oracle for  $g \otimes f$  using a statistical query oracle for  $f$  (when learning from random examples simulation of random examples of  $g \otimes f$  is trivial). We show that this is possible by giving a procedure that uses a function  $\psi$  that approximates  $f$  in place of  $f$  to answer statistical queries for  $g \otimes f$ . To create such  $\psi$  we use a form of gradient descent to  $f$  in which the equivalent of the gradient can be generated whenever  $\psi$  cannot be used in place of  $f$  to answer a statistical query for  $g \otimes f$ . The number of steps of the gradient descent is bounded and therefore this method produces correct answers to statistical queries for  $g \otimes f$ .

We replace the second component (the hard-core lemma) with “smooth boosting,” a technique from computational learning theory which is known to be analogous to hard-core set constructions (Klivans and Servedio, 2003).

Finally, for the third component we need to show that appropriate hardness amplification can be obtained by using only one additional level of depth. By combining balanced Talagrand CNF and the complement of the “tribes” DNF with carefully chosen parameters and using analysis from (O’Donnell, 2004; Mossel and O’Donnell, 2003), we demonstrate hardness amplification from  $1 - \log^{-\alpha} n$  accuracy to  $1/2 + 2^{-\log^\beta n}$  accuracy using a small monotone CNF as a combining function, where  $\alpha$  and  $\beta$  are positive constants (Lem. 16).

**Notation and Terminology.** For  $g$  a  $k$ -variable Boolean function and  $f$  an  $n$ -variable Boolean function, we write  $g \otimes f$  to denote the  $nk$ -variable function  $g(f(x_1), \dots, f(x_k))$ . For  $\mathcal{F}$  a class of  $n$ -variable functions and  $g$  a fixed  $k$ -variable combining function, we write  $\mathcal{F}^g$  to denote the class  $\{g \otimes f : f \in \mathcal{F}\}$ .

Let  $P_\delta^k$  denote the distribution of random restrictions  $\rho$  on  $k$  coordinates, in which each coordinate is mapped independently to  $\star$  with probability  $\delta$ , to 0 with probability  $(1-\delta)/2$ , and to 1 with probability  $(1-\delta)/2$ . We write  $h_\rho$  for the function given by applying restriction  $\rho$  to the function  $h$ . For a  $k$ -variable  $\pm 1$ -valued function  $h$  we write  $\text{bias}(h)$  to denote  $\max\{\Pr[h = -1], \Pr[h = 1]\}$ . The *expected bias of  $h$  at  $\delta$*  is  $\text{ExpBias}_\delta(h) = \mathbf{E}_\rho[\text{bias}(h_\rho)]$ , where  $\rho$  is a random restriction from  $P_\delta^k$ .

#### 4.1. Hardness Amplification in the PAC Setting

The most significant use of non-uniformity in the first component of O’Donnell’s proof is the lemma asserting that if one can predict a Boolean function on the hypercube noticeably better than the function’s bias then there exist two adjacent points of the hypercube on which predictions are noticeably different (O’Donnell, 2004). We start by showing an average-case version of this lemma by proving that predictions need to be different on average over all edges of the hypercube.

**Lemma 10** *Given two functions  $h : \{0, 1\}^k \rightarrow \{-1, 1\}$  and  $p : \{0, 1\}^k \rightarrow [0, 1]$ , suppose that*

$$\frac{1}{2^k} \left( \sum_{x:h(x)=1} p(x) + \sum_{x:h(x)=-1} (1-p(x)) \right) \geq \text{bias}(h) + \epsilon. \quad (2)$$

*Then  $\mathbf{E}_{(x,y)}[|p(x) - p(y)|] \geq 4\epsilon^2/k$  where  $(x, y)$  is a randomly and uniformly chosen edge in the Boolean hypercube  $\{0, 1\}^k$ .*

**Proof** Let us assume without loss of generality that  $h$  is biased towards 1. By the Poincaré inequality over the discrete cube we know that for any function  $p$  over  $\{0, 1\}^k$ :

$$\mathbf{Var}[p] = \mathbf{E}[p^2] - \mathbf{E}[p]^2 \leq \frac{k}{4} \mathbf{E}_{(x,y)}[(p(x) - p(y))^2].$$

The range of  $p$  is  $[0, 1]$ , so  $\mathbf{E}_{(x,y)}[|p(x) - p(y)|] \geq \mathbf{E}_{(x,y)}[(p(x) - p(y))^2] \geq 4\mathbf{Var}[p]/k$ . It is now sufficient to prove that  $\mathbf{Var}[p] \geq \epsilon^2$ .

Let  $b := \text{bias}(h) = \Pr[h = 1] \geq 1/2$ . We can rewrite Equation 2 as

$$\begin{aligned} b + \epsilon &\leq \frac{1}{2^k} \left( \sum_{x:h(x)=1} p(x) + \sum_{x:h(x)=-1} (1 - p(x)) \right) \\ &= \mathbf{E}[h(x)(p(x) - \mathbf{E}[p(x)])] + \mathbf{E}[p(x)]b + (1 - \mathbf{E}[p(x)])(1 - b). \end{aligned}$$

As  $b \geq 1/2$ ,  $b\mathbf{E}[p] + (1 - b)(1 - \mathbf{E}[p]) < b$ , and thus  $\mathbf{E}[h(x)(p(x) - \mathbf{E}[p(x)])] \geq \epsilon$ . Because  $h(x) \in \{-1, 1\}$  we obtain  $\mathbf{E}[|p - \mathbf{E}[p]|] \geq \epsilon$ . Using the Cauchy-Schwarz inequality, we get  $\mathbf{Var}[p] = \mathbf{E}[(p - \mathbf{E}[p])^2] \geq \mathbf{E}[|p - \mathbf{E}[p]|]^2 \geq \epsilon^2$ . ■

Suppose we are given a circuit  $C$  that approximates  $g \otimes f$  sufficiently well that it outperforms the expected bias of  $g$ . Roughly speaking, the following lemma shows that for any large enough set  $S$ , from  $C$  we can extract a circuit  $C'$  that weakly approximates  $f$  over the inputs in  $S$ .

**Lemma 11** *There is a randomized algorithm **Extract** with the following property: For any:*

1. *Parameters  $0 < \epsilon \leq 1/2$ ,  $0 < \eta < 1$ , subset  $S \subseteq \{0, 1\}^n$  such that  $|S| = \eta 2^n$ , Boolean function  $g$  over  $\{0, 1\}^k$ , and*
2. *Boolean function  $f$  such that  $\text{bias}(f) \leq 1/2 + \epsilon/(8k)$  and  $\text{bias}(f|_S) \leq 1/2 + \epsilon^2/(4k)$ ,*

*given a circuit  $C$  over  $\{0, 1\}^{k \times n}$  s.t.*

$$\begin{aligned} \mathbf{Pr}_{\mathcal{U}^k}[C = g \otimes f] &= \mathbf{Pr}_{(x_1, \dots, x_k) \in \{0, 1\}^{k \times n}}[C(x_1, \dots, x_k) = g(f(x_1), \dots, f(x_k))] \\ &\geq \text{ExpBias}_\eta(g) + \epsilon, \end{aligned}$$

*the algorithm **Extract** returns an  $n$ -input circuit  $C'$  such that with probability at least  $\epsilon^2/k$  (over the randomness of **Extract**) we have  $\mathbf{Pr}_{x \in S}[C'(x) = f(x)] \geq 1/2 + \epsilon^2/(2k)$ . The algorithm **Extract** runs in time  $O(nk + |C|)$  and the circuit  $C'$  is of size at most  $|C|$ .*

The proof of Lemma 11 appears in the full version.

As we will see below, two key properties of this lemma are that **Extract** is efficient and is oblivious of both  $f$  and  $S$ . The second property is crucial for hardness amplification in the SQ model. In the second part of the proof, we show how an algorithm  $A$  that learns the combined class  $\mathcal{F}^g$  to moderate accuracy can be used to obtain an algorithm  $B$  that learns the original class  $\mathcal{F}$  to high accuracy. This is exactly the well-studied ‘weak

learning  $\implies$  strong learning” paradigm of *boosting* in computational learning theory (see Schapire, 1990, 2001 for introductions to boosting). Roughly speaking, boosting algorithms are automatic procedures that can be used to convert any weak learning algorithm (that only achieves low accuracy slightly better than  $1/2$ ) into a strong learning algorithm (that achieves high accuracy close to  $1$ ). Boosting algorithms work by repeatedly running the weak learning algorithm under a sequence of carefully chosen probability distributions  $\mathcal{D}_1, \mathcal{D}_2, \dots$ , obtaining weak hypotheses  $h_1, h_2, \dots$ . If each hypothesis  $h_i$  has non-negligible accuracy under the distribution  $\mathcal{D}_i$  that was used to generate it, then the boosting guarantee ensures that the final hypothesis  $h$  (which combines  $h_1, h_2, \dots$ ) has high accuracy under the original distribution.

Since we require the set  $|S|$  to be “large” (recall the statement of Lemma 11), we will need to use a so-called “smooth boosting algorithm” such as the algorithm by Servedio (2003). A  $1/\delta$ -smooth boosting algorithm is a boosting algorithm with the following property: if the original distribution is uniform over a finite domain  $X$  (as is the case for us here), then in learning to final accuracy  $\delta$ , every distribution  $\mathcal{D}_i$  that the smooth boosting algorithm constructs will be “ $1/\delta$ -smooth,” meaning that it puts probability weight at most  $\frac{1}{\delta} \cdot \frac{1}{|X|}$  on any example  $x \in X$ . Such  $1/\delta$ -smooth distributions correspond naturally to large sets  $S$  (of size  $\delta 2^n$ ) in Lemma 11.

So at a high level, we use a smooth boosting algorithm, and for each smooth distribution that it constructs we use **Extract** several times to generate a set of candidate weak hypotheses (recall that **Extract** constructs a “good”  $C$  only with some nonnegligible probability). These hypotheses are then tested using uniform examples (filtered according to the current smooth distribution; since the distribution is smooth this does not incur much overhead), and we identify one which has the required nonnegligible accuracy. The boosting guarantee ensures that the combined hypothesis has accuracy  $1 - \delta$  under the original (uniform) distribution.

Having sketched the intuition for the second stage, we now state the main hardness amplification theorem for PAC learning.

**Theorem 12** *Let  $\mathcal{F}$  be a class of functions such that for every  $f \in \mathcal{F}$ ,  $\text{bias}(f) \leq 1/2 + \epsilon/(8k)$ . Let  $A$  be a uniform distribution PAC learning algorithm that learns  $\mathcal{F}^g$  to accuracy  $\text{ExpBias}_\delta(g) + \epsilon$ . There exists a uniform-distribution PAC learning algorithm  $B$  that learns  $\mathcal{F}$  to accuracy  $1 - \delta$  in time  $O(T_1 \cdot T_2 \cdot \text{poly}(n, k, 1/\epsilon, 1/\delta))$  where  $T_1$  is the time required to evaluate  $g$  and  $T_2$  is the running time of  $A$ .*

**Proof** Let  $f$  denote the unknown target function. We will first simulate  $A$  to obtain a circuit  $C$  that is  $(\text{ExpBias}_\delta(g) + \epsilon)$ -close to  $g \otimes f$ . To generate a random example of  $g \otimes f$  we simply draw  $k$  random examples of  $f$ :  $(x_1, \ell_1), \dots, (x_k, \ell_k)$  and give the example  $((x_1, \dots, x_k), g(\ell_1, \dots, \ell_k))$  to  $A$ .

We now use  $C$  to produce weak hypotheses on distributions produced by the  $1/\delta$ -smooth boosting algorithm by Servedio (2003) (here  $\delta$  refers to the desired accuracy parameter).

Let  $D_t(x)$  denote the distribution obtained at step  $t$  of boosting and let  $h_1, \dots, h_{t-1}$  be the hypotheses obtained in the previous stages of boosting. Let  $M = 2^n L_\infty(D_t)$  and let  $S_t$  denote a set obtained by including each point  $x \in \{0, 1\}^n$  randomly with probability  $D_t(x)/M$ . As it is easy to see (e.g., Impagliazzo, 1995), for any function  $h$  fixed independently of the random choices that determine  $S_t$ , with probability at least  $1 - 2^{-n/2}$  (over

the choice of  $S_t$ )  $|\Pr_{D_t}[h = f] - \Pr_{S_t}[h = f]| \leq 2^{-n/2}$ . Therefore for our purposes we can treat  $\Pr_{S_t}[h = f]$  as equal to  $\Pr_{D_t}[h = f]$ .

If  $\text{bias}(f|_{S_t}) \geq 1/2 + \epsilon^2/(4k)$  then  $\Pr_{S_t}[f = b] \geq 1/2 + \epsilon^2/(4k)$  for  $b \in \{-1, 1\}$ . Otherwise, by Lemma 11, with probability at least  $\epsilon^2/k$  the algorithm **Extract** returns a circuit  $C_1$  such that  $\Pr_{S_t}[C_1 = f] \geq 1/2 + \epsilon^2/(2k)$ . As it is easy to see from the analysis by Servedio (2003), the value  $D_t(x)/M$  equals  $\mu_t(f(x), h_1(x), \dots, h_{t-1}(x))$  for a fixed function  $\mu_t$  defined by the boosting algorithm. This allows the learning algorithm to generate random examples from  $D_t(x)$  by filtering random and uniform examples using  $\mu_t$ . In particular, we can estimate  $\Pr_{D_t}[h = f]$  to accuracy  $\epsilon^2/(12k)$  and confidence  $1/2$  using  $\tilde{O}(k^2/\epsilon^4\delta)$  random and uniform examples in order to test whether either  $-1, 1$  or  $C'$  give a good weak hypothesis (the  $1/\delta$  factor in the number of examples suffices because we are using a  $1/\delta$ -smooth boosting algorithm). By repeating the execution of **Extract** a total of  $O(\epsilon^{-2} \cdot k \log(k/\epsilon\delta))$  times we can ensure that with probability at least  $2/3$  this weak learning step is successful in all  $O(k^2/(\epsilon^4\delta))$  boosting stages that the booster by Servedio (2003) requires. This implies that the boosting algorithm produces a  $(1 - \delta)$ -accurate hypothesis with probability at least  $2/3$ . It is easy to verify that the running time of this algorithm is as claimed. ■

**Remark 13** *This hardness amplification also applies to algorithms using membership queries since membership queries to  $g \otimes f$  can be easily simulated using membership queries to  $f$ .*

## 4.2. Hardness amplification in the Statistical Query setting

We now establish the SQ version of hardness amplification.

**Theorem 14** *Let  $\mathcal{F}$  be a class of functions such that for every  $f \in \mathcal{F}$ ,  $\text{bias}(f) \leq 1/2 + \epsilon/(8k)$ . Let  $A$  be a uniform-distribution SQ-learning algorithm that learns  $\mathcal{F}^g$  to accuracy  $\text{ExpBias}_\delta(g) + \epsilon$  using queries of tolerance  $\tau$ . There exists a uniform-distribution SQ learning algorithm  $B$  that learns  $\mathcal{F}$  to accuracy  $1 - \delta$  in time  $O(T_1 \cdot T_2 \cdot \text{poly}(n, k, 1/\epsilon, 1/\delta))$  using queries of tolerance  $\Omega(\delta \cdot \min\{\tau/k, \epsilon^2/k\})$ , where  $T_1$  is the time required to evaluate  $g$  and  $T_2$  is the running time of  $A$ .*

**Proof** The main challenge in translating the result to SQ-learning is to simulate SQs for  $g \otimes f$  using SQs for  $f$ . Given the circuit  $C$  we can proceed exactly as in the proof of Theorem 12 but use SQs of tolerance  $\Omega(\delta\epsilon^2/k)$  to estimate the bias of  $f$  on  $S_t$  or to test whether the output of **Extract** is a weak hypothesis.

We now describe how to simulate statistical queries to  $g \otimes f$ . The distribution is known to be uniform therefore it is sufficient to answer only correlational statistical queries of  $A$  (Bshouty and Feldman, 2002), namely, it is sufficient to be able to estimate  $\mathbf{E}_{\mathcal{U}^k}[\phi \cdot (g \otimes f)]$  within  $\tau/2$ , where  $\phi$  is a Boolean function over  $\{0, 1\}^{kn}$ . To estimate  $\mathbf{E}_{\mathcal{U}^k}[\phi \cdot (g \otimes f)]$  we plan to use random sampling with an approximation to  $f$  used in place of  $f$ . We refer to the approximation as  $\psi_r(x)$ . However, before doing so, we first test whether  $\psi_r$  is suitable to be used as a replacement. In the main technical claim we prove that if  $\psi_r(x)$  cannot be used to replace  $f$  then we can find a way to update  $\psi_r$  to  $\psi_{r+1}$  which is closer to  $f$  than  $\psi_r$  in  $L_2$  distance. The number of such updates will be bounded and therefore we will eventually obtain  $\psi_r$  that can be used in place of  $f$ . Formally, let  $\psi_0(x) \equiv 0$  and for a function

$\psi_r(x) \in \mathcal{F}_1^\infty$  we denote by  $\Psi_r(x)$  the random  $\{-1, 1\}$  variable with expectation  $\psi_r(x)$ . We also denote by  $g \otimes \Psi_r$  the random variable obtained by applying  $g$  to  $k$  evaluations of  $\Psi_r$ .

**Lemma 15** *For  $i \in [k]$  and  $y = y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_k \in \{0, 1\}^n$  and any function  $\phi$  over  $\{0, 1\}^{kn}$ , we denote  $\phi_{i,y}(x_i) = \phi(y_1, y_2, y_{i-1}, x_i, y_{i+1}, \dots, y_k)$ . Let  $\lambda = |\mathbf{E}_{\mathcal{U}^k}[\phi \cdot g \otimes f] - \mathbf{E}_{\mathcal{U}^k, \Psi}[\phi \cdot g \otimes \Psi]|$ . Then for randomly and uniformly chosen  $i, y$ , with probability at least  $\lambda/(4k)$ ,  $|\mathbf{E}_{\mathcal{U}}[\phi_{i,y}(x_i) \cdot f(x_i)] - \mathbf{E}_{\mathcal{U}}[\phi_{i,y}(x_i) \cdot \psi(x_i)]| \geq \lambda/(2k)$ .*

**Proof** First we denote by  $g \otimes f^{i,\Psi}$  the  $i$ -th hybrid between  $g \otimes f$  and  $g \otimes \Psi$ . Namely, the randomized function  $g \otimes f^{i,\Psi}(x) = g(f(x_1), \dots, f(x_i), \Psi(x_{i+1}), \dots, \Psi(x_k))$ . Now,  $g \otimes f^{k,\Psi} = g \otimes f$  and  $g \otimes f^{0,\Psi} = g \otimes \Psi$ . Hence we can write,

$$|\mathbf{E}_{\mathcal{U}^k}[\phi \cdot g \otimes f] - \mathbf{E}_{\mathcal{U}^k, \Psi}[\phi \cdot g \otimes \Psi]| = k \cdot |\mathbf{E}_{i \in [k]} [\mathbf{E}_{\mathcal{U}^k, \Psi}[\phi \cdot g \otimes f^{i,\Psi}] - \mathbf{E}_{\mathcal{U}^k, \Psi}[\phi \cdot g \otimes f^{i-1,\Psi}]]|$$

We now split the random and uniform choice over  $\{0, 1\}^{kn}$  into choosing  $y = y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_k \in \{0, 1\}^n$  and  $x_i \in \{0, 1\}^n$  randomly and uniformly.

$$|\mathbf{E}_{i \in [k], y} [\mathbf{E}_{\mathcal{U}, \Psi}[(\phi \cdot g \otimes f^{i,\Psi})_{i,y}(x_i)] - \mathbf{E}_{\mathcal{U}, \Psi}[(\phi \cdot g \otimes f^{i-1,\Psi})_{i,y}(x_i)]]| \geq \lambda/k . \quad (3)$$

We claim that

$$\mathbf{E}_{\mathcal{U}, \Psi}[(\phi \cdot g \otimes f^{i,\Psi})_{i,y}(x_i)] = \mathbf{E}_{\mathcal{U}} [\phi_{i,y} \cdot \mathbf{E}_{\Psi}[(g \otimes f^{i,\Psi})_{i,y}(x_i)]] = \mathbf{E}_{\mathcal{U}}[\phi_{i,y} \cdot (\alpha_{i,y} f(x_i) + \beta_{i,y})] .$$

Here  $\alpha_{i,y}$  and  $\beta_{i,y}$  are constants in  $[-1, 1]$ .

To see this assume for simplicity that  $\Psi$  is deterministic. Then

$$(g \otimes f^{i,\Psi})_{i,y}(x_i) = g(f(y_1), \dots, f(y_{k-1}), f(x_i), f(y_{k-1}), \dots, f(y_k)) .$$

All the variables of  $g$  are fixed except for the  $i$ -th and therefore this restriction of  $g$  equals  $1, -1, f(x_i)$  or  $-f(x_i)$ . This corresponds to  $\alpha_{i,y}, \beta_{i,y} \in \{-1, 0, 1\}$  and exactly one of them is non-zero. For randomized  $\Psi$  we obtain a fixed convex combination of the deterministic cases that can be represented by  $\alpha_{i,y}, \beta_{i,y} \in [-1, 1]$ . Similarly,

$$\mathbf{E}_{\mathcal{U}, \Psi}[(\phi \cdot g \otimes f^{i-1,\Psi})_{i,y}(x_i)] = \mathbf{E}_{\mathcal{U}}[\phi_{i,y} \cdot (\alpha_{i,y} \psi(x_i) + \beta_{i,y})] .$$

By substituting this into equation (3), we obtain

$$|\alpha_{i,y} \cdot \mathbf{E}_{i \in [k], y} [\mathbf{E}_{\mathcal{U}}[\phi_{i,y} \cdot f(x_i)] - \mathbf{E}_{\mathcal{U}}[\phi_{i,y} \cdot \psi(x_i)]]| \geq \lambda/k .$$

By the averaging argument, we obtain that with probability at least  $\lambda/(4k)$  over the choice of  $i$  and  $y$ ,  $|\mathbf{E}_{\mathcal{U}}[\phi_{i,y} \cdot f(x_i)] - \mathbf{E}_{\mathcal{U}}[\phi_{i,y} \cdot \psi(x_i)]| \geq \lambda/(2k)$ . ■

If  $|\mathbf{E}_{\mathcal{U}^k}[\phi \cdot g \otimes f] - \mathbf{E}_{\mathcal{U}^k, \Psi_r}[\phi \cdot g \otimes \Psi_r]| \geq \tau/3$  then with probability at least  $\tau/(12k)$  for a randomly chosen  $\phi_{i,y}$ ,  $|\mathbf{E}_{\mathcal{U}}[\phi_{i,y} \cdot f(x_i)] - \mathbf{E}_{\mathcal{U}}[\phi_{i,y} \cdot \psi(x_i)]| \geq \tau/(6k)$ . Let  $\tau' = \tau/(6k)$ . Now we sample  $\phi_{i,y}$  and test if  $|\mathbf{E}_{\mathcal{U}}[\phi_{i,y} \cdot f(x_i)] - \mathbf{E}_{\mathcal{U}}[\phi_{i,y} \cdot \psi(x_i)]| \leq 2\tau'/3$  using a single SQ of tolerance  $\tau'/6$  and an estimate of  $\mathbf{E}_{\mathcal{U}}[\phi_{i,y} \cdot \psi(x_i)]$  within  $\tau'/6$  obtained using random sampling. It is easy to see that by repeating this procedure  $O(k \log(1/\Delta)/\tau)$  times and using  $O(k \log(1/\Delta)/\tau)$  random samples we can ensure that with probability at least  $1 - \Delta$

some  $\phi_{i',y'}$  will pass the test whenever  $|\mathbf{E}_{\mathcal{U}^k}[\phi \cdot g \otimes f] - \mathbf{E}_{\mathcal{U}^k, \Psi_r}[\phi \cdot g \otimes \Psi_r]| \geq \tau/3$  and also that  $|\mathbf{E}_{\mathcal{U}}[\phi_{i',y'} \cdot f(x_{i'})] - \mathbf{E}_{\mathcal{U}}[\phi_{i',y'} \cdot \psi(x_{i'})]| \geq 2\tau'/3 - \tau'/3 = \tau'/3$  whenever  $\phi_{i',y'}$  passes the test.

If the test was not passed then we estimate  $\mathbf{E}_{\mathcal{U}^k, \Psi}[\phi \cdot g \otimes \Psi]$  within  $\tau/6$  using random sampling and return the estimate as the answer to the query. Using  $O(k \log(1/\Delta)/\tau)$  random samples we can ensure that with probability  $1 - 2\Delta$  the returned estimate is  $\tau/2$  close to  $\mathbf{E}_{\mathcal{U}^k}[\phi \cdot g \otimes f]$ .

Otherwise, we use such  $\phi_{i',y'}$  to update  $\psi_r$  using the idea from Feldman's (2010, Th.3.5) strong SQ characterization:  $\psi_{r+1} = P_1(\psi_r + (\tau'/3) \cdot \phi_{i',y'})$ . Here  $P_1(a)$  is the function that equals  $a$  when  $a \in [-1, 1]$  and equals  $\text{sign}(a)$  otherwise.

As is proved by Feldman (2010, Cl.3.6),  $|\mathbf{E}_{\mathcal{U}}[\phi_{i',y'} \cdot (f(x_{i'}) - \psi(x_{i'}))]| \geq \tau'/3$  implies that  $\mathbf{E}_{\mathcal{U}}[(f - \psi_{r+1})^2] \leq \mathbf{E}_{\mathcal{U}}[(f - \psi_r)^2] - (\tau'/3)^2$ . Therefore at most  $O(k^2/\tau^2)$  such updates are possible giving an upper bound on the additional time required to produce the desired estimates to all the SQs of  $A$  (a similar bound can also be obtained using a different update method of Trevisan et al. (2009)). For an appropriate  $\Delta = \text{poly}(k, 1/\tau)$  we can make sure that the success probability is at least  $2/3$ . ■

## 5. Amplified Hardness for SQ Learning of Depth-4 Monotone Formulas

We begin this section by showing how a refinement of the constructions and analysis from (O'Donnell, 2004; Mossel and O'Donnell, 2003) can be used to obtain a small monotone CNF with low expected bias. Specifically, we prove the following lemma.

**Lemma 16** *For every  $0 < \gamma < 1/2$ , there exists a circuit  $C_{k,m}$  over  $k$  variables such that:*

$$\text{ExpBias}_{1/\sqrt{m}}(C_{k,m}) \leq \frac{1}{2} + 2^{-(\log n)^\gamma},$$

where  $k = 2^{(\log n)^\alpha}$  and  $m = \log^{2-\beta}(n)$  for  $\gamma < \alpha < \beta/2 < 1/2$ , and  $C_{k,m}$  is computable by a monotone CNF of size  $n^{o(1)}$ .

The proof of Lemma 16 appears in the full version.

Coupled with our hardness result for depth-3 monotone formulas, Lemma 16 gives the claimed lower bound for depth-4 monotone formulas.

**Theorem 17** *For every  $0 < \gamma < 1/2$ , the class of  $n^{o(1)}$ -size, depth-4 monotone formulas is not SQ-learnable to accuracy  $\frac{1}{2} + 2^{-(\log n)^\gamma}$  in  $\text{poly}(n)$  time.*

The proof of Theorem 17 appears in the full version.

## Acknowledgments

Homin K. Lee is supported by NSF grant 1019343 subaward CIF-B-108. Rocco A. Servedio is supported by NSF grants CCF-0347282, CCF-0523664 and CNS-0716245, and by DARPA award HR0011-08-1-0069.

## References

- K. Amano and A. Maruoka. On learning monotone boolean functions under the uniform distribution. In *Proceedings of the 13th International Conference on Algorithmic Learning Theory (ALT)*, pages 57–68, 2002.
- Avrim Blum. Learning a function of  $r$  relevant variables. In *Proc. of the 16th Annual Conference on Computational Learning Theory (COLT)*, volume 2777 of *Lecture Notes in Computer Science*, pages 731–733. Springer-Verlag, 2003.
- Avrim Blum, Merrick L. Furst, Jeffrey Jackson, Michael J. Kearns, Yishay Mansour, and Steven Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proc. 26th Annual ACM Symposium on Theory of Computing (STOC)*, pages 253–262. ACM Press, 1994.
- Avrim Blum, Carl Burch, and John Langford. On learning monotone boolean functions. In *Proc. 39th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 408–415. IEEE Computer Society Press, 1998.
- Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM*, 50(4):506–519, July 2003. Prelim. ver. in *Proc. of STOC’00*.
- Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the SuLQ framework. In *PODS*, pages 128–138, 2005.
- D. Boneh and R. Lipton. Amplification of weak learning over the uniform distribution. In *Proceedings of the Sixth Annual Workshop on Computational Learning Theory*, pages 347–351, 1993.
- Nader H. Bshouty and Vitaly Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2:359–395, 2002. Prelim. ver. in *Proc. of COLT’01*.
- Nader H. Bshouty and Christino Tamon. On the Fourier spectrum of monotone functions. *Journal of the ACM*, 43(4):747–770, 2006.
- T. Bylander. Learning noisy linear threshold functions. Available at: <http://ringer.cs.utsa.edu/research/AI/bylander/pubs/pubs.html>, 1998.
- Tom Bylander. Learning linear threshold functions in the presence of classification noise. In *Proc. of the 7th Annual Conference on Computational Learning Theory (COLT)*, pages 340–347, 1994.
- D. Dachman-Soled, H. Lee, T. Malkin, R. Servedio, A. Wan, and H. Wee. Optimal cryptographic hardness of learning monotone functions. In *Proc. 35th International Colloquium on Algorithms, Languages and Programming (ICALP)*, pages 36–47, 2008.
- John Dunagan and Santosh Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. In *Proc. 36th Annual ACM Symposium on Theory of Computing (STOC)*, pages 315–320. ACM Press, 2004.

- V. Feldman. A complete characterization of statistical query learning with applications to evolvability. *CoRR*, abs/1002.3183, 2010. Prelim. ver. in *Proc. of FOCS'09*.
- Vitaly Feldman. Evolvability from learning algorithms. In *Proc. 40th Annual ACM Symposium on Theory of Computing (STOC)*. ACM Press, 2008.
- T. Hancock and Y. Mansour. Learning monotone  $k$ - $\mu$  DNF formulas on product distributions. In *Proceedings of the Fourth Annual Conference on Computational Learning Theory*, pages 179–193, 1991.
- R. Impagliazzo. Hard-core distributions for somewhat hard problems. In *Proceedings of the Thirty-Sixth Annual Symposium on Foundations of Computer Science*, pages 538–545, 1995.
- Jeffrey C. Jackson, Homin K. Lee, Rocco A. Servedio, and Andrew Wan. Learning random monotone DNF. In *11th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems and 12th International Workshop on Randomization and Computation (RANDOM-APPROX)*, pages 483–497. Springer-Verlag, 2008.
- Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In *FOCS*, pages 531–540, 2008.
- Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998. Prelim. ver. in *Proc. of STOC'93*.
- Michael J. Kearns, Ming Li, and Leslie G. Valiant. Learning Boolean formulas. *Journal of the ACM*, 41(6):1298–1328, 1994. Prelim. ver. in *Proc. of STOC'87*.
- Maria M. Klawe, Wolfgang J. Paul, Nicholas Pippenger, and Mihalis Yannakakis. On monotone formulae with restricted depth. In *Proc. 16th Annual ACM Symposium on Theory of Computing (STOC)*, pages 480–487. ACM Press, 1984.
- Adam Klivans and Rocco A. Servedio. Boosting and hard-core sets. *Machine Learning*, 53(3):217–238, 2003. Prelim. ver. in *Proc. of FOCS'99*.
- Adam R. Klivans and Alexander A. Sherstov. Unconditional lower bounds for learning intersections of halfspaces. In *Proc. of the 19th Annual Conference on Computational Learning Theory (COLT)*, 2006.
- E. Mossel and R. O'Donnell. On the noise sensitivity of monotone functions. *Random Structures and Algorithms*, 23(3):333–350, 2003.
- Elchanan Mossel, Ryan O'Donnell, and Rocco A. Servedio. Learning functions of  $k$  relevant variables. *SIAM Journal on Computing*, 37(3):421–434, 2007. Prelim. ver. in *Proc. of STOC'03*.
- R. O'Donnell and R. Servedio. Learning monotone decision trees in polynomial time. *SIAM J. Comput.*, 37(3):827–844, 2007.
- Ryan O'Donnell. Hardness amplification within NP. *Journal of Computer and System Sciences*, 69(1):68–94, 2004. Prelim. ver. in *Proc. of STOC'02*.

- Ryan O'Donnell and Karl Wimmer. KKL, Kruskal-Katona, and monotone nets. In *Proc. 50th IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE Computer Society Press, 2009.
- Y. Sakai and A. Maruoka. Learning monotone log-term DNF formulas under the uniform distribution. *Theory of Computing Systems*, 33:17–33, 2000.
- Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990. Prelim. ver. in *Proc. of FOCS'1989*.
- Robert E. Schapire. The boosting approach to machine learning: An overview. In *MSRI Workshop on Nonlinear Estimation and Classification*, 2001.
- R. Servedio. Smooth boosting and learning with malicious noise. *Journal of Machine Learning Research*, 4:633–648, 2003.
- R. Servedio. On learning monotone DNF under product distributions. *Information and Computation*, 193(1):57–74, 2004.
- Alexander A. Sherstov. Communication complexity under product and nonproduct distributions. *Computational Complexity, Annual IEEE Conference on*, 0:64–70, 2008. ISSN 1093-0159. doi: <http://doi.ieeecomputersociety.org/10.1109/CCC.2008.10>.
- Hans Ulrich Simon. A characterization of strong learnability in the statistical query model. In *Proc. 24th Annual Symposium on Theoretical Aspects of Computer Science (STACS)*, pages 393–404, 2007.
- B. Szörényi. Characterizing statistical query learning:simplified notions and proofs. In *ALT*, pages 186–200, 2009.
- Luca Trevisan. List decoding using the XOR lemma. In *Proc. 44th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 126–135. IEEE Computer Society Press, 2003.
- Luca Trevisan. On uniform amplification of hardness in NP. In *Proc. 37th Annual ACM Symposium on Theory of Computing (STOC)*, pages 31–38. ACM Press, 2005.
- Luca Trevisan. Personal communication, 2010.
- Luca Trevisan, Madhur Tulsiani, and Salil P. Vadhan. Regularity, boosting, and efficiently simulating every high-entropy distribution. In *Proc. 24th Annual IEEE Conference on Computational Complexity (CCC)*, pages 126–136. IEEE Computer Society Press, 2009.
- Leslie G. Valiant. Evolvability. *Journal of the ACM*, 56(1):3.1–3.21, 2009.
- Ke Yang. New lower bounds for statistical query learning. *Journal of Computer and System Sciences*, 70(4):485–509, 2005. Prelim. ver. in *Proc. of COLT'02*.

FELDMAN LEE SERVEDIO

# Complexity-Based Approach to Calibration with Checking Rules

**Dean P. Foster**

*Department of Statistics  
University of Pennsylvania*

**Alexander Rakhlin**

*Department of Statistics  
University of Pennsylvania*

**Karthik Sridharan**

*TTI-Chicago*

**Ambuj Tewari**

*Department of Computer Science  
University of Texas at Austin*

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We consider the problem of forecasting a sequence of outcomes from an unknown source. The quality of the forecaster is measured by a family of checking rules. We prove upper bounds on the value of the associated game, thus certifying the existence of a calibrated strategy for the forecaster. We show that complexity of the family of checking rules can be captured by the notion of a sequential cover introduced in (Rakhlin et al., 2010a). Various natural assumptions on the class of checking rules are considered, including finiteness of Vapnik-Chervonenkis and Littlestone’s dimensions.

## 1. Introduction

As many other papers on calibration, we start with the following motivating example: Consider a weatherman who predicts the probability of rain tomorrow and then observes the binary “rain/no rain” outcome. How can we measure the weatherman’s performance? If we make no assumption on the way Nature selects outcomes, defining a notion of performance is a non-trivial matter. One approach, familiar to the learning community, is to prove regret bounds with respect to some class of strategies. However, in the absence of any assumptions on the sequence, the performance of the comparator will not be favorable, rendering the bounds meaningless. An alternative measure of performance is to ask that the forecaster satisfies certain properties with respect to the sequence. One such natural property is *calibration*. It posits that for all the days that the forecaster predicted a probability  $p$  of rain, the empirical frequency of rain was indeed close to  $p$ . It is not obvious, a priori, that there exists a forecasting strategy calibrated with respect to every  $p$ , no matter what sequences Nature presents. The question was raised in the Bayesian setting by Dawid (1982), followed by the negative result of Oakes (1985), who showed that no deterministic

calibration strategy exists. The first positive result was shown by Foster and Vohra (1998), who provided a randomized calibration strategy.

Calibration is indeed the absolute minimum we should expect from a forecaster. Clearly a forecaster who makes a constant prediction of .6 on the binary sequence 11.001001000011111... for  $\pi$  (which empirically is one half ones and believed by most to be half ones in the limit) should be fired at some point for a failure to be calibrated (Lehrer, 2004). However, forecasting the right overall frequency might not be enough. Indeed, consider a binary sequence “010101...” of “rain/no rain” outcomes. A forecaster predicting 0.5 chance of rain is calibrated, yet such a lousy weatherman should be fired immediately! To cope with the obvious shortcoming of calibration, one may introduce more complex *checking rules* (Kalai et al., 1999; Sandroni et al., 2003; Cesa-Bianchi and Lugosi, 2006), such as “the forecaster should be calibrated on all even rounds.” This additional rule clearly disallows a constant prediction of 0.5 since within the even rounds the empirical frequency is 1. While resolving the problem with the particular sequence “010101...,” the forecaster’s performance might still appear unacceptable (by our standards) on other sequences. We refer to (Sandroni et al., 2003) for further discussion on checking rules.

How rich can we make the set of checking rules while being able to satisfy all of them at the same time? Of course, if checking rules are completely arbitrary, there is no hope, as the rule can be tailored to the particular sequence presented. It is then natural to ask the following questions: What is a sufficient restriction on the class of checking rules? What are the relevant measures of complexity of infinite classes of checking rules? What governs the rates of convergence in calibration? In addressing these matters, we come to questions of martingale convergence for function classes. In particular, this allows us to make a connection to the Vapnik-Chervonenkis theory which measures the complexity of the class using a combinatorial parameter. We can view the classical calibration results as a particular instance of checking rules with a finite VC dimension. To the best of our knowledge, the connection between calibration and statistical learning has not been previously observed.

Our results are based on tools recently developed in (Rakhlin et al., 2010b,a). These papers consider abstract repeated zero-sum games (subsuming Online Learning) and obtain upper bounds on the minimax value via the process of sequential symmetrization. Interestingly, these bounds are attained without explicitly talking about algorithms, and instead focusing on the inherent complexity of the problem. Analogously, in the present paper we prove convergence results which depend on the complexity of the class of checking rules without providing a computationally efficient algorithm (the inefficient algorithm can be recovered from the minimax formulation). We argue that an understanding of what is attainable in terms of satisfying checking rules is necessary before looking for an efficient implementation. Once the inherent complexity of calibration with checking rules is understood, algorithmic questions will arise. While there is an efficient algorithm for classical calibration with two actions (see Foster and Vohra (1998); Abernethy et al. (2011)), the question is still open for more complex classes of checking rules.

Classical decision theory typically divides problems into two pieces, probability and loss, and then combines these (via expectation) for making decisions. Calibrated forecasts allow this same division to be done in the setting of individual sequences: a probabilistic forecast can be made and then a loss function can be optimized as if these probabilities were in fact correct. These decisions can be made in a game theoretic setting, in which case calibrated

forecasts can lead to equilibria in games (Foster and Vohra, 1997; Kakade and Foster, 2008). But unlike traditional decision theory which has viewed this division of decisions into probability and loss as having zero cost, there is a huge cost when using calibration in this way for individual sequences. Namely, the rates of convergence for a calibrated forecast have often been much poorer than the ones generated by optimizing the decisions directly, as is typically done in the experts literature. The cause of this rate difference is that calibration tries to optimize over details that the experts approach would ignore. We present alternative definitions of calibration that address this by focusing attention only on the parts of calibration that translate into difference at the decision-making level. We refer to (Young, 2004) for connections between calibration, decision making, and games.

Another motivation for studying checking rules comes from recent research at the intersection of game theory, learning, and economics, which often involves multiple agents acting in the world (Kakade et al., 2003). Being able to calibrate with respect to a class of checking rules can lead to good guarantees on the quality of actions taken by agents. For instance, one can consider multi-agent decision-making problems in large environments, where the agents only need to calibrate with respect to a small set of checking rules relevant to their decision making.

## 2. Notation

Let  $\mathbb{E}_{x \sim p}$  denote expectation with respect to a random variable  $x$  with a distribution  $p$ . A Rademacher random variable is a symmetric  $\pm 1$ -valued random variable. The notation  $x_{a:b}$  denotes the sequence  $x_a, \dots, x_b$ . The indicator of an event  $A$  is denoted by  $\mathbf{1}\{A\}$ . The set  $\{1, \dots, T\}$  is denoted by  $[T]$ , while the  $(k - 1)$ -dimensional probability simplex in  $\mathbb{R}^k$  is denoted by  $\Delta_k$ . Let  $E_k$  denote the  $k$  vertices of  $\Delta_k$ . The set of all functions from  $\mathcal{X}$  to  $\mathcal{Y}$  is denoted by  $\mathcal{Y}^\mathcal{X}$ , and the  $t$ -fold product  $\mathcal{X} \times \dots \times \mathcal{X}$  is denoted by  $\mathcal{X}^t$ . Whenever a supremum (or infimum) is written in the form  $\sup_a$  without  $a$  being quantified, it is assumed that  $a$  ranges over the set of all possible values which will be understood from the context.

Following (Rakhlin et al., 2010a), we define binary trees as follows. Consider a binary tree of uniform depth  $T$  where every interior node and every leaf is labeled with a value  $X$  chosen from some set  $\mathcal{X}$ . More precisely, given some set  $\mathcal{X}$ , an  $\mathcal{X}$ -valued tree of depth  $T$  is a sequence  $(\mathbf{x}_1, \dots, \mathbf{x}_T)$  of  $T$  mappings  $\mathbf{x}_i : \{\pm 1\}^{i-1} \mapsto \mathcal{X}$ . Unless specified otherwise,  $\epsilon = (\epsilon_1, \dots, \epsilon_T) \in \{\pm 1\}^T$  will define a path. For brevity, we will write  $\mathbf{x}_t(\epsilon)$  instead of  $\mathbf{x}_t(\epsilon_{1:t-1})$ .

## 3. The Setting

In this paper we consider the  $k$ -outcome calibration game (in the weatherman example,  $k = 2$ ). Each outcome is represented by an element of  $E_k$ , whereas the forecast is represented by a point in  $\Delta_k$ . More precisely, the protocol can be viewed as the  $T$ -round game between player (learner) and the adversary (Nature):

**FOR** round  $t = 1, \dots, T$ ,

- the player chooses a mixed strategy  $q_t \in \Delta(\Delta_k)$  (distribution on  $\Delta_k$ )
- the adversary picks outcome  $x_t \in E_k$

- the learner draws  $f_t \in \Delta_k$  from  $q_t$  and observes outcome  $x_t$

**ENDFOR**

Both opponents can base their next move on the history of actions observed so far. In particular, this makes the adversary *adaptive*. Throughout the paper,  $z_t \in \mathcal{Z}$  is given by  $z_t = ((f_1, x_1), \dots, (f_{t-1}, x_{t-1}))$ , the history of actions by both players at round  $t$ . Define the set of all possible histories by  $\mathcal{Z} = \bigcup_{t=1}^T (\Delta_k \times E_k)^t$ .

**Definition 1** A forecast-based checking rule is a binary-valued function  $c : \mathcal{Z} \times \Delta_k \mapsto \{0, 1\}$ .

In other words, a checking rule depends on both the history and the current forecast. For simplicity, we only consider binary-valued checking rules; however, the results can be extended to real-valued functions and will appear in the full version of the paper.

Let  $\zeta$  be a family of checking rules. The goal of the player is to minimize the performance metric

$$\mathbf{R}_T := \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t, f_t) \cdot (f_t - x_t) \right\|$$

for some norm  $\|\cdot\|$  on  $\mathbb{R}^k$ . While the  $\ell_1$  norm is typically used for calibration (Mannor and Stoltz, 2010), we can consider a general  $\ell_p$  norm for  $1 \leq p \leq \infty$ . Informally,  $\mathbf{R}_T$  says that the player needs to be calibrated (that is, average of forecasts close to the actual frequency) for any rule  $c$  that becomes active only on certain rounds. In the asymptotic sense, any rule that is not active infinitely often does not matter for the player.

**Example 1** For classical  $\epsilon$ -calibration, choose  $\zeta = \{c_p(z_t, f_t) = \mathbf{1}\{\|f_t - p\| \leq \epsilon : p \in \Delta_k\}\}$ . In particular,  $\epsilon$ -calibration captures the weather forecasting example discussed earlier. We refer to (Cesa-Bianchi and Lugosi, 2006; Mannor and Stoltz, 2010) for the details on the relationship between  $\epsilon$ -calibration and well-calibration.

**Example 2** Let  $\mathcal{G}$  be an  $\epsilon$  grid of the  $\Delta_k$ . Define

$$\zeta = \{c_A(z_t, f_t) = \mathbf{1}\{\|f_t - a\| \leq \epsilon \text{ for some } a \in A\}\}_{A \in 2^{\mathcal{G}}}.$$

That is,  $c_A$  captures the set of forecasts for which  $f_t$  either over-forecasts or under-forecasts the correct probability of the outcome. This is a much richer set of rules than the previous example and is the implicit set used in the Brier quadratic calibration score used in (Foster and Vohra, 1998). As we will show later, the rate of convergence is much slower than for classical calibration.

**Example 3** Let  $\hat{p}_{\theta,t}$  be the forecast made by a probabilistic model  $P_\theta$ . Using  $\zeta = \{c_{\theta,p}(z_t, f_t) = \mathbf{1}\{\|\hat{p}_{\theta,t} - p\| \leq \epsilon\}\}$  will test if the model  $P_\theta$  is a much better fit to the data than the forecasting rule  $f_t$ . If complexity of the set of models  $\{P_\theta\}$  is controlled, then theorems we will discuss later will guarantee existence of a rule that can do well against this family of tests. This connects to the testing of experts literature (Olszewski and Sandroni, 2009).

Given the set  $\zeta$  of checking rules, when is it possible to find a strategy for the forecaster such that  $\mathbf{R}_T$  goes to zero as  $T$  increases? Instead of using, for instance, Blackwell's approachability to provide a calibration strategy with respect to the class  $\zeta$  (as done in (Foster and Vohra, 1998; Sandroni et al., 2003)), we directly attack the value of the game. Given a  $\theta > 0$ , we define the value of the calibration game as

$$\mathcal{V}_T^\theta(\zeta) := \inf_{q_1} \sup_{x_1} \mathbb{E}_{f_1 \sim q_1} \dots \inf_{q_T} \sup_{x_T} \mathbb{E}_{f_T \sim q_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t, f_t) \cdot (f_t - x_t) \right\| > \theta \right\} \right],$$

where  $q_t$ 's range over all distributions over  $\Delta_k$  and  $x_t$  range over  $E_k$ . Note that the value can be interpreted as the probability of the performance metric  $\mathbf{R}_T$  being larger than  $\theta$  under the stochastic process arising from the successive infima, suprema, and expectations. An upper bound on  $\mathcal{V}_T^\theta(\zeta)$  implies *existence* of a strategy for the learner such that the calibration metric  $\mathbf{R}_T$  is smaller than  $\theta$  with probability at least  $1 - \mathcal{V}_T^\theta(\zeta)$ . Or put more colloquially, our bound on  $\mathcal{V}_T$  is an upper bound on the probability of the weatherman being fired for failure to be calibrated to accuracy  $\theta$ . Alternatively, lower bounds on  $\mathcal{V}_T^\theta(\zeta)$  imply impossibility results for the learner. Note that the definition of value of the game is for a fixed  $\theta$  and number of rounds  $T$ . Thus, it is not obvious how to use the so-called “doubling trick” to get a player strategy that is Hannan consistent for the calibration game. The main difficulty is the dependence of the game (and hence the optimal player strategy) on  $\theta$ . It is possible to define a game where the optimal player strategy will work *uniformly* over all  $\theta$  (see (Rakhlin et al., 2010b)). Once this is done, we can proceed along similar lines as in Mannor and Stoltz (2010) to guarantee the existence of a Hannan consistent strategy for calibration with only an extra logarithmic factor on number of rounds played. However, for simplicity, we stick to the fixed  $\theta, T$  definition above in this paper.

#### 4. General Upper Bound on the Value $\mathcal{V}_T^\theta(\zeta)$

Let  $\delta > 0$  and let  $C_\delta$  be a minimal  $\delta$ -cover of  $\Delta_k$  in the norm  $\|\cdot\|$ . The size of the  $\delta$ -cover can be bounded as

$$|C_\delta| \leq (c_1/(2\delta))^{k-1}. \quad (1)$$

where  $c_1$  is some constant independent of  $k$ , but varying with the choice of the norm  $\|\cdot\|$ . This constant will appear throughout the paper. Further, for any  $p_t \in \Delta_k$ , let  $p_t^\delta \in C_\delta$  be a point in  $C_\delta$  such that  $\|p_t - p_t^\delta\| \leq \delta$ . Slightly abusing the notation, define  $z_t^\delta = ((p_1^\delta, x_1), \dots, (p_{t-1}^\delta, x_{t-1})) \in \mathcal{Z}^\delta \subseteq \mathcal{Z}$  where  $\mathcal{Z}^\delta := \bigcup_{t=1}^T (C_\delta \times E_k)^{t-1}$ . (For the proofs of Lemmata 2–4, see Sec. 7 & Appendix.)

**Lemma 2** *For any  $\theta > 0$ ,*

$$\mathcal{V}_T^\theta(\zeta) \leq \sup_{p_1} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T \sim p_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x_t) \right\| > \theta/2 \right\} \right] \quad (2)$$

for any  $\delta \leq \theta/2$ .

The interleaved suprema and expectations on the right-hand side of (2) can be written more succinctly as

$$\sup_{\mathbf{p}} \mathbb{E} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x_t) \right\| > \theta/2 \right\} \right] \quad (3)$$

where  $\mathbf{p}$  can be either thought of as a joint distribution over sequences  $(x_1, \dots, x_T)$  or as a sequence of conditional distributions  $\{p_t : E_k^{t-1} \rightarrow \Delta_k\}$ . Using the notation of conditional distributions, the expectation in (3) can be expanded as  $\mathbb{E}_{x_1 \sim p_1} \mathbb{E}_{x_2 \sim p_2(\cdot | x_1)} \mathbb{E}_{x_T \sim p_T(\cdot | x_{1:T-1})}$ . Of course, expected value of an indicator is just the probability of the event. The goal is to relate (3) to the probability that the norm  $\|\cdot\|$  of the average of a martingale difference sequence is large. The latter probability is exponentially small by a concentration of measure result which we present next.

**Lemma 3** *For any  $\mathbb{R}^k$ -valued martingale difference sequence  $\{d_t\}_{t=1}^T$  with  $\|d_t\| \leq 1$  a.s. for all  $t \in [T]$ , there exists a  $k$ -dependent constant  $c_k$  such that*

$$\mathbb{P} \left( \left\| \frac{1}{T} \sum_{t=1}^T d_t \right\| > \theta \right) \leq 2 \exp \left( -\frac{T\theta^2}{c_k} \right).$$

In particular,  $c_k = 8k$  for any  $\ell_p$  norm with  $1 \leq p \leq \infty$ .

Armed with a concentration result for martingales, we apply the sequential symmetrization technique (see (Rakhlin et al., 2010b) for the high-probability version). In the lemma below, the supremum is over all binary  $E_k$ -valued trees  $\mathbf{x}$  of depth  $T$ , as well as all binary  $C_\delta$ -valued trees  $\mathbf{p}^\delta$  of depth  $T$ . Given  $\mathbf{x}, \mathbf{p}^\delta$ , let the  $\mathcal{Z}^\delta$ -valued tree  $\mathbf{z}^\delta$  be defined by

$$\mathbf{z}_t^\delta(\epsilon) = ((\mathbf{p}_1^\delta(\epsilon), \mathbf{x}_1(\epsilon)), \dots, (\mathbf{p}_{t-1}^\delta(\epsilon), \mathbf{x}_{t-1}(\epsilon)))$$

for any  $t \in [T]$ . We also write  $\mathbf{z}^{(\mathbf{x}, \mathbf{p}^\delta)}$  instead of  $\mathbf{z}^\delta$  to make the dependence on  $\mathbf{x}, \mathbf{p}^\delta$  explicit.

**Lemma 4** *For  $T > \frac{16c_k \log(4)}{\theta^2}$  and  $\delta \leq \theta/2$ ,*

$$\mathcal{V}_T^\theta(\zeta) \leq 4 \sup_{\mathbf{x}, \mathbf{p}^\delta} \mathbb{P}_\epsilon \left( \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t c(\mathbf{z}_t^\delta(\epsilon), \mathbf{p}_t^\delta(\epsilon)) \mathbf{x}_t(\epsilon) \right\| > \theta/8 \right),$$

where the probability is over an i.i.d. draw of Rademacher random variables  $\epsilon_1, \dots, \epsilon_T$ .

What has been achieved by this lemma? We were able to pass from the quantity in (3) which is defined with respect to a complicated stochastic process to a much simpler process. It is defined by fixing the worst-case trees (in the spaces of moves of the adversary and the player) and then generating the process by coin flips  $\epsilon_t$ . The resulting quantity is a symmetrized one and can be seen as a sequential version of the classical Rademacher complexity. In particular, the symmetrized upper bound of Lemma 4 allows us to define appropriate covering numbers and thus analyze infinite classes of checking rules.

The definitions of a sequential *cover* and *covering number* below are from (Rakhlin et al., 2010a). Note that they differ from the corresponding classical “static” notions.

**Definition 5** Consider a binary-valued function class  $\mathcal{G} \subseteq \{0,1\}^{\mathcal{Y}}$  over some set  $\mathcal{Y}$ . For any given  $\mathcal{Y}$ -valued tree  $\mathbf{y}$  of depth  $T$ , a set  $V$  of binary-valued trees of depth  $T$  is called a 0-cover of  $\mathcal{G}$  on  $\mathbf{y}$  if

$$\forall g \in \mathcal{G}, \forall \epsilon \in \{\pm 1\}^T, \exists \mathbf{v} \in V \text{ s.t. } \forall t \in [T], g(\mathbf{y}_t(\epsilon)) = \mathbf{v}_t(\epsilon). \quad (4)$$

The *covering number* at scale 0 of a class  $\mathcal{G}$  (the 0-covering number) on a given tree  $\mathbf{y}$  is defined as

$$N(\mathcal{G}, \mathbf{y}) = \min \{ |V| : V \text{ is a 0-cover of } \mathcal{G} \text{ on } \mathbf{y} \}.$$

Also define the worst-case covering number for all depth- $T$  trees as  $N(\mathcal{G}, T) = \sup_{\mathbf{y}} N(\mathcal{G}, \mathbf{y})$ .

We point out that the order of quantifiers in (4) is crucial: For a given function  $g$ , the covering tree  $\mathbf{v}$  can be chosen based on the path  $\epsilon$  itself. It is thus not correct to think of the 0-cover as the number of distinct trees obtained by evaluating all functions from  $\mathcal{G}$  on the given  $\mathbf{y}$ . Indeed, as described in (Rakhlin et al., 2010a), it is possible for an exponentially-large set of functions  $\mathcal{G}$  to have a 0-cover of size 2, capturing the temporal structure of  $\mathcal{G}$ .

**Definition 6** Define the minimal checking covering number of  $\zeta$  over depth  $T$  trees as

$$\mathcal{N}_{\text{ch}}(\zeta, T) = \sup_{\mathbf{x}, \mathbf{p}^\delta} N(\zeta, (\mathbf{z}^{(\mathbf{x}, \mathbf{p}^\delta)}, \mathbf{p}^\delta))$$

and the minimal checking cover of  $\zeta$  on  $\mathbf{x}, \mathbf{p}^\delta$  as the set of size  $N(\zeta, (\mathbf{z}^{(\mathbf{x}, \mathbf{p}^\delta)}, \mathbf{p}^\delta))$  that provides the cover. Here, abusing notation,  $(\mathbf{z}^{(\mathbf{x}, \mathbf{p}^\delta)}, \mathbf{p}^\delta)$  is the  $\mathcal{Z}^\delta \times C_\delta$ -valued tree obtained by pairing the trees  $\mathbf{z}^{(\mathbf{x}, \mathbf{p}^\delta)}$  and  $\mathbf{p}^\delta$  together (and note that  $\zeta$  is a class of binary functions on  $\mathcal{Z}^\delta \times C_\delta$ ).

Importantly, the minimal checking covering number is defined only over history trees  $\mathbf{z}^{(\mathbf{x}, \mathbf{p}^\delta)}$  consistent with the chosen trees  $\mathbf{x}, \mathbf{p}^\delta$ . Clearly, we can upper bound the minimal checking covering number by the minimal cover  $N(\zeta, T)$  over  $\mathcal{Z}^\delta \times C_\delta$ . It is immediate that  $\mathcal{N}_{\text{ch}}(\zeta, T) \leq N(\zeta, T)$ .

**Theorem 7** For  $T > \frac{16c_k \log(4)}{\theta^2}$  and  $\delta \leq \theta/2$ ,

$$\mathcal{V}_T^\theta(\zeta) \leq 8 \mathcal{N}_{\text{ch}}(\zeta, T) \exp\left(-\frac{T\theta^2}{64 c_k}\right)$$

**Proof [Theorem 7]** Given any trees  $\mathbf{x}, \mathbf{p}^\delta$ , let the set of binary valued trees  $V$  be a (finite) minimal checking cover of  $\zeta$  on  $\mathbf{x}, \mathbf{p}^\delta$ . For any  $c \in \zeta$ , let  $\mathbf{v}[c, \epsilon] \in V$  be the member of the minimal checking cover that matches  $c$  on the tree  $(\mathbf{x}, \mathbf{p}^\delta)$  over the path  $\epsilon$ . Then we see that

$$\begin{aligned} \mathbb{P}_\epsilon \left( \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t c(\mathbf{z}_t^\delta(\epsilon), \mathbf{p}_t^\delta(\epsilon)) \mathbf{x}_t(\epsilon) \right\| > \theta/8 \right) &= \mathbb{P}_\epsilon \left( \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \mathbf{v}[c, \epsilon]_t(\epsilon) \mathbf{x}_t(\epsilon) \right\| > \theta/8 \right) \\ &\leq \mathbb{P}_\epsilon \left( \max_{\mathbf{v} \in V} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \mathbf{v}_t(\epsilon) \mathbf{x}_t(\epsilon) \right\| > \theta/8 \right) \end{aligned}$$

Since  $|V|$  is finite, by union bound we pass to the upper bound of

$$|V| \max_{\mathbf{v} \in V} \mathbb{P}_\epsilon \left( \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \mathbf{v}_t(\epsilon) \mathbf{x}_t(\epsilon) \right\| > \theta/8 \right) \leq \mathcal{N}_{\text{ch}}(\zeta, T) \max_{\mathbf{v} \in V} \mathbb{P}_\epsilon \left( \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \mathbf{v}_t(\epsilon) \mathbf{x}_t(\epsilon) \right\| > \theta/8 \right)$$

We now appeal to Lemma 3. Note that  $\mathbf{v}$  is binary-valued and  $\mathbf{x}$  is  $E_k$ -valued, and, hence,  $\|\mathbf{v}_t(\epsilon) \mathbf{x}_t(\epsilon)\| \leq 1$  for any  $t$ . Also,  $\epsilon_t \mathbf{v}_t(\epsilon) \mathbf{x}_t(\epsilon)$  is a martingale difference sequence since  $\mathbf{x}_t$  and  $\mathbf{v}_t$  by definition only depend on  $\epsilon_{1:t-1}$ . Hence, for any  $\mathbf{x}$  and  $\mathbf{v}$ ,

$$\mathbb{P}_\epsilon \left( \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \mathbf{v}_t(\epsilon) \mathbf{x}_t(\epsilon) \right\| > \theta/8 \right) \leq 2 \exp \left( -\frac{T\theta^2}{64 c_k} \right)$$

Combining with Lemma 4, we have that

$$\mathcal{V}_T^\theta(\zeta) \leq 4 \sup_{\mathbf{x}, \mathbf{p}^\delta} \mathbb{P}_\epsilon \left( \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t c(\mathbf{z}_t^\delta(\epsilon), \mathbf{p}_t^\delta(\epsilon)) \mathbf{x}_t(\epsilon) \right\| > \theta/8 \right) \leq 8 \mathcal{N}_{\text{ch}}(\zeta, T) \exp \left( -\frac{T\theta^2}{64 c_k} \right) .$$

■

## 5. Families of Checking Rules

The main objective of this paper is to find general sufficient conditions on the set of checking rules that guarantee existence of a calibrated strategy. Theorem 7 guarantees decay of  $\mathcal{V}_T^\theta(\zeta)$  if checking covering numbers of  $\zeta$  can be controlled. In this section, we show control of these numbers under various assumptions on  $\zeta$ , along with the resulting rates of convergence.

### 5.1. Finite Class of Checking Rules

The first straightforward consequence of Theorem 7 is that, for a finite class  $\zeta$ ,

$$\mathcal{V}_T^\theta(\zeta) \leq 8 |\zeta| \exp \left( -\frac{T\theta^2}{64 c_k} \right) \quad (5)$$

for  $T > \frac{16c_k \log(4)}{\theta^2}$ . We can convert this statement into a probability of  $\mathbf{R}_T$  being large. To this end, setting the right-hand side of (5) to  $\eta$  and solving for  $\theta$ , we obtain

$$\theta = \sqrt{\frac{64c_k \log(8|\zeta|/\eta)}{T}}.$$

For this value, the condition  $T > \frac{16c_k \log(4)}{\theta^2}$  is automatically satisfied. We can then state the result for finite  $\zeta$  as follows: There exists a randomized strategy for the player such that

$$\mathbb{P} \left( \mathbf{R}_T \leq \sqrt{\frac{64c_k \log(8|\zeta|/\eta)}{T}} \right) \geq 1 - \eta$$

for any  $\eta > 0$ , no matter how Nature chooses the outcomes.

As an example, consider the classic problem of digit identification, with the images of digits presented as “side information”. A system that generates a prediction and gets scored against the true digit is then being effectively tested by a total of 10 checking rules.

## 5.2. History Invariant Checking Rules

A finite class of checking rules is, in some sense, too easy for the forecaster. Once we go to infinite classes, much of the difficulty arises from potentially complicated dependence of the rules on the history. Before attacking infinite classes of history-dependent rules, we consider the case of history-independence. The classical notion of calibration is an example of such a class of checking rules.

Formally, assume that  $\zeta$  is a class of checking rules such that for all  $c \in \zeta$ , pair of histories  $z, z' \in \mathcal{Z}$  and  $p \in \Delta_k$  :

$$c(z, p) = c(z', p)$$

Abusing notation, we can write each  $c \in \zeta$  as a function  $c : \Delta_k \mapsto \{0, 1\}$ .

The next lemma recovers the rates obtained by Mannor and Stoltz (2010). For  $k = 2$ , the rate  $T^{-1/3}$  has been also found previously by a variety of algorithms that reduced calibration on an  $\epsilon$ -grid to the experts problem of no-internal regret with  $O(1/\epsilon)$  experts.

**Lemma 8** *For any class  $\zeta$  of history invariant measurable checking rules, for any  $\theta \in (0, 1]$  we have that*

$$\mathcal{V}_T^\theta(\zeta) \leq 8 \exp\left(-\frac{T\theta^2}{64 c_k} + \left(\frac{c_1}{\theta}\right)^{k-1}\right)$$

for  $T > \frac{16c_k \log(4)}{\theta^2}$ . This leads to

$$\mathbb{P}\left(\mathbf{R}_T \leq c'_k T^{-1/(k+1)} \sqrt{\log(8/\eta)}\right) \leq 1 - \eta$$

for an appropriate constant  $c'_k$ .

**Proof** From Eq. (1), the total number of different labelings of set  $C_\delta$  by  $\zeta$  is bounded by  $2^{(c_1/(2\delta))^{k-1}}$  (that is, the number of binary functions over set of size  $|C_\delta|$ ). For  $\delta = \theta/2$ , we have that the size is bounded by  $2^{(c_1/\theta)^{k-1}}$ . By Theorem 7 we conclude that

$$\mathcal{V}_T^\theta(\zeta) \leq 8 2^{(\frac{c_1}{\theta})^{k-1}} \exp\left(-\frac{T\theta^2}{64 c_k}\right).$$

Over-bounding, we obtain the first statement. Now, set  $\theta = c'_k T^{-1/(k+1)} \sqrt{\log(8/\eta)}$  for some appropriate constant  $c'_k$ . For this value of  $\theta$ , it holds that  $\mathcal{V}_T^\theta(\zeta) \leq \eta$ . We conclude that

$$\mathbb{P}\left(\mathbf{R}_T \leq c'_k T^{-1/(k+1)} \sqrt{\log(8/\eta)}\right) \leq 1 - \eta.$$

■

While the rate for all measurable history-invariant checking rules decays with  $k$ , we can get  $\tilde{O}(\sqrt{T})$  rates as soon as we restrict the class of checking rules to have a finite combinatorial dimension. A finite combinatorial dimension limits the effective size of  $\zeta$  as applied on  $C_\delta$ . The first result we present holds for Vapnik-Chervonenkis classes.

**Lemma 9** *For any class  $\zeta$  of history invariant checking rules with VC dimension  $\text{VCdim}(\zeta)$ , we have that*

$$\mathcal{V}_T^\theta(\zeta) \leq 8 \left( \frac{e c_1}{\theta} \right)^{(k-1) \text{VCdim}(\zeta)} \exp \left( -\frac{T\theta^2}{64 c_k} \right)$$

for  $T > \frac{16c_k \log(4)}{\theta^2}$ . We therefore obtain

$$\mathbb{P} \left( \mathbf{R}_T \leq c' \sqrt{\frac{k \text{VCdim}(\zeta) \cdot c_k \log(8/\eta) \log T}{T}} \right) \leq 1 - \eta$$

for an appropriate constant  $c'_k$ .

**Proof** By the Vapnik-Chervonenkis-Sauer-Shelah lemma, the number of different labelings of the set  $C_\delta$  by  $\zeta$  is bounded by  $(e |C_\delta|)^{\text{VCdim}(\zeta)}$ . Clearly, the size of the minimal 0-cover cannot be more than the number of different labelings on the set  $C_\delta$ . Using  $|C_\delta| \leq (c_1/(2\delta))^{k-1}$  with  $\delta = \theta/2$  and Theorem 7 we conclude that

$$\mathcal{V}_T^\theta(\zeta) \leq 8 \left( \frac{e c_1}{\theta} \right)^{(k-1) \text{VCdim}(\zeta)} \exp \left( -\frac{T\theta^2}{64 c_k} \right)$$

which concludes the first statement. For the probability version, set

$$\theta = c' \sqrt{\frac{k \text{VCdim}(\zeta) \cdot c_k \log(8/\eta) \log T}{T}}$$

For this setting,  $\mathcal{V}_T^\theta(\zeta) \leq \eta$  for some appropriate  $k$ -independent constant  $c'$ . The second statement follows.  $\blacksquare$

For the classical calibration problem, the VC dimension of the set of  $\ell_1$ -balls is at most  $k^2$  and the constant  $c_k = 8k$  for the  $\ell_1$  norm (as shown in Lemma 3). Combining, we obtain the following corollary, which, to the best of our knowledge, does not appear in the literature.

**Corollary 10** *For classical calibration with  $k$  actions and  $\ell_1$  norm, the rate of convergence is*

$$O \left( k^2 \sqrt{\frac{\log(T) \log(1/\eta)}{T}} \right)$$

Next, we consider an alternative combinatorial parameter, called Littlestone's dimension (Littlestone, 1988; Ben-David et al., 2009). This dimension captures the sequential "richness" of the function class.

**Definition 11** *An  $\mathcal{X}$ -valued tree  $\mathbf{x}$  of depth  $d$  is shattered by a function class  $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$  if for all  $\epsilon \in \{\pm 1\}^d$ , there exists  $f \in \mathcal{F}$  such that  $f(\mathbf{x}_t(\epsilon)) = \epsilon_t$  for all  $t \in [d]$ . The Littlestone dimension  $\text{Ldim}(\mathcal{F}, \mathcal{X})$  is the largest  $d$  such that  $\mathcal{F}$  shatters some  $\mathcal{X}$ -valued tree of depth  $d$ .*

We use  $\text{Ldim}(\mathcal{F})$  for  $\text{Ldim}(\mathcal{F}, \mathcal{X})$  if the domain  $\mathcal{X}$  is clear from context. As shown in (Rakhlin et al., 2010a), the Littlestone's dimension can be used to upper bound sequential covering numbers in a way similar to VC dimension upper bounding the classical covering numbers.

**Lemma 12** *For any class  $\zeta$  of history invariant checking rules with Littlestone's dimension  $\text{Ldim}(\zeta)$ ,*

$$\mathcal{V}_T^\theta(\zeta) \leq 8 (eT)^{\text{Ldim}(\zeta)} \exp\left(-\frac{T\theta^2}{64 c_k}\right)$$

**Proof** Note that for any history invariant family of checking rules  $\zeta$ , the definition of covering number here coincides with the definition of covering number in (Rakhlin et al., 2010a) for binary class of functions  $\zeta$  on space  $C_\delta$ . Therefore,

$$\mathcal{N}_{\text{ch}}(\zeta, T) \leq (eT)^{\text{Ldim}(\zeta, C_\delta)}.$$

The Littlestone's dimension on the set  $C_\delta$  can be upper bounded by the Littlestone's dimension  $\text{Ldim}(\zeta)$  over the whole simplex  $\Delta_k$ . Using Theorem 7 concludes the proof. ■

In the above lemma and in the rest of the paper, it will be assumed that  $T$  is large enough that  $T > \frac{16c_k \log(4)}{\theta^2}$  so that we can appeal to Theorem 7.

### 5.3. Time Dependent Checking Rules

We now turn to richer classes of checking rules. Of particular interest are classes of history-invariant rules that have mild dependence on time. Our results have a flavor of “shifting experts” results in individual sequence prediction. Suppose the checking rules can be written as a family of functions  $c : [T] \times \Delta_k \mapsto \{0, 1\}$  (i.e. the checking rule only depends on the length of the history and not the history itself). More specifically, given a family  $\zeta$  of time invariant checking rules, we consider the family of time dependent checking rules  $\zeta^n$  given by checking rules that are allowed to change at most  $n \leq T$  times over the  $T$  rounds (checking rule for each round is chosen from  $\zeta$ ). Formally,

$$\begin{aligned} \zeta^n = \{c^n | \exists 1 = i_0 \leq \dots \leq i_n \leq T \text{ and } c_1, \dots, c_n \in \zeta \quad \text{s.t.} \\ \forall s \geq 0, \forall i_s \leq t \leq t' < i_{s+1}, c^n(t, \cdot) = c^n(t', \cdot) = c_s\} \end{aligned}$$

and  $i_{n+1}$  is assumed to be  $T + 1$ .

**Lemma 13** *For any class  $\zeta$  of history invariant measurable checking rules, we have that*

$$\mathcal{V}_T^\theta(\zeta^n) \leq 8 \exp\left(-\frac{T\theta^2}{64 c_k} + n \left(\frac{c_1}{\theta}\right)^{k-1} + n \log T\right)$$

**Proof** For any  $t$ , the total number of different labelings of set  $C_\delta$  by  $\zeta$  is bounded by  $2^{(c_1/(2\delta))^{k-1}}$ . To account for all the possibilities, we need to consider all possible ways of choosing  $n$  shifts out of  $T$  rounds, and then to choose a constant function for each interval

out of the  $2^{(c_1/(2\delta))^{k-1}}$  possibilities. Choosing  $\delta = \theta/2$ , the effective size of  $\zeta$  on  $C_\delta$  is bounded by  $\binom{T}{n} \left(2^{(c_1/\theta)^{k-1}}\right)^n$ . Hence by Theorem 7 we conclude that

$$\mathcal{V}_T^\theta(\zeta) \leq 8 \binom{T}{n} 2^{n(\frac{c_1}{\theta})^{k-1}} \exp\left(-\frac{T\theta^2}{64 c_k}\right)$$

which concludes the proof. ■

The corresponding statement in probability is analogous to that in Lemma 8 if  $n$  is constant. If  $n$  grows with  $T$ , a non-trivial rate in probability can still be shown as long as  $n = o(T)$ . Hence, there exists a calibration strategy for arbitrary sets of history-independent measurable checking rules which change  $o(T)$  of times.

**Lemma 14** *For any class  $\zeta$  of history invariant checking rules with VC dimension  $\text{VCdim}(\zeta)$ ,*

$$\mathcal{V}_T^\theta(\zeta^n) \leq 8 \binom{T}{n} \left(\frac{e c_1}{\theta}\right)^{n(k-1)\text{VCdim}(\zeta)} \exp\left(-\frac{T\theta^2}{64 c_k}\right)$$

**Proof** For any  $t \in [T]$  the number of different labelings of the set  $C_\delta$  by  $\zeta$  is bounded by  $(e |C_\delta|)^{\text{VCdim}(\zeta)}$ . Hence the total possible number of different labelings of set  $C_\delta$  by  $\zeta$  in the  $T$  different rounds can be bounded by  $\binom{T}{n} (e |C_\delta|)^{n\text{VCdim}(\zeta)} \leq \binom{T}{n} \left(\frac{e c_1}{\theta}\right)^{n(k-1)\text{VCdim}(\zeta)}$ . By Theorem 7 we conclude that

$$\mathcal{V}_T^\theta(\zeta) \leq 8 \binom{T}{n} \left(\frac{e c_1}{\theta}\right)^{n(k-1)\text{VCdim}(\zeta)} \exp\left(-\frac{T\theta^2}{64 c_k}\right)$$

which concludes the proof. ■

Similarly to Lemma 9, we obtain  $\tilde{O}(\sqrt{T})$  rate of convergence for the class  $\zeta^n$  constructed from a VC class of history-independent checking rules.

#### 5.4. General Checking Rules

In this section we study checking rules that depend on history. We start with an assumption on the form of these rules: history is represented by some potentially smaller set. Such a smaller set can arise from a bound on the available memory, or from limited precision.

Formally, assume that for some set  $\mathcal{Y}$  there exists a mapping  $\phi : \mathcal{Z}^\delta \mapsto \mathcal{Y}$  and a class of binary functions  $\mathcal{G} \subseteq \{0, 1\}^{\mathcal{Y} \times \Delta_k}$  with the following property: For any  $c \in \zeta$  there exists  $g \in \mathcal{G}$  such that

$$c(z, p) = g(\phi(z), p) \quad \text{for any } z \in \mathcal{Z} \text{ and } p \in \Delta_k.$$

Clearly, if we set  $\mathcal{Y} = \mathcal{Z}^\delta$  and  $\phi$  the identity mapping,  $\mathcal{G}$  and  $\zeta$  coincide.

**Lemma 15** *For any set  $\mathcal{Y}$  and class of binary functions  $\mathcal{G}$  satisfying the above mentioned assumption with mapping  $\phi$ , we have that*

$$\mathcal{V}_T^\theta \leq 8 (eT)^{\text{Ldim}(\mathcal{G})} \exp\left(-\frac{T\theta^2}{64 c_k}\right)$$

**Proof** Note that

$$\mathcal{N}_{\text{ch}}(\zeta, T) = \sup_{\mathbf{x}, \mathbf{p}^\delta} N(\zeta, (\mathbf{z}^{(\mathbf{x}, \mathbf{p}^\delta)}, \mathbf{p}^\delta)) = \sup_{\mathbf{x}, \mathbf{p}^\delta} N(\mathcal{G}, (\phi(\mathbf{z}^{(\mathbf{x}, \mathbf{p}^\delta)}), \mathbf{p}^\delta)) \leq \sup_{\mathbf{y}, \mathbf{p}^\delta} N(\mathcal{G}, (\mathbf{y}, \mathbf{p}^\delta)) \leq (eT)^{\text{Ldim}(\mathcal{G})}.$$

Using this with Theorem 7 we conclude the proof.  $\blacksquare$

**Corollary 16** For any class of checking rules  $\zeta$ ,

$$\mathcal{V}_T^\theta \leq 8(eT)^{\text{Ldim}(\zeta, \mathcal{Z}^\delta \times C_\delta)} \exp\left(-\frac{T\theta^2}{64c_k}\right)$$

**Proof** Use previous lemma with  $\mathcal{G} = \zeta$ ,  $\mathcal{Y} = \mathcal{Z}^\delta$  and  $\phi$  the identity mapping.  $\blacksquare$

### 5.5. Checking Rules With Limited History Lookback

We now consider a family of checking rules that only depend on at most  $m$  of the most recent pairs of actions played by the two players. We call such a class of rules an *m-look back family*. Specifically, for  $0 \leq m \leq T - 1$ , define  $\mathcal{Y} = \bigcup_{t=0}^m (C_\delta \times E_k)^t \subset \mathcal{Z}^\delta$ ,  $\mathcal{G} = \zeta$  and  $\phi : \mathcal{Z}^\delta \mapsto \mathcal{Y}$  is given by:

$$\phi(z) = \begin{cases} z & \text{if } z \in \mathcal{Y} \\ (z_{t-m-1}, \dots, z_t) & \text{if } z \in (C_\delta \times E_k)^t \text{ for some } m < t \leq T \end{cases}$$

The first bound we can get here directly is the one implied by Lemma 15 for the  $\mathcal{G}$  and  $\mathcal{Y}$  mentioned above.

**Lemma 17** For any *m-look back family* of checking rules  $\zeta$ ,

$$\mathcal{V}_T^\theta \leq 8 \cdot 2^{m k^m} \left(\frac{c_1}{\theta}\right)^{km} \exp\left(-\frac{T\theta^2}{64c_k}\right)$$

**Proof** Note that

$$|\mathcal{Y}| = \sum_{t=0}^m |(C_\delta \times E_k)^t| \leq \sum_{t=0}^m (|C_\delta| \cdot k)^t \leq \sum_{t=0}^m \left(\left(\frac{c_1}{2\delta}\right)^{(k-1)} \cdot k\right)^t \leq m k^m \left(\frac{c_1}{2\delta}\right)^{(k-1)m}$$

So for  $\delta = \theta/2$  we have  $|\mathcal{Y}| \leq m k^m \left(\frac{c_1}{\theta}\right)^{(k-1)m}$ . This implies that the total number of different possible binary labelings of elements of the set  $\mathcal{Y} \times C_\delta$  (and hence  $\mathcal{N}_{\text{ch}}(\zeta, T)$ ) is bounded by

$$\mathcal{N}_{\text{ch}}(\zeta, T) \leq 2^{m k^m} \left(\frac{c_1}{\theta}\right)^{km}$$

Hence using Theorem 7 we conclude the theorem statement.  $\blacksquare$

Note that the above bound gives polynomial convergence for any  $m \leq \frac{\log T}{1+\epsilon}$  for any  $\epsilon > 0$ . That is, there exists a forecasting strategy that can calibrate against any family of measurable checking rules which have dependence on a logarithmic (in  $T$ ) number of past forecasts and outcomes.

**Lemma 18** *For any  $m$ -look back family of checking rules  $\zeta$ , if VC dimension of the class as applied on input space  $\mathcal{Y} \times C_\delta$  is given by  $\text{VCdim}(\zeta, \mathcal{Y} \times C_\delta)$  then,*

$$\mathcal{V}_T^\theta \leq 2 \left( e m k^m \left( \frac{c_1}{\theta} \right)^{km} \right)^{\text{VCdim}(\zeta, \mathcal{Y} \times C_\delta)} \exp \left( - \frac{T\theta^2}{64c_k} \right)$$

**Proof** By VC lemma the number of different labelings of the set  $\mathcal{Y} \times C_\delta$  by the class  $\zeta$  is bounded by  $(e|\mathcal{Y} \times C_\delta|)^{\text{VCdim}(\zeta, \mathcal{Y} \times C_\delta)}$ . However

$$|\mathcal{Y} \times C_\delta| \leq m k^m \left( \frac{c_1}{\theta} \right)^{km}$$

Hence

$$\mathcal{N}(\zeta, T) \leq \left( e m k^m \left( \frac{c_1}{\theta} \right)^{km} \right)^{\text{VCdim}(\zeta, \mathcal{Y} \times C_\delta)}$$

We conclude the proof by appealing to Theorem 7. ■

The above bound guarantees existence of a calibration strategy whenever  $m = o(T)$ . That is, as long as the checking rule with bounded VC only looks back up to  $o(T)$  steps in history, the forecaster has a successful strategy.

### 5.6. Checking Rules with Bounded Computation

Whenever the number of arithmetic operations required to compute each function in a class is bounded by some constant, the VC dimension of the class can be bounded from above Goldberg and Jerrum (1995). Specifically result in Goldberg and Jerrum (1995) states that for binary function class  $\zeta$  over domain  $\mathcal{X} \subset \mathbb{R}^n$  defined by algorithms of description length bounded by  $\ell$  and which run in time  $U$  using only the operations of conditional jumps and  $+, -, \times$  and  $/$  (in constant time), the VC dimension of the function class is bounded by  $O(\ell U)$ . Using this with Lemma 18 we make the following observation.

For  $m$ -look back family of checking rules  $\zeta$  defined by algorithms with description length bounded by  $\ell$  and runtime bounded by  $U$ , applying Lemma 18, the value of the game is bounded by

$$\mathcal{V}_T^\theta \leq 2 \left( e m k \left( \frac{c_1}{\theta} \right)^k \right)^{O(m\ell U)} \exp \left( - \frac{T\theta^2}{64c_k} \right)$$

Hence we can guarantee calibration against set of all checking rules defined by algorithms of description length bounded by  $\ell$  and whose run times are bounded by  $U$  as long as  $m\ell U = o(T)$ .

## 6. Lower Bounds

In this section we show that the  $\sqrt{T}$  rate for classical calibration cannot be improved. While the argument is not difficult, we could not find it in the literature.

**Lemma 19** *For two actions, the rate for the classical calibration game is lower bounded for any  $\theta > 0$  as*

$$\mathcal{V}_T^\theta \geq \mathbb{P} \left( \frac{1}{T} \sum_{t=1}^T x_t \geq 2\theta \right)$$

where  $x_1, \dots, x_T$  are independent Rademacher random variables.

**Proof** Note that for  $k = 2$ , the vector notation for the outcomes is no longer necessary. Indeed, the difference of any two vectors in the simplex is  $|(a, 1-a) - (b, 1-b)| = 2|a-b|$ , and thus the value of the game can be written as

$$\mathcal{V}_T^\theta(\zeta) := \inf_{q_1} \sup_{x_1} \mathbb{E}_{f_1 \sim q_1} \dots \inf_{q_T} \sup_{x_T} \mathbb{E}_{f_T \sim q_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left| \frac{1}{T} \sum_{t=1}^T c(z_t, f_t) \cdot (f_t - x_t) \right| > \theta \right\} \right]$$

where  $q_t$  is a distribution over  $[0, 1]$ ,  $f_t \in [0, 1]$ , and  $x_t \in \{0, 1\}$ . In fact, the mathematical exposition is easier if  $q_t$  is a distribution on  $[-1, 1]$ ,  $f_t \in [-1, 1]$ , and  $x_t \in \{-1, 1\}$ . The problem is not changed, as one can easily translate between the two formulations. We consider a particular  $\zeta$  consisting of two rules:  $c_1(z_t, f_t) = \mathbf{1}\{f_t \geq 0\}$  and  $c_2(z_t, f_t) = \mathbf{1}\{f_t < 0\}$ . Note that we can equivalently write these rules as being  $1/4$ -close to the centers  $1/4$  and  $3/4$ . Hence, this is genuinely a classical  $\epsilon$ -calibration problem with  $\epsilon = 1/4$ . We can then write the value of the game as

$$\begin{aligned} & \inf_{q_1} \sup_{x_1} \mathbb{E}_{f_1 \sim q_1} \dots \inf_{q_T} \sup_{x_T} \mathbb{E}_{f_T \sim q_T} \\ & \quad \mathbf{1} \left\{ \max \left\{ \left| \frac{1}{T} \sum_{t=1}^T (x_t - f_t) \mathbf{1}\{f_t \geq 0\} \right|, \left| \frac{1}{T} \sum_{t=1}^T (x_t - f_t) \mathbf{1}\{f_t < 0\} \right| \right\} > \theta \right\} \end{aligned}$$

Let  $\text{sign}(b)$  denote the sign of  $b \in \mathbb{R}$ , and  $\text{sign}(0) = 1$ . Let us write

$$A(f_{1:T}, x_{1:T}) := \frac{1}{T} \sum_{t=1}^T (x_t - f_t) \mathbf{1}\{f_t \geq 0\} \quad \text{and} \quad B(f_{1:T}, x_{1:T}) := \frac{1}{T} \sum_{t=1}^T (x_t - f_t) \mathbf{1}\{f_t < 0\}.$$

The suprema over  $x_t$ 's can equivalently be written as suprema over all distributions on  $\{-1, 1\}$ . The lower bound is then achieved by choosing  $x_t$  to be i.i.d. Rademacher random variables. The lower bound on the value of the game can thus be written as

$$\begin{aligned} \mathcal{V}_T^\theta & \geq \inf_{q_1} \mathbb{E}_{f_1 \sim q_1} \mathbb{E}_{x_1} \dots \inf_{q_T} \mathbb{E}_{f_T \sim q_T} \mathbb{E}_{x_T} [\mathbf{1} \{ \max \{|A(f_{1:T}, x_{1:T})|, |B(f_{1:T}, x_{1:T})|\} > \theta \}] \\ & = \inf_{f_1} \mathbb{E}_{x_1} \dots \inf_{f_T} \mathbb{E}_{x_T} [\mathbf{1} \{ \max \{|A(f_{1:T}, x_{1:T})|, |B(f_{1:T}, x_{1:T})|\} > \theta \}] \\ & = \inf_{f_1} \sup_{a_1 \in \{\pm 1\}} \mathbb{E}_{x_1} \dots \inf_{f_T} \sup_{a_T \in \{\pm 1\}} \mathbb{E}_{x_T} [\mathbf{1} \{ \max \{|A(f_{1:T}, \{a_t x_t\}_{t=1}^T)|, |B(f_{1:T}, \{a_t x_t\}_{t=1}^T)|\} > \theta \}] \end{aligned}$$

The last equality holds because  $x_t$  have the same distribution as  $a_t x_t$ . Now, choosing  $a_t = \text{sign}(f_t)$ , we get

$$\begin{aligned} \mathcal{V}_T^\theta & \geq \inf_{f_1} \mathbb{E}_{x_1} \dots \inf_{f_T} \mathbb{E}_{x_T} [\mathbf{1} \{ \max \{|A(f_{1:T}, \{\text{sign}(f_t) x_t\}_{t=1}^T)|, |B(f_{1:T}, \{\text{sign}(f_t) x_t\}_{t=1}^T)|\} > \theta \}] \\ & = \inf_{f_1} \mathbb{E}_{x_1} \dots \inf_{f_T} \mathbb{E}_{x_T} [\mathbf{1} \{ \max \{|A(f_{1:T}, x_{1:T})|, |B(f_{1:T}, -x_{1:T})|\} > \theta \}]. \end{aligned}$$

Observe that

$$\begin{aligned}
 A(f_{1:T}, x_{1:T}) - B(f_{1:T}, -x_{1:T}) &= \frac{1}{T} \sum_{t=1}^T (x_t - f_t) \mathbf{1}\{f_t \geq 0\} - \frac{1}{T} \sum_{t=1}^T (-x_t - f_t) \mathbf{1}\{f_t < 0\} \\
 &= \frac{1}{T} \sum_{t=1}^T x_t - \frac{1}{T} \sum_{t=1}^T f_t \mathbf{1}\{f_t \geq 0\} + \frac{1}{T} \sum_{t=1}^T f_t \mathbf{1}\{f_t < 0\} \\
 &\leq \frac{1}{T} \sum_{t=1}^T x_t.
 \end{aligned}$$

Hence,

$$\mathbf{1}\{\max\{|A(f_{1:T}, x_{1:T})|, |B(f_{1:T}, -x_{1:T})|\} > \theta\} > \mathbf{1}\left\{\frac{1}{T} \sum_{t=1}^T x_t < -2\theta\right\}.$$

We conclude

$$\mathcal{V}_T^\theta \geq \mathbb{P}\left(\frac{1}{T} \sum_{t=1}^T x_t < -2\theta\right).$$

■

The lower bound of Lemma 19 can be immediately extended to  $k > 2$  actions and history-invariant checking rules that change  $O(k)$  times. This can be done by dividing  $T$  rounds into  $\lfloor k/2 \rfloor$  equal-length periods and then constructing the lower bound for each period based on two actions.

## 7. Proofs

**Proof [Lemma 2]** The first step is replacing the suprema over  $x_t$  with suprema over distributions  $p_t$  on  $E_k$ . The second step is exchanging each infimum and supremum by appealing to the minimax theorem.

$$\begin{aligned}
 \mathcal{V}_T^\theta(\zeta) &= \inf_{q_1} \sup_{p_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \inf_{q_T} \sup_{p_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t, f_t) \cdot (f_t - x_t) \right\| > \theta \right\} \right] \\
 &= \sup_{p_1} \inf_{q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T} \inf_{q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t, f_t) \cdot (f_t - x_t) \right\| > \theta \right\} \right] \\
 &= \sup_{p_1} \inf_{f_1 \in \Delta_k} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \inf_{f_T \in \Delta_k} \mathbb{E}_{x_T \sim p_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t, f_t) \cdot (f_t - x_t) \right\| > \theta \right\} \right]
 \end{aligned}$$

Now since  $C_\delta \subset \Delta_k$  we have

$$\begin{aligned} \mathcal{V}_T^\theta(\zeta) &= \sup_{p_1} \inf_{f_1 \in \Delta_k} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \inf_{f_T \in \Delta_k} \mathbb{E}_{x_T \sim p_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t, f_t) \cdot (f_t - x_t) \right\| > \theta \right\} \right] \\ &\leq \sup_{p_1} \inf_{f_1 \in C_\delta} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \inf_{f_T \in C_\delta} \mathbb{E}_{x_T \sim p_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t, f_t) \cdot (f_t - x_t) \right\| > \theta \right\} \right] \\ &\leq \sup_{p_1} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T \sim p_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t^\delta - x_t) \right\| > \theta \right\} \right] \end{aligned} \quad (6)$$

where the last inequality is obtained by replacing each  $\inf_{f_t \in C_\delta}$  by the (possibly) sub-optimal choice of  $p_t^\delta$ , thus only increasing the value.

By triangle inequality

$$\left\| \frac{1}{T} \sum_{t=1}^T c(z_t, p_t^\delta) \cdot (p_t^\delta - x_t) \right\| \leq \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t^\delta - p_t) \right\| + \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x_t) \right\|$$

and the first term above is further bounded above by

$$\frac{1}{T} \sum_{t=1}^T \left\| c(z_t^\delta, p_t^\delta) \cdot (p_t^\delta - p_t) \right\| \leq \frac{1}{T} \sum_{t=1}^T \|p_t^\delta - p_t\| \leq \delta .$$

Using this in Equation 6, we get

$$\begin{aligned} \mathcal{V}_T^\theta(\zeta) &\leq \sup_{p_1} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T \sim p_T} \left[ \mathbf{1} \left\{ \delta + \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x_t) \right\| > \theta \right\} \right] \\ &\leq \mathbf{1} \{\delta > \theta/2\} + \sup_{p_1} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T \sim p_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x_t) \right\| > \theta/2 \right\} \right] \end{aligned}$$

Choosing  $\delta \leq \theta/2$  concludes the proof. ■

### Proof [Lemma 3]

The result is a straightforward consequence of concentration results for 2-smooth functions of an average of a martingale difference sequence due to Pinelis (1994). We also refer to (Rakhlin et al., 2010b) for a short but detailed proof. The result states that, for a 2-smooth norm (in particular,  $\|\cdot\|_2$ ),

$$\mathbb{P} \left( \left\| \frac{1}{T} \sum_{t=1}^T d_t \right\|_2 \geq \epsilon \right) \leq 2 \exp \left( -\frac{\epsilon^2 T}{8B^2} \right)$$

if  $\|d_t\|_2 \leq B$  almost surely for all  $t$ . It remains to pass from our norm  $\|\cdot\|$  to the  $\ell_2$  norm. Here, we make this transition explicit for any  $\ell_p$  norm ( $1 \leq p \leq \infty$ ), but it can also be done for any appropriately normalized norm on  $\mathbb{R}^k$ .

For  $p \leq 2$ ,  $\|\cdot\|_2 \leq \|\cdot\|_p$  and thus the condition  $\|d_t\|_p \leq 1$  implies  $\|d_t\|_2 \leq 1$ . Further,  $\|\cdot\|_p \leq \sqrt{k}\|\cdot\|_2$  and so  $\|\cdot\|_p \geq \epsilon$  implies  $\|\cdot\|_2 \geq \epsilon/\sqrt{k}$ . Thus,  $c_k = 8k$ . Now, for the case  $p \geq 2$ ,  $\|\cdot\|_2 \leq \sqrt{k}\|\cdot\|_p$  and thus we set  $B = \sqrt{k}$ , leading to the value  $c_k = 8k$ . ■

## Acknowledgments

A. Rakhlin gratefully acknowledges the support of NSF under grant CAREER DMS-0954737 and Dean's Research Fund.

## References

- J. Abernethy, P. L. Bartlett, and E. Hazan. Blackwell approachability and no-regret learning are equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.
- S. Ben-David, D. Pal, and S. Shalev-Shwartz. Agnostic online learning. In *Proceedings of the 22th Annual Conference on Learning Theory*, 2009.
- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- A. P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.
- V. de la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Springer, 1998.
- D. P. Foster and R. V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40–55, October 1997.
- D. P. Foster and R. V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379, 1998.
- P.W. Goldberg and M.R. Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2):131–148, 1995. ISSN 0885-6125.
- S. Kakade, M. Kearns, J. Langford, and L. Ortiz. Correlated equilibria in graphical games. In *Proceedings of the 4th ACM Conference on Electronic Commerce*, pages 42–47. ACM, 2003. ISBN 158113679X.
- S. M. Kakade and D. P. Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115 – 130, 2008. ISSN 0022-0000. Learning Theory 2004.
- E. Kalai, E. Lehrer, and R. Smorodinsky. Calibrated Forecasting and Merging. *Games and Economic Behavior*, 29(1-2):151–169, 1999. ISSN 0899-8256.
- E. Lehrer. The game of normal numbers. *Mathematics of Operations Research*, 29(2):259–265, 2004. ISSN 0364-765X.

CALIBRATION WITH CHECKING RULES

- N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 04 1988.
- S. Mannor and G. Stoltz. A geometric proof of calibration. *Mathematics of Operations Research*, 35(4):721–727, 2010. doi: 10.1287/moor.1100.0465.
- D. Oakes. Self-calibrating priors do not exist. *Journal of the American Statistical Association*, 80(390):339–339, 1985. ISSN 0162-1459.
- W. Olszewski and A. Sandroni. A nonmanipulable test. *Annals of Statistics*, 37(2):1013–1039, 2009.
- I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.
- A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In *NIPS*, 2010a. Full version available at arXiv:1006.1138.
- A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Beyond regret. *ArXiv preprint arXiv:1011.3168v2*, 2010b.
- A. Sandroni, R. Smorodinsky, and R.V. Vohra. Calibration with many checking rules. *Mathematics of Operations Research*, 28(1):141–153, 2003. ISSN 0364-765X.
- H.P. Young. *Strategic learning and its limits*. Oxford University Press, USA, 2004. ISBN 0199269181.

## Appendix

**Proof [Lemma 4]** Fix a  $\mathbf{p}$ . If we condition on  $x_1, \dots, x_T$ , the sequence of  $p_1, \dots, p_T$  is well-defined, and we can consider a *tangent* sequence  $x'_t \sim p_t$ . This sequence is independent (see (de la Peña and Giné, 1998; Rakhlin et al., 2010a)). Note also that for any  $t$ ,  $c(z_t^\delta, p_t^\delta)$  is constant given  $x_1, \dots, x_T$ . Then for any fixed  $c \in \zeta$ ,

$$\begin{aligned} & \mathbb{E}_{x'_1 \sim p_1, \dots, x'_T \sim p_T} \left[ \mathbf{1} \left\{ \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x'_t) \right\| > \theta/4 \right\} \middle| x_1, \dots, x_T \right] \\ &= \mathbb{P} \left( \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x'_t) \right\| > \theta/4 \middle| x_1, \dots, x_T \right) \leq 2 \exp \left( -\frac{T\theta^2}{16c_k} \right) \leq \frac{1}{2} \end{aligned}$$

where the last inequality is by our assumption that  $T > \frac{16c_k \log(4)}{\theta^2}$ . Hence we can conclude that for any fixed  $c \in \zeta$ ,

$$\mathbb{P} \left( \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x'_t) \right\| \leq \theta/4 \middle| x_1, \dots, x_T \right) \geq \frac{1}{2}$$

Now since we are conditioning on  $x_1, \dots, x_T$  we can pick  $c^* \in \zeta$  as :

$$c^* = \operatorname{argmax}_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x_t) \right\|$$

and so

$$\mathbb{P} \left( \left\| \frac{1}{T} \sum_{t=1}^T c^*(z_t^\delta, p_t^\delta) \cdot (p_t - x'_t) \right\| \leq \theta/4 \middle| x_1, \dots, x_T \right) \geq \frac{1}{2} \quad (7)$$

Since the Inequality (7) holds for any  $x_1, \dots, x_T$  we assert that

$$\frac{1}{2} \leq \mathbb{P} \left( \left\| \frac{1}{T} \sum_{t=1}^T c^*(z_t^\delta, p_t^\delta) \cdot (p_t - x'_t) \right\| \leq \theta/4 \middle| \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x_t) \right\| > \theta/2 \right)$$

Hence we can conclude that for any distribution,

$$\begin{aligned} & \frac{1}{2} \mathbb{P} \left( \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x_t) \right\| > \theta/2 \right) \\ & \leq \mathbb{P} \left( \left\| \frac{1}{T} \sum_{t=1}^T c^*(z_t^\delta, p_t^\delta) \cdot (p_t - x'_t) \right\| \leq \theta/4 \middle| \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x_t) \right\| > \theta/2 \right) \\ & \quad \times \mathbb{P} \left( \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x_t) \right\| > \theta/2 \right) \\ &= \mathbb{P} \left( \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x_t) \right\| > \theta/2, \left\| \frac{1}{T} \sum_{t=1}^T c^*(z_t^\delta, p_t^\delta) \cdot (p_t - x'_t) \right\| \leq \theta/4 \right) \\ &\leq \mathbb{P} \left( \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (x_t - x'_t) \right\| > \theta/4 \right) \end{aligned}$$

Note that the probability is both with respect to the stochastic process  $x_1, \dots, x_T$  and the tangent sequence  $x'_1, \dots, x'_T$ . Furthermore, the above inequality holds for any  $\mathbf{p}$ . Thus,

$$\begin{aligned} & \frac{1}{2} \sup_{\mathbf{p}} \mathbb{E}_{x_1, \dots, x_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (p_t - x_t) \right\| > \theta/2 \right\} \right] \\ & \leq \sup_{\mathbf{p}} \mathbb{E}_{x_1, \dots, x_T} \mathbb{E}_{x'_1, \dots, x'_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (x_t - x'_t) \right\| > \theta/4 \right\} \right] \end{aligned}$$

Moving back to the expanded notation of (2) and using Lemma 2,

$$\begin{aligned} \frac{1}{2} \mathcal{V}_T^\theta & \leq \sup_{p_1} \mathbb{E}_{x_1 \sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T \sim p_T} \mathbb{E}_{x'_1 \sim p_1, \dots, x'_T \sim p_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (x_t - x'_t) \right\| > \theta/4 \right\} \right] \\ & \leq \sup_{p_1} \mathbb{E}_{x_1, x'_1 \sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T, x'_T \sim p_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (x_t - x'_t) \right\| > \theta/4 \right\} \right] \end{aligned}$$

Next, we upper bound the above expression by introducing suprema over  $p_t^\delta$  (we are slightly abusing the notation, as these variables will no longer depend on  $p_t$ ):

$$\begin{aligned} & \sup_{p_1} \sup_{p_1^\delta \in C_\delta} \mathbb{E}_{x_1, x'_1 \sim p_1} \dots \sup_{p_T} \sup_{p_T^\delta \in C_\delta} \mathbb{E}_{x_T, x'_T \sim p_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T c(z_t^\delta, p_t^\delta) \cdot (x_t - x'_t) \right\| > \theta/4 \right\} \right] \\ & = \sup_{p_1} \sup_{p_1^\delta \in C_\delta} \mathbb{E}_{x_1, x'_1 \sim p_1} \dots \sup_{p_T} \sup_{p_T^\delta \in C_\delta} \mathbb{E}_{x_T, x'_T \sim p_T} \mathbb{E}_{\epsilon_T} \\ & \quad \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^{T-1} c(z_t^\delta, p_t^\delta) \cdot (x_t - x'_t) + \epsilon_T c(z_T^\delta, p_T^\delta) \cdot (x_T - x'_T) \right\| > \theta/4 \right\} \right] \end{aligned}$$

The last step is justified because  $x_T$  and  $x'_T$  have the same distribution  $p_t$  when conditioned on  $x_1, \dots, x_{T-1}$ , and thus we can introduce the Rademacher random variable  $\epsilon_T$ . Next, we pass to the supremum over  $(x_T, x'_T)$ :

$$\begin{aligned} & \sup_{p_1} \sup_{p_1^\delta \in C_\delta} \mathbb{E}_{x_1, x'_1 \sim p_1} \dots \sup_{x_T, x'_T \in E_k} \sup_{p_T^\delta \in C_\delta} \mathbb{E}_{\epsilon_T} \\ & \quad \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^{T-1} c(z_t^\delta, p_t^\delta) \cdot (x_t - x'_t) + \epsilon_T c(z_T^\delta, p_T^\delta) \cdot (x_T - x'_T) \right\| > \theta/4 \right\} \right] \\ & = \sup_{p_1} \sup_{p_1^\delta \in C_\delta} \mathbb{E}_{x_1, x'_1 \sim p_1} \dots \sup_{p_{T-1}} \sup_{p_{T-1}^\delta \in C_\delta} \mathbb{E}_{x_{T-1}, x'_{T-1} \sim p_{T-1}} \mathbb{E}_{\epsilon_{T-1}} \sup_{x_T, x'_T \in E_k} \sup_{p_T^\delta \in C_\delta} \mathbb{E}_{\epsilon_T} \\ & \quad \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^{T-2} c(z_t^\delta, p_t^\delta) \cdot (x_t - x'_t) + \sum_{j=T-1}^T \epsilon_j c(z_j^\delta, p_j^\delta) \cdot (x_j - x'_j) \right\| > \theta/4 \right\} \right] \\ & \leq \sup_{p_1} \sup_{p_1^\delta \in C_\delta} \mathbb{E}_{x_1, x'_1 \sim p_1} \dots \sup_{x_{T-1}, x'_{T-1} \in E_k} \sup_{p_{T-1}^\delta \in C_\delta} \mathbb{E}_{\epsilon_{T-1}} \sup_{x_T, x'_T \in E_k} \sup_{p_T^\delta \in C_\delta} \mathbb{E}_{\epsilon_T} \\ & \quad \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^{T-2} c(z_t^\delta, p_t^\delta) \cdot (x_t - x'_t) + \sum_{j=T-1}^T \epsilon_j c(z_j^\delta, p_j^\delta) \cdot (x_j - x'_j) \right\| > \theta/4 \right\} \right] \end{aligned}$$

Continuing similarly all the way to the first term, we obtain an upper bound

$$\sup_{x_1, x'_1 \in E_k} \sup_{p_1^\delta \in C_\delta} \mathbb{E}_{\epsilon_1} \dots \sup_{x_T, x'_T \in E_k} \sup_{p_T^\delta \in C_\delta} \mathbb{E}_{\epsilon_T} \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t c(z_t^\delta, p_t^\delta) \cdot (x_t - x'_t) \right\| > \theta/4 \right\} \right]$$

We now pass to the tree notation. The above quantity is equal to

$$\sup_{\mathbf{x}, \mathbf{x}', \mathbf{p}^\delta} \mathbb{E}_\epsilon \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t c(\mathbf{z}_t^\delta(\epsilon), \mathbf{p}_t^\delta(\epsilon)) \cdot (\mathbf{x}_t(\epsilon) - \mathbf{x}'_t(\epsilon)) \right\| > \theta/4 \right\} \right]$$

where  $\mathbf{x}, \mathbf{x}'$  are  $E_k$ -valued trees of depth  $T$ ,  $\mathbf{p}^\delta$  is a  $C_\delta$ -valued tree of depth  $T$ , and the  $\mathcal{Z}$ -valued history tree is defined for by

$$\mathbf{z}_t^\delta(\epsilon) := \left( (\mathbf{p}_1^\delta(\epsilon), \mathbf{x}_1(\epsilon)), \dots, (\mathbf{p}_{t-1}^\delta(\epsilon), \mathbf{x}_{t-1}(\epsilon)) \right).$$

Here,  $\epsilon = (\epsilon_1, \dots, \epsilon_T) \in \{\pm 1\}^T$  denotes a path. The last quantity is upper bounded by

$$\begin{aligned} & \sup_{\mathbf{x}, \mathbf{x}', \mathbf{p}^\delta} \mathbb{E}_\epsilon \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t c(\mathbf{z}_t^\delta(\epsilon), \mathbf{p}_t^\delta(\epsilon)) \mathbf{x}_t(\epsilon) \right\| + \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t c(\mathbf{z}_t^\delta(\epsilon), \mathbf{p}_t^\delta(\epsilon)) \mathbf{x}'_t(\epsilon) \right\| > \theta/4 \right\} \right] \\ & \leq \sup_{\mathbf{x}, \mathbf{x}', \mathbf{p}^\delta} \mathbb{E}_\epsilon \left\{ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t c(\mathbf{z}_t^\delta(\epsilon), \mathbf{p}_t^\delta(\epsilon)) \mathbf{x}_t(\epsilon) \right\| > \theta/8 \right\} \right. \\ & \quad \left. + \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t c(\mathbf{z}_t^\delta(\epsilon), \mathbf{p}_t^\delta(\epsilon)) \mathbf{x}'_t(\epsilon) \right\| > \theta/8 \right\} \right\} \\ & \leq 2 \sup_{\mathbf{x}, \mathbf{p}^\delta} \mathbb{E}_\epsilon \left[ \mathbf{1} \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t c(\mathbf{z}_t^\delta(\epsilon), \mathbf{p}_t^\delta(\epsilon)) \mathbf{x}_t(\epsilon) \right\| > \theta/8 \right\} \right] \\ & = 2 \sup_{\mathbf{x}, \mathbf{p}^\delta} \mathbb{P}_\epsilon \left( \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t c(\mathbf{z}_t^\delta(\epsilon), \mathbf{p}_t^\delta(\epsilon)) \mathbf{x}_t(\epsilon) \right\| > \theta/8 \right) \end{aligned}$$

■

# Concentration-Based Guarantees for Low-Rank Matrix Reconstruction

**Rina Foygel**

*Department of Statistics, University of Chicago*

RINA@UCHICAGO.EDU

**Nathan Srebro**

*Toyota Technological Institute at Chicago*

NATI@TTIC.EDU

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We consider the problem of approximately reconstructing a partially-observed, approximately low-rank matrix. This problem has received much attention lately, mostly using the trace-norm as a surrogate to the rank. Here we study low-rank matrix reconstruction using both the trace-norm, as well as the less-studied max-norm, and present reconstruction guarantees based on existing analysis on the Rademacher complexity of the unit balls of these norms. We show how these are superior in several ways to recently published guarantees based on specialized analysis.

**Keywords:** Matrix completion, low-rank matrices, trace norm, nuclear norm, max norm, Rademacher complexity

## 1. Introduction

We consider the problem of (approximately) reconstructing an (approximately) low-rank matrix based on observing a random subset of entries. That is, we observe  $s$  randomly chosen entries of an unknown matrix  $Y \in \mathbb{R}^{n \times m}$ , where we assume either  $Y$  is of rank at most  $r$ , or there exists  $X \in \mathbb{R}^{n \times m}$  of rank at most  $r$  that is close to  $Y$ . Based on these  $s$  observations, we would like to construct a matrix  $\hat{X}$  that is as close as possible to  $Y$ .

There has been much interest recently in computationally efficient methods for reconstructing a partially-observed, possibly noisy, low-rank matrix, and on accompanying guarantees on the quality of the reconstruction and the required number of observations. Since directly searching for a low-rank matrix minimizing the empirical reconstruction error is NP-hard (Chistov and Grigoriev, 1984), most work has focused on using the trace-norm (a.k.a. nuclear norm, or Schatten-1-norm) as a surrogate for the rank. The trace-norm of a matrix is the sum (i.e.  $\ell_1$ -norm) of its singular values, and thus relaxing the rank (i.e. the number of non-zero singular values) to the trace-norm is akin to relaxing the sparsity of a vector to its  $\ell_1$ -norm, as is frequently done in compressed sensing. The analysis of the quality of reconstruction has also been largely driven by ideas coming from compressed sensing, typically studying the optimality conditions of the empirical optimization problem, and often requiring various “incoherence”-type assumptions on the underlying low-rank matrix.

In this paper we provide simple guarantees on approximate low-rank matrix reconstruction using a different surrogate regularizer: the  $\gamma_{2:\ell_1 \rightarrow \ell_\infty}$  norm, which we refer to simply as the “max-norm”. This regularizer was first suggested by Srebro et al. (2005), though it has

not received much attention since. Here we show how this regularizer can yield guarantees that are superior in some ways to recent state-of-the-art. In particular, we show that when the entries are uniformly bounded, i.e.  $|X|_\infty = \mathbf{O}(1)$  (this corresponds to the “no spikiness” assumption of Negahban and Wainwright (2010), and is also assumed by Koltchinskii et al. (2010) and in the approximate reconstruction guarantee of Keshavan et al. (2010)), then the max-norm regularized predictor requires a sample size of

$$s = \mathbf{O} \left( \frac{r(n+m)}{\epsilon} \cdot \frac{\sigma^2 + \epsilon}{\epsilon} \cdot \log^3(1/\epsilon) \right) \quad (1)$$

to achieve mean-squared reconstruction error  $\frac{1}{nm}|\hat{X} - Y|_2^2 = \sigma^2 + \epsilon$ , where  $\sigma^2$  is the mean-squared-error of the best rank- $r$  approximation of  $Y$ —that is,  $\sigma^2 = \frac{1}{nm}|X - Y|_2^2$ , where  $X$  is the rank- $r$  approximation. When  $Y$  is exactly low-rank (the noiseless case),  $\sigma^2 = 0$  and the sample complexity is  $\mathbf{O} \left( \frac{r(n+m)}{\epsilon} \cdot \log^3(1/\epsilon) \right)$ . Compared to the three recent similar bounds mentioned above, this guarantee avoids the extra logarithmic dependence on the dimensionality, as well as the assumption of independent noise, but has a slightly worse dependence on  $\epsilon$ . We emphasize that we do not make any assumptions about the noise, nor about incoherence properties of the underlying low-rank matrix  $X$ .

We also provide a guarantee on the mean-absolute-error of the reconstruction, and discuss guarantees for reconstruction using the trace-norm as a surrogate. Using the trace-norm allows us to provide mean-absolute-error guarantees also for matrices where the magnitudes are *not* uniformly bounded (i.e. “spiky” matrices). We further show that a spikiness assumption is necessary for squared-error approximate reconstruction of low-rank matrices, regardless of the estimator used.

Instead of focusing on optimality conditions as in previous work, our guarantees follow from generic generalization guarantees based on the Rademacher complexity, and an analysis of the Rademacher complexity of the max-norm and trace-norm balls conducted by Srebro and Shraibman (2005). To obtain the desired low rank reconstruction guarantees, we combine these with bounds on the max-norm and trace-norm in terms of the rank. The point we make here is that these fairly simple arguments, mostly based on the work of Srebro and Shraibman (2005), are enough to obtain guarantees that are in many ways better and more general than those presented in recent years.

**Notation.** We use  $|M|$  to denote the elementwise norms of a matrix  $M$ :  $|M|_1 = \sum_{ij} |M_{ij}|$ ,  $|M|_2$  is the Frobenius norm, and  $|M|_\infty = \max_{ij} |M_{ij}|$ . We discuss  $n \times m$  matrices, and without loss of generality always assume  $n \geq m$ .

## 2. The Max-Norm and Trace-Norm

We will consider the following two matrix norms, which are both surrogates for the rank:

**Definition 1** *The **trace-norm** of a matrix  $X \in \mathbb{R}^{n \times m}$  is given by:*

$$\|X\|_\Sigma = \sum (\text{singular values of } X) = \min_{U,V: X=UV^T} |U|_2 |V|_2 .$$

**Definition 2** The **max-norm** of a matrix  $X \in \mathbb{R}^{n \times m}$  is given by:

$$\|X\|_{\max} = \min_{U,V:X=UV^T} \left( \max_i |U_{(i)}|_2 \right) \left( \max_j |V_{(j)}|_2 \right),$$

where  $U_{(i)}$  and  $V_{(j)}$  denote the  $i^{\text{th}}$  row of  $U$  and the  $j^{\text{th}}$  row of  $V$ , respectively.

Both the trace-norm and the max-norm are semi-definite representable (Fazel et al., 2002; Srebro et al., 2005). Consequently, optimization problems involving a constraint on the trace-norm or max-norm, a linear or quadratic objective, and possibly additional linear constraints, are solvable using semi-definite programming. We will consider estimators which are solutions to such problems.

Srebro and Shraibman (2005) and later Sherstov (2007) studied the max-norm and trace-norm as surrogates for the rank in a classification setting, where one is only concerned with the signs of the underlying matrix. They showed that a sign matrix might be realizable with low rank, but realizing it with unit margin might require exponentially high max-norm or trace-norm. Based on this analysis, they argued that the max-norm and trace-norm *cannot* be used to obtain reconstruction guarantees for sign matrices of low rank matrices.

Here, we show that in a regression setting, the situation is quite different, and the max-norm and trace-norm *are* good convex surrogates for the rank. The specific relationship between these surrogates and the rank is determined by how we control the scale of the matrix  $X$  (i.e. the magnitude of its entries). This will be made explicit in the next section, but for now we state the bounds on the trace-norm and max-norm in terms of the rank which we will leverage in Section 3.

By bounding the  $\ell_1$  norm of the singular values (i.e. the trace-norm) by their  $\ell_2$  norm (i.e. the Frobenius norm) and the number of non-zero values (the rank), we obtain the following relationship between the trace-norm and Frobenius norm:

$$|X|_2 \leq \|X\|_{\Sigma} \leq \sqrt{\text{rank}(X)} \cdot |X|_2. \quad (2)$$

Interpreting the Frobenius norm as specifying the *average* entry magnitude,  $\frac{1}{nm} |X|_2^2$ , we can view the above as upper bounding the trace-norm with the square root of the rank, when the average entry magnitude is fixed.

An analogous bound for the max norm, substituting  $\ell_{\infty}$  norm (maximal entry magnitude) for Frobenius norm (average entry magnitude), can be obtained as follows:

**Lemma 3** For any  $X \in \mathbb{R}^{n \times m}$ ,  $|X|_{\infty} \leq \|X\|_{\max} \leq \sqrt{\text{rank}(X)} \cdot |X|_{\infty}$ .

**Proof** Consider the minimizing factorization  $X = UV^T$  and let  $X_{ij}$  be the largest magnitude entry in  $X$ , then:  $\|X\|_{\max} \geq |U_{(i)}| \cdot |V_{(j)}| \geq |X_{ij}| = |X|_{\infty}$ .

To obtain the upper bound we first write the max-norm as (Lee et al., 2008):

$$\|X\|_{\max} = \sup_{p,q} \|\text{diag}(p)X\text{diag}(q)^2\|_{\Sigma}, \quad (3)$$

where the supremum is over nonnegative unit vectors  $p, q$ . We can now continue using (2):

$$\begin{aligned} &\leq \sup_{p,q} \sqrt{\text{rank}(\text{diag}(p)X\text{diag}(q))} \cdot |\text{diag}(p)X\text{diag}(q)|_2 \\ &\leq \sup_{p,q} \sqrt{\text{rank}X} \cdot \sqrt{\sum_{ij} p_i^2 q_j^2 X_{ij}^2} = \sqrt{\text{rank}X} |X|_{\infty}. \end{aligned}$$

■

### 3. Reconstruction Guarantees

The theorems below provide reconstructions guarantees, first under the a mean-absolute-error reconstruction measure (Theorem 4) and then under a mean-squared-error reconstruction measure (Theorem 6). Since the guarantees are for *approximate* reconstruction, we must impose some notion of scale. In other words, we can think of measuring the error relative to the scale of the data—if  $Y$  is multiplied by some constant, then obviously the reconstruction error would also be multiplied by this constant. In the theorems below we refer to two notions of scale: the *average* squared magnitude of matrix entries, i.e.  $\frac{1}{nm} |X|_2^2$ , and the *maximal* magnitude of matrix entries, i.e.  $|X|_\infty$ . For simplicity and without loss of generality, the results are stated for unit scale.

An issue to take note of is whether the  $s$  observed entries of  $Y$  are chosen with or without replacement, i.e. whether we choose a set  $S$  of entries uniformly at random over all sets of exactly  $s$  entries (no replacements), or whether we make  $s$  independent uniform choices of entries, possibly observing the same entry twice. Our results apply in both cases.

**Theorem 4** *For any  $M, Y \in \mathbb{R}^{n \times m}$  where  $M$  is of rank at most  $r$ :*

a. **Entry magnitudes bounded on-average.** Consider the estimator<sup>1</sup>

$$\hat{X}(S) = \arg \min_{\|X\|_\Sigma \leq \sqrt{rnm}} \sum_{(i,j) \in S} |Y_{ij} - X_{ij}| .$$

If  $\frac{1}{nm} |M|_2^2 \leq 1$  and  $s \geq \mathbf{O}\left(\frac{r(n+m)\log(n)}{\epsilon^2}\right)$ , then in expectation over a sample  $S$  chosen either uniformly over sets of size  $s$  (without replacements) or by choosing  $s$  entries uniformly and independently (with replacements):

$$\frac{1}{nm} |Y - \hat{X}(S)|_1 \leq \frac{1}{nm} |Y - M|_1 + \epsilon .$$

b. **Entry magnitudes bounded uniformly.** Consider the estimator

$$\hat{X}(S) = \arg \min_{\|X\|_{\max} \leq \sqrt{r}} \sum_{(i,j) \in S} |Y_{ij} - X_{ij}| .$$

If  $|M|_\infty \leq 1$  and  $s \geq \mathbf{O}\left(\frac{r(n+m)}{\epsilon^2}\right)$ , then in expectation over a sample  $S$  of size  $s$  chosen either with or without replacements as above:

$$\frac{1}{nm} |Y - \hat{X}(S)|_1 \leq \frac{1}{nm} |Y - M|_1 + \epsilon .$$

**Remark 5** The above results can also be shown to hold in high probability over the sample  $S$ , rather than in expectation. Specifically, to ensure that the results of Theorem 4 hold with probability at least  $1 - n^{-\beta}$  (for sampling with replacement) or  $1 - n^{-(\beta-2)}$  (for sampling without replacement), it is sufficient to change the sample size requirement to  $s \geq \mathbf{O}\left(\frac{r(n+m)\log(n)+\beta\log(n)}{\epsilon^2}\right)$  (in the trace-norm case) or  $s \geq \mathbf{O}\left(\frac{r(n+m)+\beta\log(n)}{\epsilon^2}\right)$  (in the max-norm case).

---

1. If  $S$  is chosen with replacements, it is a multiset, and the summation  $\sum_{(i,j) \in S}$  should be interpreted as summation with repetitions.

**Theorem 6** For any  $Y = M + Z \in \mathbb{R}^{n \times m}$  where  $|Z|_\infty \leq \sqrt{\frac{rn}{\log n}}$  and  $M$  is of rank at most  $r$  with  $|M|_\infty \leq 1$ , denote  $\sigma^2 = \frac{1}{nm} |Z|_2^2$ . Consider the estimator

$$\hat{X}(S) = \arg \min_{\|X\|_{\max} \leq \sqrt{r}} \sum_{(i,j) \in S} (Y_{ij} - X_{ij})^2. \quad (4)$$

If  $s \geq \mathbf{O}\left(\frac{r(n+m)}{\epsilon} \cdot \frac{\sigma^2 + \epsilon}{\epsilon} \cdot (\log^3(r/\epsilon) + \beta)\right)$ , then, with probability at least  $1 - n^{-\beta}$  over a sample  $S$  of size  $s$  chosen with replacement, or with probability at least  $1 - n^{-(\beta-2)}$  over a sample  $S$  of size  $s$  chosen without replacement,

$$\frac{1}{nm} |Y - \hat{X}(S)|_2^2 \leq \sigma^2 + \epsilon. \quad (5)$$

If we instead use the estimator:

$$\hat{X}(S) = \arg \min_{\substack{\|X\|_{\max} \leq \sqrt{r} \\ |X|_\infty \leq 1}} \sum_{(i,j) \in S} (Y_{ij} - X_{ij})^2, \quad (6)$$

then we obtain (5) when  $s \geq \mathbf{O}\left(\frac{r(n+m)}{\epsilon} \cdot \frac{\sigma^2 + \epsilon}{\epsilon} \cdot (\log^3(1/\epsilon) + \beta)\right)$ .

The estimator (6) is SDP-representable, though potentially more cumbersome.

**Remark 7** The requirement on the maximal magnitude of the error in Theorem 6,  $|Z|_\infty \leq \sqrt{\frac{rn}{\log n}}$ , is very generous, and easily holds with high probability for sub-exponential noise. A stricter requirement, e.g.  $\mathbf{O}(\sqrt{r \log n})$ , which still holds with high probability for subgaussian noise, yields a guarantee with exponentially high probability  $1 - e^{-n/\log n}$ , without a sample-complexity dependence on  $\beta$ .

**Remark 8** A guarantee similar to Theorem 6 can be obtained if we can ensure  $\|M\|_{\max} \leq A$ , for some  $A$ , without requiring  $|M|_\infty \leq 1$ . For  $\hat{X}(S) = \arg \min_{\|X\|_{\max} \leq A} \sum_{ij \in S} (Y_{ij} - X_{ij})^2$ , we have (5) with a sample of size

$$s \geq \mathbf{O}\left(\frac{A^2(n+m)}{\epsilon} \cdot \frac{\sigma^2 + \epsilon}{\epsilon} \cdot (\log^3(A^2/\epsilon) + \beta)\right).$$

In Section 4.2.3, we will see how certain incoherence assumptions used in previous bounds yield a bound on  $\|M\|_{\max}$ , and compare the max-norm based reconstruction guarantee to the previously published results.

In Theorems 4 and 6 we do not assume the noise, i.e. the entries of  $Z = Y - M$ , are independent or zero-mean—in fact, we make no assumption on  $Z$ , other than the very generous upper bound  $|Z|_\infty \leq \sqrt{\frac{rn}{\log n}}$  discussed above. When entries of  $Z$  can be arbitrary, it is not possible to ensure reconstruction of  $M$  (e.g. we can set things up so  $Y$  actually has lower rank than  $M$ , and so it is impossible to identify  $M$ ). Consequently, in Theorems 4 and 6 we instead bound the excess error in predicting  $Y$  itself. If entries of  $Z$  are independent and zero-mean, then we may give the following guarantee about reconstructing the underlying matrix  $M$ :

**Theorem 9** For  $(i, j) \in [n] \times [m]$ , let  $\mathcal{F}_{(i,j)}$  be any mean-zero distribution. Suppose that the observed entries of  $Y$  are given by  $Y_{(i_t, j_t)} = M_{(i_t, j_t)} + Z_t$  for  $t = 1, 2, \dots, s$ , where  $(i_t, j_t) \stackrel{iid}{\sim} \text{Unif}([n] \times [m])$  and  $Z_t | (i_t, j_t) \sim \mathcal{F}_{(i_t, j_t)}$  independently for each  $t$ . That is, the noise is independent and zero-mean (though its distribution is allowed to depend on the location of the observation), the sample is drawn with replacement, and if an entry of the matrix is observed more than once, then the noise on the entry is drawn independently each time.

Assume  $|M|_\infty \leq 1$ ,  $\text{rank}(M) \leq r$ , and  $\sup_{t \in [s]} |Z_t| \leq \mathbf{o}\left(\sqrt{\frac{rn}{\log n}}\right)$  with high probability. Denote

$$\sigma^2 = \frac{1}{nm} \sum_{i,j} E_{Z_{ij} \sim \mathcal{F}_{ij}} (Z_{ij}^2) .$$

For the estimator given in Equation (4), with high probability over the sample  $S$  of size  $s \geq \mathbf{O}\left(\frac{r(n+m)}{\epsilon} \cdot \frac{\sigma^2 + \epsilon}{\epsilon} \cdot \log^3(r/\epsilon)\right)$ ,

$$\frac{1}{nm} |M - \hat{X}(S)|_2^2 \leq \epsilon . \quad (7)$$

Alternatively, if  $S$  is sampled uniformly without replacements, with the same assumptions and sample size, and as long as  $s \leq \frac{K+1}{e} (nm)^{1-\frac{1}{K+1}}$ , we have  $\frac{1}{nm} |M - \hat{X}(S)|_2^2 \leq 4K\epsilon$ .

**Remark 10** When sampling without replacement, we imposed both a lower bound and an upper bound on the sample size. For these two bounds to be compatible (in an asymptotic sense) for a fixed  $K$ , we need  $m = \Omega(n^a)$  for some positive power  $a$ , and make  $\epsilon$  arbitrarily small. Alternately, we can set  $K = \mathbf{O}(\log n)$ , ensuring the upper bound on  $s$  always holds (since  $s \leq nm$  necessarily), yielding  $\frac{1}{nm} |M - \hat{X}(S)|_2^2 \leq \epsilon$  whenever  $s \geq \mathbf{O}\left(\frac{r(n+m) \log(n)}{\epsilon} \cdot \frac{\sigma^2 + \epsilon}{\epsilon} \cdot \log^3(r/\epsilon)\right)$ .

The remainder of this Section is organized as follows: In Section 3.1, we prove Theorems 4 and 6 in the case where the sample is drawn without replacement. In Section 3.2, we discuss possible bounds of the mean-squared-error, as in Theorem 6, but using the trace-norm. In Section 3.3, we compare sampling with and without replacement, establishing Theorems 4 and 6 also for sampling with replacement. In Section 3.4, we turn to the setting of independent mean-zero noise, and prove Theorem 9 in both the sampling-with-replacement and sampling-without-replacement settings.

### 3.1. Proof of Theorems 4 and 6 when $S$ is drawn with replacement

We first establish the Theorems for a sample chosen i.i.d. with replacements. In this case, following Srebro and Shraibman (2005), we may view matrix reconstruction as a prediction problem, by regarding a matrix  $X \in \mathbb{R}^{n \times m}$  as a function  $[n] \times [m] \rightarrow \mathbb{R}$ . Each observation in the training set consists of a covariate  $(i, j) \in [n] \times [m]$  and an observed noisy response  $Y_{ij} \in \mathbb{R}$ . Here, we assume that the distribution over  $[n] \times [m]$  is uniform, and the joint distribution over  $(i, j)$  and its response is determined by the unknown  $Y$ . The hypothesis class is then a set of matrices bounded in either trace-norm or max-norm, and for a particular hypothesis  $X \in \mathbb{R}^{n \times m}$ , the averaged error  $\frac{1}{nm} |Y - X|_1$  or  $\frac{1}{nm} |Y - X|_2^2$  is equal to the

expected loss  $L(X) = \mathbf{E}_{ij} [\text{loss}(X_{ij}, Y_{ij})]$  under either the absolute-error or squared-error loss, respectively.

Srebro and Shraibman (2005) established bounds on the Rademacher complexity of the trace-norm and max-norm balls. For any sample of size  $s$ , the empirical Rademacher complexity of the max-norm ball is bounded by

$$\hat{\mathcal{R}}_s (\{X \in \mathbb{R}^{n \times m} \mid \|X\|_{\max} \leq A\}) \leq 12 \sqrt{\frac{A^2(n+m)}{s}}. \quad (8)$$

Although the empirical Rademacher complexity of the trace-norm ball might be fairly high, the *expected* Rademacher complexity, for a random sample of  $s$  independent *uniformly* chosen index pairs (with replacements) can be bounded as

$$\mathbf{E} [\hat{\mathcal{R}}_s (\{X \in \mathbb{R}^{n \times m} \mid \|X\|_{\Sigma} \leq A\})] \leq K \sqrt{\frac{\frac{A^2}{nm}(n+m)\log(n)}{s}} \quad (9)$$

for some numeric constant  $K$  (this is a slightly better bound than the one given by Srebro and Shraibman (2005), and is proved in Appendix B).

Since the absolute error loss,  $\text{loss}(x, y) = |x - y|$ , is 1-Lipschitz, these Rademacher complexity bounds immediately imply (Bartlett and Mendelson, 2001):

$$\frac{1}{nm} \left| Y - \hat{X}(S) \right|_1 \leq \inf_{\|X\|_{\max} \leq A} \left( \frac{1}{nm} |Y - X|_1 \right) + 24 \sqrt{\frac{A^2(n+m)}{s}} \quad (10)$$

for  $\hat{X}(S) = \arg \min_{\|X\|_{\max} \leq A} \sum_{(i,j) \in S} |Y_{ij} - X_{ij}|$ , and:

$$\frac{1}{nm} \left| Y - \hat{X}(S) \right|_1 \leq \inf_{\|X\|_{\Sigma} \leq A} \left( \frac{1}{nm} |Y - X|_1 \right) + 2K \sqrt{\frac{\frac{A^2}{nm}(n+m)\log(n)}{s}} \quad (11)$$

for  $\hat{X}(S) = \arg \min_{\|X\|_{\Sigma} \leq A} \sum_{(i,j) \in S} |Y_{ij} - X_{ij}|$ . These provide guarantees on reconstructing matrices with bounded max-norm or trace-norm. Choosing  $A = \sqrt{r}$  for the max-norm and  $A = \sqrt{rnm}$  for the trace-norm, Theorem 4 (for sampling with replacement) follows from Equation (2) and Lemma 3. (Remark 5 follows from the results of Bartlett and Mendelson (2001) with identical arguments for the sampling-with-replacement case.)

In order to obtain Theorem 6, we use a recent bound on the excess error with respect to a *smooth* (rather than Lipschitz) loss function, such as the squared loss. Specifically, Theorem 1 of Srebro et al. (2010) states that, for a class of predictors  $X : \mathcal{I} \rightarrow [-B, B]$  and a loss function bonded by  $b$  with second derivative bounded by  $H$ , with probability at least  $1 - \delta$  over a random sample of size  $s$ ,

$$L(\hat{X}) \leq L^* + O \left( \sqrt{L^* \tilde{\mathcal{R}}_s} + \tilde{\mathcal{R}}_s \right), \quad (12)$$

$$L^* = \inf_X L(X),$$

$$\tilde{\mathcal{R}}_s = H \mathcal{R}_s^2 \log^3 \left( \frac{B}{\mathcal{R}_s} \right) + \frac{b \log(\log(s)/\delta)}{s}, \quad (13)$$

where the infimum is over predictors in the class,  $\hat{X}$  is the empirical error minimizer in the class, and  $\mathcal{R}_s$  is an upper bound on the Rademacher complexity for all samples of size  $s$ .

In our case, for the class  $\{X \mid \|X\|_{\max} \leq A\}$  and the squared loss, we have  $B = \sup_X \sup_{ij} |X_{ij}| = \sup_X \|X\|_{\infty} \leq \sup_X \|X\|_{\max} \leq A$  and  $b = \sup_X |X - Y|_{\infty}^2 \leq \sqrt{\frac{4A^2(n+m)}{\log(n+m)}}$ , when we assume  $|Z|_{\infty} \leq A\sqrt{\frac{n+m}{\log(n+m)}}$ . Applying the bound (8) on the Rademacher complexity yields:

$$\tilde{\mathcal{R}}_s = \mathbf{O}\left(\frac{A^2(n+m)}{s} \log^3\left(\frac{s}{n}\right) + \frac{A^2(n+m) \log \log s}{s \log(n+m)} + \frac{A^2(n+m) \log(1/\delta)}{s \log n}\right) \quad (14)$$

$$= \mathbf{O}\left(\frac{A^2(n+m)}{s} \left(\log^3\left(\frac{s}{n+m}\right) + \frac{\log(1/\delta)}{\log n}\right)\right). \quad (15)$$

Here the last inequality uses the fact that  $s \leq n^2$ , while the next-to-last inequality assumes  $s \geq e^3(n+m)$ , and applies the fact that  $x^2 \log^3(1/x)$  is an increasing function for  $x < e^{-1.5}$ , where in this case  $x = \sqrt{\frac{n+m}{s}}$ .

Remark 8 follows immediately. The first claim in Theorem 6 follows when we assume  $|M|_{\infty} \leq 1$  and  $\text{rank}(M) \leq r$  and set  $A = \sqrt{r}$  (since, by Lemma 3,  $\|M\|_{\max} \leq A$ ). If we instead consider the class  $\{X : \|X\|_{\max} \leq \sqrt{r}, |X|_{\infty} \leq 1\}$ , then in the notation of (12), we may define  $B = 1$  instead of  $B = A = \sqrt{r}$ , and thus obtain

$$\tilde{\mathcal{R}}_s = \mathbf{O}\left(\frac{r(n+m)}{s} \left(\log^3\left(\frac{s}{r(n+m)}\right) + \frac{\log(1/\delta)}{\log n}\right)\right), \quad (16)$$

which yields the second claim of Theorem 6.

Finally, we prove the claim Remark 7. If instead we assume  $|Z|_{\infty} \leq \sqrt{r \log n}$ , then in the the notation of (12), we may define  $b = r \log n$  instead of  $b = \frac{4A^2n}{\log n} = \frac{4rn}{\log n}$ , and thus obtain

$$\tilde{\mathcal{R}}_s = \mathbf{O}\left(\frac{r(n+m)}{s} \left(\log^3\left(\frac{s}{(n+m)}\right) + \frac{\log n \cdot \log(1/\delta)}{n+m}\right)\right). \quad (17)$$

For  $\delta \leq e^{-n/\log n}$ , the second term is dominated by the first; therefore the sample complexity no longer depends on  $\beta$ .

### 3.2. Bounds on $\ell_2$ error using the trace norm

In Theorem 4, we saw that for mean-absolute-error matrix reconstruction, using the trace-norm instead of the max-norm allows us to forgo a bound on the spikiness, and rely only on the average squared magnitude  $\frac{1}{nm} |Y|_2^2$ . One might hope that we can similarly get a squared-error reconstruction guarantee using the trace-norm and without a spikiness bound that was required in Theorem 6. Unfortunately, this is not possible.

In fact, as the following example demonstrates, it is not possible to reconstruct a low-rank matrix to within much-better-than-trivial squared-error without a spikiness assumption, and relying only on  $\frac{1}{nm} |Y|_2 \leq 1$ . Specifically, consider an  $n \times m$  matrix

$$Y = \sqrt{m/r} (A | 0_{n \times (m-r)})$$

where  $A \in \{\pm 1\}^{n \times r}$  is an arbitrary sign matrix. The matrix  $Y$  has rank at most  $r$  and average squared magnitude  $\frac{1}{nm} \|Y\|_2^2 = 1$  (but maximal squared magnitude  $\|Y\|_\infty^2 = m/r$ ). Now, with even half the entries observed (i.e.  $s = nm/2$ ), we have no way of reconstructing the unobserved entries of  $A$ , as any values we choose for these entries would be consistent with the rank- $r$  assumption, yielding an expected average squared error of at least  $1/2$ . We can conclude that regardless of the estimator, controlling the average squared magnitude is not enough here, and we cannot expect to obtain a squared-error reconstruction guarantee based on  $\frac{1}{nm} \|Y\|_2^2$ , even if we use the trace-norm.

We note that if  $|M|_\infty, |Y|_\infty = \mathbf{O}(1)$ , then the squared-loss in the relevant regime has a bounded Lipschitz constants, and Theorem 4a applies. In particular, if  $|M|_\infty, |Y|_\infty \leq 1$ , then we can consider the estimator

$$\hat{X}(S) = \arg \min_{\substack{\|X\|_\Sigma \leq \sqrt{rnm} \\ \|X\|_\infty \leq 1}} \sum_{(i,j) \in S} (Y_{ij} - X_{ij})^2. \quad (18)$$

Since we now only need to consider  $X$  where  $|X_{ij} - Y_{ij}| \leq 2$ , the squared-loss in the relevant domain is 4-Lipschitz. We can therefore use the standard generalization results for Lipschitz loss as in Theorem 4, and obtain that with high probability over a sample of size

$$s \geq \mathbf{O} \left( \frac{r(n+m) \log n}{\epsilon^2} \right), \quad (19)$$

we have  $\frac{1}{nm} \|Y - \hat{X}(S)\|_2^2 \leq \sigma^2 + \epsilon$ . However, this result gives a dependence on  $\epsilon$  that is quadratic, as opposed to the more favorable dependence (at least when  $\epsilon = \Omega(\sigma^2)$ ) of Theorem 6.

We believe that, when  $|M|_\infty, |Y|_\infty \leq \mathbf{O}(1)$ , it is possible to improve the dependence on  $\epsilon$  to a dependence similar to that of Theorem 6 (this would require a more delicate analysis than that of Srebro et al. (2010), as their techniques rely on bounding the worst-case Rademacher complexity). But even this would not give any advantage over the max-norm, since the bound on  $|M|_\infty$  could not be relaxed, while an additional factor of  $\log n$  would be introduced into the sample complexity (coming from the Rademacher complexity calculation for the trace-norm). It seems then, that at least in terms of the quantities and conditions considered in this paper, as well as elsewhere in the low-rank reconstruction literature we are familiar with, there is no theoretical advantage for the trace-norm over the max-norm in terms of squared-error approximate reconstruction, though there could be an advantage for the max-norm in avoiding a logarithmic factor.

### 3.3. Sampling with or without replacement in Theorems 4 and 6

Theorems 4 and 6 give results that hold for either sampling with replacement or sampling without replacement. When an entry of the matrix  $Y$  is sampled twice, the same value is observed each time—no new information about the matrix is observed, and so intuitively, sampling without replacement should yield strictly better results than sampling with replacement. The two lemmas below, proved in the Appendix, establish that sampling without replacement is indeed as at least as good as sampling with replacement (up to a constant).

Before stating the lemmas, we briefly introduce some notation. Let  $L(X)$  denote the loss for an estimated matrix  $X$ ; that is,  $L(X) = \frac{1}{nm} \|Y - X\|_1$  or  $L(X) = \frac{1}{nm} \|Y - X\|_2^2$ , as

appropriate. Let  $\hat{L}_S(X)$  denote the empirical loss,  $\hat{L}_S(X) = \sum_{(i,j) \in S} |Y_{ij} - X_{ij}|^p$  (where  $p \in \{1, 2\}$  and the sum includes repeated elements in  $S$ ). Let  $\mathcal{D}^s$  and  $\mathcal{D}_{w/o}^s$  denote the distributions of a sample of size  $s$  drawn uniformly at random from the matrix, either with or without replacement, respectively.

**Lemma 11** *Let  $\mathcal{X}$  denote any class of matrices, with  $\mathcal{D}^s$  and  $\mathcal{D}_{w/o}^s$  defined as above. Then*

$$E_{S \sim \mathcal{D}_{w/o}^s} \left[ \sup_{X \in \mathcal{X}} L(X) - \hat{L}_S(X) \right] \leq E_{S \sim \mathcal{D}^s} \left[ \sup_{X \in \mathcal{X}} L(X) - \hat{L}_S(X) \right].$$

**Lemma 12** *Let  $\mathcal{X}$  denote any class of matrices, with  $\mathcal{D}^s$  and  $\mathcal{D}_{w/o}^s$  defined as above. Then for any  $c \in \mathbb{R}$ , and for any function  $g$ ,*

$$P_{S \sim \mathcal{D}_{w/o}^s} \left\{ \left( \sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \right) \geq c \right\} \leq 4s \cdot P_{S \sim \mathcal{D}^s} \left\{ \left( \sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \right) \geq c \right\}.$$

For the  $\ell_1$ -loss case, the Rademacher bounds (10) and (11) are derived from Bartlett and Mendelson (2001) by bounding  $E_{S \sim \mathcal{D}^s} (\sup_{X \in \mathcal{X}} L(X) - \hat{L}_S(X))$  (or by bounding  $P_{S \sim \mathcal{D}^s} (\sup_{X \in \mathcal{X}} L(X) - \hat{L}_S(X) \geq c)$ , for the proof of Remark 5). By Lemma 11, the same bound then holds for the same expectation taken over  $S \sim \mathcal{D}_{w/o}^s$ , and therefore (10) and (11) must hold for this case as well. This implies that the results of Theorem 4 (and Remark 5) hold for sampling without replacement as well as sampling with replacement.

Similarly, for the  $\ell_2$ -loss case, the Rademacher bound (12) is derived in Srebro et al. (2010) by bounding  $\sup_{X \in \mathcal{X}} L(X) - \sqrt{a \cdot L(X)} - \hat{L}_S(X)$  for some constant  $a$ , with probability at least  $1 - \delta$  over  $S \sim \mathcal{D}^s$ . Defining  $g(L) = L - \sqrt{a \cdot L}$ , the same bound must therefore hold with probability at least  $1 - 4s\delta \geq 1 - 4n^2\delta$  over  $S \sim \mathcal{D}_{w/o}^s$ , and therefore (12) holds for this case also. This implies that the results of Theorem 6 (and the subsequent remarks) hold for sampling without replacement as well as sampling with replacement.

### 3.4. Proof of Theorem 9: independent errors in the $\ell_2$ -loss setting.

First, we prove the theorem when sampling with replacement. For a matrix  $X$ , let  $L(X)$  denote the expected squared error for a randomly sampled entry, that is,

$$L(X) = \frac{1}{nm} \sum_{(i,j)} E((Y_{ij} - X_{ij})^2) = \frac{1}{nm} \sum_{(i,j)} E_{Z \sim \mathcal{F}_{(i,j)}} ((Z + M_{ij} - X_{ij})^2).$$

Now write  $\sigma^2 = \frac{1}{nm} \sum_{(i,j)} E_{Z \sim \mathcal{F}_{(i,j)}} (Z^2)$ . Then  $L(M) = \sigma^2$ .

Then, for any sample  $S$ , given  $\hat{X}(S)$  which is a random matrix depending on some observed sample, the expected loss (over a future observation of an entry in the matrix) of  $\hat{X}(S)$  satisfies the following (due to the fact that noise in a future observation of the matrix has zero mean and is independent from  $\hat{X}(S)$ ):

$$\begin{aligned} L(\hat{X}(S)) &= E_{(i,j)} \left( (Y_{ij} - \hat{X}(S)_{ij})^2 \middle| \hat{X}(S) \right) = E_{(i,j), Z \sim \mathcal{F}_{ij}} \left( (Z + M_{ij} - \hat{X}(S)_{ij})^2 \middle| \hat{X}(S) \right) \\ &= E_{(i,j), Z \sim \mathcal{F}_{ij}} \left( Z^2 + (M_{ij} - \hat{X}(S)_{ij})^2 \middle| \hat{X}(S) \right) = E_{(i,j), Z \sim \mathcal{F}_{ij}} (Z^2) + \frac{1}{nm} \|M - \hat{X}(S)\|_2^2 \\ &= \sigma^2 + \frac{1}{nm} \|M - \hat{X}(S)\|_2^2. \end{aligned}$$

Therefore, following the same reasoning as the proof of Theorem 6 (and Remark 7, we have that if  $s \geq \mathbf{O}\left(\frac{r(n+m)}{\epsilon} \cdot \frac{\sigma^2 + \epsilon}{\epsilon} \cdot \log^3(r/\epsilon)\right)$ , then with high probability,

$$L(\hat{X}) \leq \sigma^2 + \epsilon.$$

Applying the work above, we obtain

$$\frac{1}{nm}|M - \hat{X}(S)|_2^2 \leq \epsilon. \quad (20)$$

Now we turn to sampling without replacement. We first state a lemma which is proved in the appendix. (Notation: here  $\mathcal{D}^s$  and  $\mathcal{D}_{w/o}^s$  again denote sampling with or without replacement, but in this context  $\mathcal{D}^s$  represents sampling with replacement when the noise is added independently each time an entry is sampled, as in the statement of Theorem 9.)

**Lemma 13** *Let  $\mathcal{X}$  denote any class of matrices, with  $\mathcal{D}^s$  and  $\mathcal{D}_{w/o}^s$  defined as above. For any  $c$ , if  $s$  satisfies  $s \leq \frac{K+1}{e}(nm)^{1-\frac{1}{K+1}}$ , then*

$$P_{S \sim \mathcal{D}_{w/o}^s} \left( \sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq c \right) \leq 4K \cdot P_{S \sim \mathcal{D}^s} \left( \sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq (2K)^{-1}c \right).$$

As in the proof of the sampling-without-replacement case of Theorem 6, this is sufficient to show that  $\frac{1}{nm}|M - \hat{X}(S)|_2^2 \leq 4K \cdot \epsilon$  with high probability for the stated sample complexity, as long as we also have that  $s \leq \frac{K+1}{e}(nm)^{1-\frac{1}{K+1}}$ .

#### 4. Comparison to prior work

Suppose  $Y = M + Z$  where  $\text{rank}(M) \leq r$  and  $Z$  is a “noise” matrix of average squared magnitude  $\sigma^2 = \frac{1}{nm}|Z|_2^2$ , and we observe random entries of  $Y$ . One might then consider different types of reconstruction guarantees, requiring different assumptions on  $M$ ,  $Z$  and the sampling distribution:

$$\begin{aligned} \text{Exact recovery of } M : \quad & \hat{X}(S) = M. \\ \text{Near-exact recovery of } M : \quad & \frac{1}{nm}|\hat{X}(S) - M|_2^2 \leq \epsilon \cdot \sigma^2. \\ \text{Approximate recovery of } M : \quad & \frac{1}{nm}|\hat{X}(S) - M|_2^2 \leq \epsilon \cdot \text{scale}(M). \\ \text{Approximate recovery of } Y : \quad & \frac{1}{nm}|\hat{X}(S) - Y|_2^2 \leq \sigma^2 + \epsilon \cdot \text{scale}(M). \end{aligned}$$

Exact or near-exact recovery require strong incoherence-type assumptions on the matrix  $M$ , and is not possible for arbitrary low-rank matrices (see, e.g. Candès and Recht (2009)). Here we do not make any such assumptions, and show that approximate recovery is still possible. Such approximate recovery must be relative to some measure of the scale of  $M$ , and we discuss results relative to both the maximal magnitude,  $\text{scale}(M) = |M|_\infty^2$ , and the average squared magnitude  $\text{scale}(M) = \frac{1}{nm}|M|_2^2$ . Although not actually guaranteeing the same type of “recovery”, in Section 4.2 we nevertheless compare the sample complexity required for our approximate recovery results to the best sample complexity guarantee for exact and near-exact recovery (obtained by Recht (2009) and Keshavan et al. (2010), respectively), and comment on the differences between the required assumptions on  $M$ .

More directly comparable to our results are recent results by Keshavan et al. (2010), Negahban and Wainwright (2010) and Koltchinskii et al. (2010) on approximate recovery of  $M$ . These give essentially the same type of guarantee as in Theorem 9, and also rely on  $|M|_\infty^2$  as a measure of scale. In Section 4.1 we compare our guarantee to these results, discussing the different dependence on the various parameters and different assumptions on the noise. (Note that both types of results appear in Keshavan et al. (2010); in Section 4.1, we refer to the approximate recovery result stated in Theorem 1.1 of their paper, while in Section 4.2, we refer to the near-exact recovery result stated in Theorem 1.2 of their paper.)

Recovery of  $M$ , whether exact, near-exact, or approximate, also requires the noise to be independent and zero-mean, otherwise  $M$  might not be identifiable. All prior matrix reconstruction results we are aware of work in this setting. Approximate recovery of  $M$  also immediately implies an excess error bound on approximate recovery of  $Y$ . However, we also provide excess error bounds for approximate recovery of  $Y$ , that do *not* assume independent nor zero-mean noise (Theorems 4 and 6). That is, we provide reconstruction guarantees in a significantly less restrictive setting compared to other matrix reconstruction guarantees.

Another difference between different results is whether entries are sampled with or without replacement, and if replacement is allowed, whether the error is per-entry (i.e. repeat observations of the same entry are identical) or per-observation (i.e. repeat observations of the same entry are each corrupted independently). However, as we show in Sections 3.3 and 3.4, and as has also been shown for exact recovery (Recht, 2009), these differences do not significantly alter the quality of reconstruction or the required sample size.

The most common algorithm for low-rank matrix recovery in the literature is squared-error minimization subject to a penalty on trace norm. All the methods cited here prove results about some variation of this approach, with the exception of a recent result by Keshavan et al. (2010), which applies to the output of the local search procedure OPTSPACE. In contrast, our results are mostly for error minimization subject to a *max-norm* constraint.

#### 4.1. Comparison With Recent Approximate Recovery Guarantees

Negahban and Wainwright (2010) and Koltchinskii et al. (2010) recently presented guarantees on approximate recovery using trace-norm regularization, in a setting very similar to our Theorem 9. Earlier work by Keshavan et al. (2010) uses a low-rank SVD approximation to  $\tilde{Y}_S$  in the same setting to also obtain an approximate recovery guarantee. (Here  $Y_S$  is the matrix consisting of all observed entries of  $Y$ , with zeros elsewhere, and  $\tilde{Y}_S$  is the same matrix with overrepresented rows and columns removed.) In particular, each of the three guarantees provide an  $\epsilon$ -approximate reconstruction of  $M$  relative to  $|M|_\infty^2$ . That is, when  $|M|_\infty \leq 1$  as in Theorem 9, they provide the exact same guarantee  $\frac{1}{nm} |\hat{X}(S) - M| \leq \epsilon$ . (Negahban and Wainwright state the result relative to  $\frac{1}{nm} |M|_2^2$ , but have a linear dependence on the “spikiness”  $\frac{|M|_\infty}{|M|_2/\sqrt{nm}}$ , effectively giving a guarantee relative to  $|M|_\infty^2$ ).

Specifically, assuming  $|M|_\infty = 1$  without loss of generality, Negahban and Wainwright and Koltchinskii et al. assume the noise is independent and subgaussian (or subexponential) with variance  $\mathbf{O}(\sigma^2)$ , and require a sample size of:

$$s \geq \mathbf{O} \left( \frac{rn \log(n)}{\epsilon} \cdot (1 + \sigma^2) \right). \quad (21)$$

where the sample is drawn with replacement—in particular, an entry  $(i, j)$  of the matrix which is sampled multiple times gives multiple independent estimates of  $M_{ij}$ .

Keshavan et al. give a result on approximate recovery which holds with no assumption on the noise, but requires additional assumptions such as i.i.d. noise to be a meaningful bound. The estimator used is the rank- $r$  SVD approximation to  $\tilde{Y}_S$ , defined above. Specifically, they show that, for sufficiently large sample size, with high probability,  $\frac{1}{\sqrt{nm}}\|\hat{X}(S) - M\|_2 \leq \mathbf{O}\left(\frac{nr\sqrt{n/m}}{s} + \frac{nmr}{s^2}\|\tilde{Z}_S\|_2^2\right)$ , where  $\tilde{Z}_S$  is defined in the same way as  $\tilde{Y}_S$ . For this bound to be meaningful, there must be some distributional assumption on  $Z$ —otherwise, we could have  $\|Z_S\|_2 \approx |Z_S|_2 = \mathbf{O}(\sqrt{s})$ , and the bound on mean error would actually increase with  $\frac{nm}{s}$ , and is thus not a meaningful bound. In the presence of i.i.d. subgaussian noise, however, Keshavan et al. show that with high probability,  $\|\tilde{Z}_S\|_2^2 \leq \frac{\sigma^2(\sqrt{n/m}s\log(s))}{m}$ . Using this, approximate recovery of  $M$  is obtained for sample complexity

$$s \geq \mathbf{O}\left(\frac{rn}{\epsilon} \cdot (\sqrt{n/m}) \cdot (1 + \log(n)\sigma^2)\right), \quad (22)$$

where the sample is drawn *without* replacement. Therefore we may regard Keshavan et al.’s result as bounding error under the assumption of i.i.d. subgaussian noise (or perhaps some weaker assumption that gives the same result, such as independent subgaussian noise that might not be i.i.d., or similar). The guarantees (22) and (21) are therefore quite similar, even though they are for fairly different methods, with (22) being better when  $\sigma^2 = \mathbf{o}(1)$  but worse for highly rectangular matrices.

Comparing our Theorems 6 and 9 to the above, the advantages of our results are:

- We avoid the extra logarithmic dependence on  $n$ .
- Even in order to guarantee recovery of  $M$ , we assume only a much milder condition on the noise: that noise is mean-zero, and that with high probability,  $|Z_S|_\infty \leq \sqrt{\frac{rn}{\log n}}$ . We do not assume the noise is identically distributed, nor subgaussian or subexponential.
- We provide a guarantee on the excess error of recovering  $Y$ , even when the noise is *not* zero-mean nor independent.

The deficiency of our result is a possible slower rate of error decrease: when  $\sigma > 0$  and  $\epsilon = \mathbf{o}(\sigma^2)$  (i.e. to get “estimation error” significantly lower than the “approximation error”), our sample complexity scales as  $\tilde{\mathbf{O}}(1/\epsilon^2)$  compared to just  $\mathbf{O}(1/\epsilon)$  in the other results. We do not know if this difference represents a real consequence of not assuming zero-mean independent noise in our analysis, or just looseness in the proof. Our results also include an additional  $\log^3(1/\epsilon)$  factor, which we believe is purely an artifact of the proof technique.

A strength of our analysis, as compared to that of Negahban and Wainwright and Koltchinskii et al., is that the cases of sampling with and without replacement are both covered, including the case of per-entry noise when sampling with replacement, while the results of Negahban and Wainwright and Koltchinskii et al. are for sampling with replacement with per-observation noise. This is an important improvement because in many applications, the observed entries are drawn from a fixed matrix which was randomly generated, meaning that it is not possible to obtain multiple independent observations of any  $M_{ij}$ .

## 4.2. Comparison of results on exact and near-exact recovery

The results of Recht and of Keshavan et al. show that exact or near-exact recovery of the underlying low-rank matrix  $M$  can be obtained with high probability, when strong conditions on  $M$  are assumed, and when the observations are either noiseless (for Recht's exact recovery result) or are corrupted by i.i.d. subgaussian noise (for Keshavan et al.'s near-exact recovery result).

These results cannot be directly compared to the results we obtain in this paper, because the guarantees on recovery given by this work and by our work are fundamentally different—for instance, the error bound  $\epsilon$  has completely different meanings in our definitions of near-exact recovery and approximate recovery above. These two incomparable types of guarantees are linked to very different conditions on the data—exact and near-exact recovery cannot be obtained without strict assumptions about how the observations are generated.

Nonetheless, one comparison between these methods which can be made, is in the magnitude of the required sample complexities to obtain some meaningful bound via each result—exact recovery for Recht's result, near-exact recovery for Keshavan et al.'s result, and approximate recovery relative to  $|M|_\infty^2$  for our result. The rest of this section is organized as follows: we summarize the results in the literature in Section 4.2.1, compare sample complexities in Section 4.2.2, and describe how incoherence is sufficient but not necessary for approximate recovery relative to  $\frac{1}{nm} |M|_2^2$  (instead of  $|M|_\infty^2$ ) in Sections 4.2.3 and 4.2.4.

### 4.2.1. DETAILS ON EXACT AND NEAR-EXACT RESULTS IN THE LITERATURE

Let  $M = U\Sigma V^T$  be a reduced SVD of  $M$ . Let  $\kappa$  be the condition number of  $\Sigma$ . Define also the incoherence parameters for matrix  $M$  (Candès and Recht, 2009):

$$\begin{aligned}\mu_0 &= \max \left\{ \frac{n}{r} \cdot \max_i |U_{(i)}|_2^2, \frac{m}{r} \cdot \max_j |V_{(j)}|_2^2 \right\}, \\ \mu_1 &= \sqrt{\frac{nm}{r}} \cdot \max_{i,j} |U_{(i)}^T V_{(j)}|,\end{aligned}$$

where  $U_{(i)}$  denotes the  $i$ th row of  $U$  and  $V_{(j)}$  denotes the  $j$ th row of  $V$ .

Suppose that  $M$  has low incoherence parameters and  $Z = 0$ . Improving on the earlier results of Candès and Recht (2009) and Candes and Tao (2010), Recht proves that  $\hat{X}(S) = M$  (that is, exact recovery is obtained) with high probability if

$$s \geq \mathbf{O}(rn \max\{\mu_0, \mu_1^2\} \log^2 n). \quad (23)$$

In the case of noisy observations, Keshavan et al. give conditions on low  $\ell_2$  error in recovery (with high probability) in the setting of i.i.d. subgaussian noise with incoherent  $M$ , improving on Candes and Plan (2010) earlier work on the noisy case. (More precisely, Keshavan et al. give a result which holds with no assumption on the noise, but requires additional assumptions such as i.i.d. noise to be a meaningful bound. We therefore regard their result as assuming i.i.d. subgaussian noise—see the discussion of their approximate reconstruction result above in Section 4.1.) Their OPTSPACE algorithm is a method for finding the rank- $r$  matrix  $\hat{X}$  minimizing squared error on the observed entries. Let  $\hat{X}(S)$

denote the matrix recovered by this algorithm. When the entries of  $Z$  are i.i.d. subgaussian, Keshavan et al. show that, with high probability, if  $s$  satisfies

$$s \geq \mathbf{O} \left( rn\kappa^4 \cdot \max \left\{ \frac{1}{\epsilon} \log \left( \frac{rn\kappa^4}{\epsilon} \right), r\kappa^2\mu_0^2, r\kappa^2\mu_1^2 \right\} \right), \quad (24)$$

then  $|\hat{X}(S) - M|_2^2 \leq |Z|_2^2 \cdot \epsilon$ . (For simplicity of the comparison, we use a slightly relaxed form of their required sample complexity, and ignore  $\sqrt{n/m}$  in their error and sample bounds.)

#### 4.2.2. COMPARING SAMPLE COMPLEXITIES

Ignoring the dependence on  $\epsilon$ , which as we discussed earlier is in any case incomparable between approximate and exact and near-exact recovery, our sample complexity for approximate recovery using the max-norm is  $\mathbf{O}(rn)$ . Even with “perfect” incoherence parameters, this is a factor of  $\log^2(n)$  less than the sample complexity established by Recht for exact recovery (23), and a factor of  $r$  less than the sample complexity established by Keshavan et al. for near-exact recovery (24). Of course, “bad” incoherence parameters may sharply increase the sample complexity for exact or near-exact recovery, but do not affect our sample complexity for approximate recovery.

#### 4.2.3. APPROXIMATE RECOVERY RELATIVE TO AVERAGE SIGNAL MAGNITUDE, IN THE PRESENCE OF INCOHERENCE CONDITIONS

It is interesting to note that the incoherence assumptions, used by Recht and by Keshavan et al., enable approximate recovery with the max-norm relative to the average magnitude  $\frac{1}{nm}|M|_2^2$ , and not only the maximal magnitude, as in Theorem 6. This is based on the following observation:

**Lemma 14** *Let  $M \in \mathbb{R}^{n \times m}$  and let  $\kappa$  and  $\mu_0$  be defined as before. Then*

$$\|M\|_{\max} \leq \min\{\kappa, \sqrt{r}\}\mu_0\sqrt{r} \cdot \frac{|M|_2}{\sqrt{nm}}.$$

*In particular, by Lemma 3, the above expression is also an upper bound for  $|M|_\infty$ .*

**Proof** First, observe that

$$\|M\|_{\max} \leq \max_{i,j} |(U\Sigma)_{(i)}|_2 \cdot |V_{(j)}|_2 \leq \sigma_1 \cdot \max_{i,j} |U_{(i)}|_2 \cdot |V_{(j)}|_2 \leq \sigma_1 \cdot \frac{\mu_0 r}{\sqrt{nm}}.$$

Also,

$$\sigma_1 \leq \kappa\sqrt{\sigma_r^2} \leq \frac{\kappa}{\sqrt{r}} \sqrt{\sigma_1^2 + \dots + \sigma_r^2} = \frac{\kappa|M|_2}{\sqrt{r}} \text{ and } \sigma_1 \leq \sqrt{\sigma_1^2 + \dots + \sigma_r^2} = |M|_2.$$

■

Now, based on Remark 8, if  $\frac{1}{nm}|M|_2^2 \leq 1$  (and with a mild bound on  $|Z|_\infty$ ), with high probability over a sample of size

$$s \geq \mathbf{O} \left( \frac{rn}{\epsilon} \cdot \frac{\sigma^2 + \epsilon}{\epsilon} \cdot \min\{\kappa^2, r\}\mu_0^2 \cdot \log^3 \left( \frac{\mu_0^2 r}{\epsilon} \right) \right), \quad (25)$$

we have  $|Y - \hat{X}(S)|_2^2 \leq \sigma^2 + \epsilon$ . Up to log factors and the dependence on  $\epsilon$ , this sample complexity is at most as much as the sample complexity required by Keshavan et al., given in (24).

#### 4.2.4. APPROXIMATE RECOVERY RELATIVE TO AVERAGE SIGNAL MAGNITUDE, IN THE ABSENCE OF INCOHERENCE CONDITIONS

We make note of several special cases where using max-norm and the concentration result, and bounding excess error relative to  $\frac{1}{nm}|M|_2^2$ , may compare more favorably to other methods than the results above would indicate.

- If  $U = V$  (that is,  $M$  is symmetric), then  $\mu_1 = \mu_0\sqrt{r}$  and so our sample complexity compares more favorably to the sample complexities obtained by Recht and Keshavan et al. (which both involve  $\mu_1^2$ ).
- Our sample complexity uses Lemma 14 to bound  $\|M\|_{\max}$  relative to  $\frac{1}{\sqrt{nm}}|M|_2$ . An example where  $\kappa = 1$  and  $\|M\|_{\max} \ll \frac{\mu_0\sqrt{r}|M|_2}{\sqrt{nm}}$  (i.e. the bound in Lemma 14 is extremely loose) is the case where the spiky columns of  $U$  do not align with the spiky columns of  $V$ , for example writing  $n = m = N + 1$  we have:

$$M = \begin{pmatrix} 1 & 0 \\ 0 & N^{-1/2} \\ 0 & N^{-1/2} \\ \dots & \dots \\ 0 & N^{-1/2} \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ N^{-1/2} & 0 \\ N^{-1/2} & 0 \\ \dots & \dots \\ N^{-1/2} & 0 \end{pmatrix}^T = \begin{pmatrix} N^{-1/4} & 0 \\ 0 & N^{-1/4} \\ 0 & N^{-1/4} \\ \dots & \dots \\ 0 & N^{-1/4} \end{pmatrix} \cdot \begin{pmatrix} 0 & N^{-1/4} \\ N^{-1/4} & 0 \\ N^{-1/4} & 0 \\ \dots & \dots \\ N^{-1/4} & 0 \end{pmatrix}^T.$$

Since the left-hand factorization is an SVD of  $M$  (omitting  $\Sigma = I_2$ ), we therefore have  $\mu_0\sqrt{r} \cdot \frac{|M|_2}{\sqrt{nm}} = 1$  while the right-hand factorization shows that  $\|M\|_{\max} \leq \frac{1}{\sqrt{n-1}}$ .

- Large condition numbers  $\kappa$  can often lead to the same situation, in which the max norm is far lower than the bound implied by Lemma 14. For example, if low-rank  $M$  is a matrix where  $\|M\|_{\max} \approx \frac{\kappa\mu_0\sqrt{r}\cdot|M|_2}{\sqrt{nm}}$ , but if we perturb  $M$  slightly and add an extremely low singular value, then  $\kappa$  becomes extremely high while  $\|M\|_{\max}$  is only slightly perturbed.

## 5. Summary

We presented low rank matrix reconstruction guarantees based on an existing analysis of the Rademacher complexity of low trace-norm and low max-norm matrices, and carefully compared these to other recently presented results. We view the main contributions of this paper as:

- Following a string of results on low-rank matrix reconstruction, showing that an existing Rademacher complexity analysis combined with simple arguments on the relationship between the rank, max-norm, and trace-norm, can yield guarantees that are in several ways better, and relying on weaker assumptions.

- Pointing out that the max-norm can yield superior reconstruction guarantees over the more commonly used trace-norm.
- Studying the issue of sampling with and without replacement, and establishing rigorous generic results relating the two settings. This has been done before for exact recovery (Recht, 2009), but is done here for the more delicate situation of approximate recovery of either  $M$  or  $Y$ .

The main deficiency of our approach is a worse dependence on the approximation parameter  $\epsilon$ , when  $\sigma > 0$  (i.e. the approximately low rank case) and  $\epsilon = \mathbf{o}(\sigma^2)$  (i.e. estimation error less than approximation error). Although this dependence is tight for general classes with bounded Rademacher complexity, we do not know if it can be improved in Theorem 6. In particular, we do not know whether the less favorable dependence is a consequence of not relying on zero-mean i.i.d. noise, or not relying on  $M$  having low-rank (instead of only assuming low max-norm), or on relying only on the Rademacher complexity of the class of low max-norm matrices—perhaps better bounds can be obtained with a more careful analysis.

## References

- P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. In *Computational Learning Theory*, pages 224–240. Springer, 2001.
- E.J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.
- E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.
- E.J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.
- A. Chistov and D. Grigoriev. Complexity of quantifier elimination in the theory of algebraically closed fields. In *Proceedings of the 11th Symposium on Mathematical Foundations of Computer Science*, volume 176 of *Lecture Notes in Computer Science*, pages 17–31. Springer, 1984.
- M. Fazel, H. Hindi, and S.P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference*, volume 6, pages 4734–4739. IEEE, 2002.
- R.H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010.
- V. Koltchinskii, A.B. Tsybakov, and K. Lounici. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *arXiv:1011.6256*, 2010.
- T. Lee, A. Shraibman, and R. Spalek. A direct product theorem for discrepancy. In *23rd Annual IEEE Conference on Computational Complexity (CCC'08)*, pages 71–80. IEEE, 2008.

- S. Negahban and M.J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *arXiv:1009.2118*, 2010.
- B. Recht. A simpler approach to matrix completion. *arXiv:0910.0651*, 2009.
- R. Salakhutdinov and N. Srebro. Collaborative Filtering in a Non-Uniform World: Learning with the Weighted Trace Norm. *Advances in Neural Information Processing Systems*, 23: 2056–2064, 2010.
- Y. Segev. The expected norm of random matrices. *Combinatorics, Probability and Computing*, 9(2):149–166, 2000.
- A.A. Sherstov. Halfspace matrices. In *22nd Annual IEEE Conference on Computational Complexity (CCC'07)*, pages 83–95. IEEE, 2007.
- N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. *18th Annual Conference on Learning Theory (COLT)*, pages 545–560, 2005.
- N. Srebro, J.D.M. Rennie, and T.S. Jaakkola. Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems*, 17:1329–1336, 2005.
- N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. *Advances in Neural Information Processing Systems*, 23:2199–2207, 2010.
- J.A. Tropp. User-friendly tail bounds for sums of random matrices. *arXiv:1004.4389*, 2010.

## Appendix A. Proof of Sampling-Without-Replacement Lemmas

**Proof (Lemmas 11 and 12).** Let  $\mathbb{S}_r = \{S \in \mathcal{X}^s : \text{each } x \in \mathcal{X} \text{ appears at most } r \text{ times in } S\}$ . Let  $S \sim \mathcal{D}_r^s$  denote a sample  $S$  drawn uniformly from  $\mathbb{S}_r$ . In particular,  $\mathcal{D}_0^s = \mathcal{D}_{w/o}^s$  and  $\mathcal{D}_s^s = \mathcal{D}^s$ . By Lemma 15 (proved below), for any  $r$ ,

$$E_{S \sim \mathcal{D}_{w/o}^s} \left( \sup_{h \in \mathcal{H}} L(h) - \hat{L}_S(h) \right) \leq E_{S \sim \mathcal{D}_r^s} \left( \sup_{h \in \mathcal{H}} L(h) - \hat{L}_S(h) \right),$$

$$P_{S \sim \mathcal{D}_{w/o}^s} \left( \sup_{h \in \mathcal{H}} g(L(h)) - \hat{L}_S(h) \geq c \right) \leq r! \cdot P_{S \sim \mathcal{D}_r^s} \left( \sup_{h \in \mathcal{H}} g(L(h)) - \hat{L}_S(h) \geq c \right).$$

Taking the first inequality with  $r = s$ , this completes the proof for Lemma 11.

Now we complete the proof of Lemma 12. Take  $S \sim \mathcal{D}^s$  and write  $S = \{e_1, \dots, e_s\}$ . For any  $i_1 < i_2 < \dots < i_{K+1}$ ,

$$P(e_{i_1} = e_{i_2} = \dots = e_{i_{K+1}}) = \frac{1}{(nm)^K},$$

and so for any  $K$  with  $(K+1)! \geq 2s$ , the probability that any entry of the matrix appears at least  $(K+1)$  times in  $S$  is bounded by

$$\binom{s}{K+1} \cdot \frac{1}{(nm)^K} \leq \frac{s^{K+1}}{(K+1)!(nm)^K} \leq \frac{s}{(K+1)!} \leq \frac{1}{2}.$$

Fix the smallest  $K$  such that  $(K+1)! \geq 2s$ . This implies  $K! < 2s$ . We then have

$$\begin{aligned} & P_{S \sim \mathcal{D}_{w/o}^s} \left( \sup_{h \in \mathcal{H}} g(L(h)) - \hat{L}_S(h) \geq c \right) \\ & \leq K! \cdot P_{S \sim \mathcal{D}_K^s} \left( \sup_{h \in \mathcal{H}} g(L(h)) - \hat{L}_S(h) \geq c \right) \\ & \leq K! \cdot (P_{S \sim \mathcal{D}^s} (\text{each } x \in \mathcal{X} \text{ appears } \leq K \text{ times in } S))^{-1} \cdot P_{S \sim \mathcal{D}^s} \left( \sup_{h \in \mathcal{H}} g(L(h)) - \hat{L}_S(h) \geq c \right) \\ & \leq 2K! \cdot P_{S \sim \mathcal{D}^s} \left( \sup_{h \in \mathcal{H}} g(L(h)) - \hat{L}_S(h) \geq c \right) \\ & \leq 4s \cdot P_{S \sim \mathcal{D}^s} \left( \sup_{h \in \mathcal{H}} g(L(h)) - \hat{L}_S(h) \geq c \right). \end{aligned}$$

This completes the proof for Lemma 12. ■

**Lemma 15** *Using the notation of the proof above, for any  $r$ ,*

$$E_{S \sim \mathcal{D}_{w/o}^s} \left( \sup_{h \in \mathcal{H}} L(h) - \hat{L}_S(h) \right) \leq E_{S \sim \mathcal{D}_r^s} \left( \sup_{h \in \mathcal{H}} L(h) - \hat{L}_S(h) \right),$$

$$P_{S \sim \mathcal{D}_{w/o}^s} \left( \sup_{h \in \mathcal{H}} g(L(h)) - \hat{L}_S(h) \geq c \right) \leq r! \cdot P_{S \sim \mathcal{D}_r^s} \left( \sup_{h \in \mathcal{H}} g(L(h)) - \hat{L}_S(h) \geq c \right).$$

**Proof**

Write  $\Omega = [n] \times [m]$ . Let  $\alpha(S)$  be any function of the sample  $S$ , where  $S$  may contain repeated entries. Assume that, for any  $S, S_1, \dots, S_r$  of equal size such that  $r \cdot S = S_1 + \dots + S_r$ ,  $\alpha(\cdot)$  satisfies the following for some function  $a(r)$ :

$$a(r) \cdot \alpha(S) \leq \sum_{i=1}^r \alpha(S_i). \quad (26)$$

Consider all samples from  $\Omega$ , drawn with replacement. For a sample set  $S$  of size  $s$ , for  $i = 1, \dots, s$ , let  $N_i(S)$  equal the number of elements of  $\Omega$  appearing exactly  $i$  times in  $S$ , which obeys  $\sum_i i N_i(S) = s$ . We call  $\mathbf{N}(S) = (N_1(S), \dots, N_s(S))$  the multiplicity vector of  $S$ ; note that, when convenient, we might write  $\mathbf{N}(S)$  to have length greater than  $s$  (filling the last terms with zeros). From this point on, we will regard these samples as ordered lists, and assume that in any sample,  $S$  is ordered in the format

$$(\omega_1^1, \dots, \omega_{N_1(S)}^1, \omega_1^2, \omega_1^2, \dots, \omega_{N_2(S)}^2, \omega_{N_2(S)}^2, \omega_1^3, \omega_1^3, \omega_1^3, \dots),$$

where for any  $i$  we might permute the  $\omega_j^i$ 's.

Let  $\mathbf{N}$  be any multiplicity vector, of the form  $(N_1, \dots, N_r, 0, \dots, 0)$  for some  $r \leq s$ . Let  $\mathbf{N}'$  and  $\mathbf{N}''$  be multiplicity vectors derived from  $\mathbf{N}$  as follows:

$$N'_i = \begin{cases} N_1 + r N_r, & i = 1 \\ N_i, & 2 \leq i \leq r - 1 \\ 0, & i \geq r \end{cases}, \quad N''_i = \begin{cases} N_i, & 1 \leq i \leq r - 1 \\ 0, & i \geq r \end{cases}$$

Define  $s = \sum_i i N_i$ . Note that  $\sum_i i N'_i = s$  and  $\sum_i i N''_i = s - r N_r$ .

Let  $\mathbb{S} = \{S : \mathbf{N}(S) = \mathbf{N}\}$ ,  $\mathbb{S}' = \{S : \mathbf{N}(S) = \mathbf{N}'\}$ ,  $\mathbb{S}'' = \{S : \mathbf{N}(S) = \mathbf{N}''\}$ . We will first prove that  $E_{S' \sim \text{Unif}(\mathbb{S}')} [\alpha(S')] \leq E_{S \sim \text{Unif}(\mathbb{S})} [\alpha(S)]$ , and then induct on  $r$ .

First consider  $\mathbb{S}'$ . We have

$$|\mathbb{S}'| E_{S' \sim \text{Unif}(\mathbb{S}')} [\alpha(S')] = \sum_{S' \in \mathbb{S}'} [\alpha(S')] = \sum_{S'' \in \mathbb{S}''} \sum_{\substack{A_1, \dots, A_r \subset \Omega \setminus S'' \\ |A_j|=N_r \\ A_j \text{ disjoint}}} [\alpha(S'' + A_1 + \dots + A_r)].$$

The last equality arises when, starting with some  $S' \in \mathbb{S}'$ , we recall that  $S''$  is an ordered sample set beginning with the  $N_1 + r N_r$  elements which appear exactly once. Let  $S''$  be the first  $N_1$  elements of  $S'$ , then let  $A_1$  be the next  $N_r$  elements of  $S'$ , let  $A_2$  be the next  $N_r$  elements of  $S'$ , etc.

Next consider  $\mathbb{S}$ . As before, we have

$$|\mathbb{S}| E_{S \sim \text{Unif}(\mathbb{S})} [\alpha(S)] = \sum_{S \in \mathbb{S}} [\alpha(S)] = \sum_{S'' \in \mathbb{S}''} \sum_{\substack{A \subset \Omega \setminus S \\ |A|=N_r}} [\alpha(S'' + r \cdot A)].$$

By counting how many times each choice of  $A$  appears in the sum below, and then rescaling accordingly, we get

$$= \left( \frac{(nm - N_1 - \dots - N_r)!}{(nm - N_1 - \dots - N_{r-1} - r N_r)!} \right)^{-1} r^{-1} \sum_{S'' \in \mathbb{S}''} \sum_{\substack{A_1, \dots, A_r \subset \Omega \setminus S'' \\ |A_j|=N_r \\ A_j \text{ disjoint}}} \sum_j [\alpha(S'' + r \cdot A_j)]$$

$$\geq \left( \frac{(nm - N_1 - \dots - N_r)!}{(nm - N_1 - \dots - N_{r-1} - rN_r)!} \right)^{-1} \frac{a(r)}{r} \sum_{S'' \in \mathbb{S}''} \sum_{\substack{A_1, \dots, A_r \subset \Omega \setminus S'' \\ |A_j| = N_r \\ A_j \text{'s disjoint}}} \alpha(S'' + A_1 + \dots + A_r) .$$

To summarize so far, we have

$$|\mathbb{S}| E_{S \sim \text{Unif}(\mathbb{S})} [\alpha(S)] \geq \left( \frac{(nm - N_1 - \dots - N_r)!}{(nm - N_1 - \dots - N_{r-1} - rN_r)!} \right)^{-1} \cdot \frac{a(r)}{r} |\mathbb{S}'| E_{S' \sim \text{Unif}(\mathbb{S}')} [\alpha(S')] .$$

Next, we see that (since sample sets are treated as ordered)

$$|\mathbb{S}| = \frac{(nm)!}{(nm - N_1 - \dots - N_r)!}, \quad |\mathbb{S}'| = \frac{(nm)!}{(nm - N_1 - \dots - N_{r-1} - rN_r)!}$$

Therefore,

$$E_{S \sim \text{Unif}(\mathbb{S})} [\alpha(S)] \geq \frac{a(r)}{r} \cdot E_{S' \sim \text{Unif}(\mathbb{S}')} [\alpha(S')] .$$

By inducting over  $r$ , we then see that

$$E_{S \sim \text{Unif}(\mathbb{S})} [\alpha(S)] \geq \frac{\prod_{i=1}^r a(i)}{r!} \cdot E_{S \sim \mathcal{D}_{w/o}^s} [\alpha(S)] ,$$

where  $\mathbb{S} = \{S : \mathbf{N}(S) = \mathbf{N}\}$  for any multiplicity vector  $\mathbf{N} = (N_1, \dots, N_r, 0, \dots, 0)$ . Therefore,

$$E_{S \sim \mathcal{D}_r^s} [\alpha(S)] \geq \frac{\prod_{i=1}^r a(i)}{r!} \cdot E_{S \sim \mathcal{D}_{w/o}^s} [\alpha(S)] ,$$

Finally, we observe that if  $\alpha(S) = \sup_{h \in \mathcal{H}} L(h) - \hat{L}_S(h)$ , then  $\alpha(S)$  satisfies (26) with  $a(r) = r$ , while if  $\alpha(S) = \mathbb{I}\{\sup_{h \in \mathcal{H}} g(L(h)) - \hat{L}_S(h) \geq c\}$ , then  $\alpha(S)$  satisfies (26) with  $a(r) = 1$ . This concludes the proof. ■

### Proof (Lemma 13.)

Suppose  $s \leq \frac{K+1}{e} (nm)^{1-\frac{1}{K+1}}$ . Then, as in the proof of Lemma 12,

$$\begin{aligned} P(\text{any entry is sampled more than } K \text{ times}) &\leq \binom{s}{K} \cdot \frac{1}{(nm)^{K-1}} \\ &\leq \frac{s^{K+1}}{(K+1)!(nm)^K} \leq \frac{(K+1)/e)^{K+1} (nm)^K}{(K+1)!(nm)^K} \leq \frac{1}{2}, \text{ by Stirling's approximation.} \end{aligned}$$

We show below that, for any  $c$ ,

$$P_{S \sim \mathcal{D}_{w/o}^s} \left( \sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq c \right) \leq 2K \cdot P_{S \sim \mathcal{D}_K^s} \left( \sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq (2K)^{-1}c \right) ,$$

where  $\mathcal{D}_{w/o}^s$  and  $\mathcal{D}_K^s$  are defined as in the proof of Lemmas 11 and 12, except with the independent noise model. As in the proof of Lemmas 11 and 12, this implies that

$$P_{S \sim \mathcal{D}_{w/o}^s} \left( \sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq c \right) \leq 4K \cdot P_{S \sim \mathcal{D}^s} \left( \sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq (2K)^{-1}c \right) .$$

We now prove that, for any  $c$ ,

$$P_{S \sim \mathcal{D}_{w/o}^s} \left( \sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq c \right) \leq 2K \cdot P_{S \sim \mathcal{D}_K^s} \left( \sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq (2K)^{-1}c \right) .$$

Write  $\Omega = [n] \times [m]$ . Consider all samples from  $\Omega$ , drawn with replacement. When a particular  $(i, j)$  is drawn multiple times, then the observed values at that entry of the matrix follow the independent noise model as described in the statement of Theorem 9.

For a sample set  $S$  of size  $s$ , for  $i = 1, \dots, s$ , define  $\mathbf{N}(S)$  as in the proof of Lemma 15. Let  $\mathbf{N}$  be any multiplicity vector, of the form  $(N_1, \dots, N_r, 0, \dots, 0)$  for some  $r \leq s$ . Let  $\mathbf{M}$  be a multiplicity vector defined from  $\mathbf{N}$  as follows:

$$\mathbf{M} = (M_i)_i, \text{ where } M_i = N_{2i-1} + 2N_i + N_{i+1} .$$

Now take any  $A_1, A_2, \dots, A_{2r}, B_2, \dots, B_{2r} \subset [n] \times [m]$ , all disjoint, with  $|A_i| = |B_i| = N_i$  for all  $i$ . Define  $B_1 = A_1$ , and

$$S_A = \sum_{i=1}^{2r} \left( \sum_{j=1}^i A_i^{(j)} \right), \quad S_B = \sum_{i=1}^{2r} \left( \sum_{j=1}^i B_i^{(j)} \right) .$$

Note that  $\mathbf{N}(S_A) = \mathbf{N}(S_B) = \mathbf{N}$ . Now define

$$T_1 = \sum_{i=1}^{2r} \left( \sum_{j=1}^{\lfloor \frac{i}{2} \rfloor} A_i^{(j)} + \sum_{j=\lfloor \frac{i}{2} \rfloor + 1}^i B_i^{(j)} \right), \quad T_2 = \sum_{i=1}^{2r} \left( \sum_{j=1}^{\lfloor \frac{i}{2} \rfloor} B_i^{(j)} + \sum_{j=\lfloor \frac{i}{2} \rfloor + 1}^i A_i^{(j)} \right) .$$

Note that  $\mathbf{N}(T_1) = \mathbf{N}(T_2) = \mathbf{M}$ , and that up to reordering,  $S_A + S_B = T_1 + T_2$ . We treat  $T_1$  and  $T_2$  as functions of  $(S_A, S_B)$ .

Write  $\alpha_c(S) = \mathbb{I} \left\{ \sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq c \right\}$ . Then  $\alpha$  satisfies the following whenever  $|S_1| = |S_2|$ :

$$\frac{1}{2} (\alpha_{2c}(S_1) + \alpha_{2c}(S_2)) \leq \alpha_c(S_1 + S_2) \leq \alpha_c(S_1) + \alpha_c(S_2) .$$

Therefore,

$$\begin{aligned}
 & 2E_{S \sim \text{Unif}(\mathbf{N})}(\alpha_c(S)) \\
 &= (\#(S_A, S_B) \text{ pairs as above})^{-1} \sum_{(S_A, S_B) \text{ as above}} \alpha_c(S_A) + \alpha_c(S_B) \\
 &\geq (\#(S_A, S_B) \text{ pairs as above})^{-1} \sum_{(S_A, S_B) \text{ as above}} \alpha_c(S_A + S_B) \\
 &= (\#(S_A, S_B) \text{ pairs as above})^{-1} \sum_{(S_A, S_B) \text{ as above}} \alpha_c(T_1 + T_2) \\
 &\geq (\#(S_A, S_B) \text{ pairs as above})^{-1} \sum_{(S_A, S_B) \text{ as above}} \frac{1}{2} (\alpha_{2c}(T_1) + \alpha_{2c}(T_2)) \\
 &= (\#(S_A, S_B) \text{ pairs as above})^{-1} \sum_{(S_A, S_B) \text{ as above}} \alpha_{2c}(T_1) \\
 &= (\#(S_A, S_B) \text{ pairs as above})^{-1} \sum_{T: \mathbf{N}(T) = \mathbf{M}} \alpha_{2c}(T) \cdot (\#(S_A, S_B) \text{ pairs such that } T = T_1)
 \end{aligned}$$

We also have the following (note that here we treat samples as unordered, unlike in the proofs of Lemmas 11 and 12):

$$(\#(S_A, S_B) \text{ pairs as above}) = \binom{nm}{N_1, N_2, N_2, N_3, N_3, \dots, N_{2r}, N_{2r}},$$

and for any  $T$  with  $\mathbf{N}(T) = \mathbf{M}$ ,

$$(\#(S_A, S_B) \text{ pairs such that } T = T_1) = \prod_{i=1}^r \binom{M_i}{N_{2i-1}, N_{2i}, N_{2i}, N_{2i+1}}.$$

Finally,

$$(\#T : \mathbf{N}(T) = \mathbf{M}(T)) = \binom{nm}{M_1, M_2, \dots, M_r},$$

and therefore, continuing from above,

$$\begin{aligned}
 & 2E_{S \sim \text{Unif}(\mathbf{N})}(\alpha_c(S)) \\
 &= (\#(S_A, S_B) \text{ pairs as above})^{-1} \sum_{T: \mathbf{N}(T) = \mathbf{M}} \alpha_{2c}(T) \cdot (\#(S_A, S_B) \text{ pairs such that } T = T_1) \\
 &= (\#T : \mathbf{N}(T) = \mathbf{M})^{-1} \sum_{T: \mathbf{N}(T) = \mathbf{M}} \alpha_{2c}(T) \\
 &= E_{T \sim \text{Unif}(\mathbf{M})}(\alpha_{2c}(T)).
 \end{aligned}$$

Inducting over  $r$ , we see that for any  $\mathbf{N} = (N_1, \dots, N_r, 0, \dots, 0)$ ,

$$2^{K(r)} E_{S \sim \text{Unif}(\mathbf{N})}(\alpha_c(S)) \geq E_{S \sim \mathcal{D}_{w/o}^s}(\alpha_{2^{K(r)}}(S)),$$

where  $K(r)$  is the number of times that the operation  $x \mapsto \lceil x/2 \rceil$  must be applied iteratively to  $r$  to obtain 1; note that  $2^{K(r)} \leq 2r$ . Therefore,

$$P_{S \sim \mathcal{D}_{w/o}^s} \left( \sup_{X \in \mathcal{X}} g(L(x)) - \hat{L}_S(X) \geq c \right) \leq 2r P_{S \sim \mathcal{D}_r^s} \left( \sup_{X \in \mathcal{X}} g(L(x)) - \hat{L}_S(X) \geq (2r)^{-1}c \right). \quad \blacksquare$$

## Appendix B. The Rademacher Complexity of the Trace-Norm Ball

Srebro and Shraibman (2005) established that for a sample  $S = \{(i_1, j_1), \dots, (i_s, j_s)\}$  of  $s$  index-pairs, the empirical Rademacher complexity of the trace-norm ball, viewed a predictor of entries, is given by:

$$\begin{aligned} \hat{\mathcal{R}}_s(\{(i, j) \mapsto X_{ij} \mid X \in \mathbb{R}^{n \times m}, \|X\|_\Sigma \leq A\}) &= \mathbf{E}_\xi \left[ \sup_{\|X\|_\Sigma \leq A} \frac{1}{s} \sum_{t=1}^s \xi_t X_{(i_t, j_t)} \right] \\ &= \frac{A}{s} \mathbf{E}_\xi \left[ \left\| \sum_{t=1}^s \xi_t e_{i_t, j_t} \right\|_2 \right], \end{aligned} \quad (27)$$

where the expectations is over independent uniformly distributed random variables  $\xi_1, \dots, \xi_s \in \pm 1$ ,  $\|X\|_2$  is the spectral norm (maximal singular value) of  $X$ , and  $e_{i,j} = e_i e_j^T$  is a matrix with a single 1 at location  $(i, j)$  and zeros elsewhere. Analyzing the Rademacher complexity then amounts to analyzing the expected spectral norm of the random matrix  $Q = \sum_{t=1}^s \xi_t e_{i_t, j_t}$ .

The worst-case Rademacher complexity, i.e. the supremum of (27) over all samples  $S$ , is  $\frac{1}{\sqrt{s}}$ , and does not lead to meaningful generalization results. Indeed, if we could meaningfully bound the worst-case Rademacher complexity, we could guarantee learning under arbitrary sampling distributions over index-pairs, but this is not the case—we know that trace-norm regularization can fail when entries are not sampled uniformly (Salakhutdinov and Srebro, 2010).

Instead, we focus on bounding the *expected* Rademacher complexity, i.e. the expectation of (27) when entries in  $S$  are chosen independently from a *uniform* distribution over index pairs. Srebro and Shraibman (2005) bounded the expected Rademacher complexity by  $\mathbf{O}\left(\frac{A}{\sqrt{nm}} \sqrt{\frac{(n+m) \log^{3/2} n}{s}}\right)$  using a bound of Seginer (2000) on the spectral norm of a matrix with fixed magnitudes and random signs, combined with arguments bounding the number of observations in each row and column. Here we present a much simpler analysis, reducing the logarithmic factor from  $\log^{3/2}(n)$  to  $\log(n)$ , using a recent result of Tropp (2010).

We now proceed to bounding  $\mathbf{E}[\|Q\|_2]$ , where the expectation is over the sample  $S$  and the random signs  $\xi_t$ . Denote  $P_t = \xi_t e_{i_t, j_t}$ , we have  $Q = \sum_t P_t$  and  $P_t$  are i.i.d. zero-mean random matrices (recall that now both  $\xi_t$  and  $(i_t, j_t)$  are random). Theorem 6.1 of Tropp (2010), combined with Remarks 6.3 and 6.5, allows us to bound the expected spectral norm of such a sum of independent random matrices by:

$$\mathbf{E}[\|Q\|] = \mathbf{O}\left(\sigma \sqrt{\log(n+m)} + R \log(n+m)\right), \quad (28)$$

where  $\|P_t\|_2 \leq R$  (almost surely) and

$$\sigma^2 = \max \left( \left\| \sum \mathbf{E} [P_t^T P_t] \right\|_2, \left\| \sum \mathbf{E} [P_t P_t^T] \right\|_2 \right).$$

For each  $t$ ,  $P_t$  is just a matrix with a single  $+1$  or  $-1$ , hence  $\|P_t\| \leq 1$ . The matrix  $P_t P_t^T \in \mathbb{R}^{n \times n}$  is equal to  $e_{i,i}$  with probability  $\frac{1}{n}$ , hence  $\mathbf{E} [P_t P_t^T] = \frac{1}{n} I_n$  and  $\left\| \sum \mathbf{E} [P_t P_t^T] \right\|_2 = \left\| \frac{s}{n} I_n \right\| = \frac{s}{n}$ . Symmetrically,  $\left\| \sum \mathbf{E} [P_t^T P_t] \right\|_2 = \frac{s}{m}$  and so  $\sigma^2 = \frac{s}{nm} \max(n, m)$ . Plugging  $\sigma$  and  $T$  into (28) we have:

$$\mathbf{E} [\|Q\|_2] = O \left( \sqrt{\frac{s(n+m) \log(n+m)}{nm}} + \log(n+m) \right) = O \left( \sqrt{\frac{s(n+m) \log(n+m)}{nm}} \right) \quad (29)$$

where in the second inequality we assume  $s \geq m$ . Plugging (29) into (27) we get:

$$\mathbf{E} \left[ \hat{\mathcal{R}}_s (\{(i,j) \rightarrow X_{ij} \mid X \in \mathbb{R}^{n \times m}, \|X\|_\Sigma \leq A\}) \right] = O \left( \frac{A}{\sqrt{nm}} \sqrt{\frac{(n+m) \log(n+m)}{s}} \right) \quad (30)$$

FOYGEL SREBRO

# On the Consistency of Multi-Label Learning

**Wei Gao**

GAOW@LAMDA.NJU.EDU.CN

**Zhi-Hua Zhou**

ZHOUZH@LAMDA.NJU.EDU.CN

*National Key Laboratory for Novel Software Technology*

*Nanjing University, Nanjing 210093, China*

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

Multi-label learning has attracted much attention during the past few years. Many multi-label learning approaches have been developed, mostly working with surrogate loss functions since multi-label loss functions are usually difficult to optimize directly owing to non-convexity and discontinuity. Though these approaches are effective, to the best of our knowledge, there is no theoretical result on the convergence of risk of the learned functions to the Bayes risk. In this paper, focusing on two well-known multi-label loss functions, i.e., *ranking loss* and *hamming loss*, we prove a necessary and sufficient condition for the consistency of multi-label learning based on surrogate loss functions. Our results disclose that, surprisingly, none convex surrogate loss is consistent with the ranking loss. Inspired by the finding, we introduce the *partial ranking loss*, with which some surrogate functions are consistent. For hamming loss, we show that some recent multi-label learning approaches are inconsistent even for deterministic multi-label classification, and give a surrogate loss function which is consistent for the deterministic case. Finally, we discuss on the consistency of learning approaches which address multi-label learning by decomposing into a set of binary classification problems.

**Keywords:** Consistency, multi-label learning, surrogate loss, ranking loss, hamming loss

## 1. Introduction

In traditional supervised learning, each instance is associated with a single label. In real-world applications, however, one object is usually relevant to multiple labels simultaneously. For example, a document about national education service may be categorized into several predefined topics, such as `government` and `education`; an image containing forests may be annotated with `trees` and `mountains`. For learning with such objects, *multi-label learning* has attracted much attention during the past few years and many effective approaches have been developed (Schapire and Singer, 2000; Elisseeff and Weston, 2002; Zhou and Zhang, 2007; Zhang and Zhou, 2007; Hsu et al., 2009; Dembczyński et al., 2010; Petterson and Caetano, 2010).

The *consistency* (also called Bayes consistency) of learning algorithms concerns if the expected risk of a learned function converges to the Bayes risk as the training sample size increases (Lin, 2002; Zhang, 2004b; Steinwart, 2005; Bartlett et al., 2006; Tewari and Bartlett, 2007; Duchi et al., 2010). Nowadays, it is well-accepted that a good learner should at least be consistent with large samples. It is noteworthy that, though many efforts have been devoted to multi-label learning, few theoretical aspects were explored; in particular,

to the best of our knowledge, the consistency of multi-label learning remains untouched although it is a very important theoretical issue.

In this paper, focusing on two well-known multi-label loss functions, i.e., *ranking loss* and *hamming loss*, we present a theoretical study on the consistency of multi-label learning. We prove a necessary and sufficient condition for the consistency of multi-label learning based on surrogate loss functions. Our analysis discloses that, surprisingly, any convex surrogate loss is inconsistent with the ranking loss. Based on this finding, we introduce the *partial ranking loss*, which is consistent with some surrogate loss functions; we also show that many current multi-label learning approaches are even not consistent with the partial ranking loss. As for hamming loss, our analysis shows that some recent multi-label learning approaches are inconsistent even for deterministic multi-label classification, and we give a surrogate loss which is consistent with hamming loss for the deterministic case. Finally, we discuss on the consistency of approaches which address multi-label learning by transforming the problem into a set of binary classification problems.

The rest of this paper is organized as follows. Section 2 briefly introduces the research background. Section 3 presents our necessary and sufficient condition for the consistency of multi-label learning. Sections 4 and 5 study the consistency of multi-label learning approaches with regard to the ranking loss and hamming loss, respectively. Section 6 gives the detailed proofs.

## 2. Background

Let  $\mathcal{X}$  be an instance space and  $\mathcal{L} = \{1, 2, \dots, Q\}$  denotes a finite set of possible labels. An instance  $X \in \mathcal{X}$  is associated with a subset of labels  $Y \subset \mathcal{L}$  which is called *relevant labels*, while the complement  $\mathcal{L} \setminus Y$  is called *irrelevant labels*. For convenience of discussion, we represent the labels as a binary vector  $Y = (y_1, y_2, \dots, y_Q)$ , where  $y_i = +1$  if label  $i$  is relevant to  $X$  and  $-1$  otherwise, and denote by  $\mathcal{Y} = \{+1, -1\}^Q$  the set of all possible labels. Let  $\mathcal{D}$  denote an unknown underlying probability distribution over  $\mathcal{X} \times \mathcal{Y}$ . For integer  $m > 0$ , we denote by  $[m] = \{1, 2, \dots, m\}$ , and for real  $r$ ,  $\lfloor r \rfloor$  denotes the greatest integer which is no more than  $r$ .

The formal description of multi-label learning in the probabilistic setting is given as follows: Given a training sample  $S = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_m, Y_m)\}$  drawn i.i.d. according to the distribution  $\mathcal{D}$ , the objective is to learn a function  $h: \mathcal{X} \rightarrow \mathcal{Y}$ , which is able to assign a set of labels to unseen instances. In general, it is not easy to learn  $h$  directly, and in practice, one instead learns a real-valued vector function

$$\mathbf{f} = (f_1, f_2, \dots, f_K): \mathcal{X} \rightarrow \mathbb{R}^K \text{ for some integer } K > 0,$$

where  $K = Q$  or  $K = 2^Q$  are common choices. Based on this vector function  $\mathbf{f}$ , a prediction function  $F: \mathbb{R}^K \rightarrow \mathcal{Y}$  can be attained for assigning the set of relevant labels to an instance. Another popular approach for multi-label learning is to learn a real-valued vector function  $\mathbf{f} = (f_1, f_2, \dots, f_Q)$  such that  $f_i(X) > f_j(X)$  if  $y_i = +1$  and  $y_j = -1$  for an instance-label pair  $(X, Y)$ , and a function  $F$  should be learned to determine the number of relevant labels.

Essentially, multi-label learning approaches try to minimize the expected risk of  $\mathbf{f}$  with regard to some loss  $L$ , i.e.,

$$R(\mathbf{f}) = \mathbb{E}_{(X, Y) \sim \mathcal{D}} [L(\mathbf{f}(X), Y)]. \quad (1)$$

Notice that  $\mathbf{f}$  could be a prediction function or a vector of real-valued functions according to different losses. Further denote by  $R^* = \inf_{\mathbf{f}} R(\mathbf{f})$  the minimal risk (e.g., the Bayes risk) over all measurable functions. In this paper, we mainly focus on loss functions which are *below-bounded* and *interval*, defined as follows:

**Definition 1** A loss function  $L$  is said to be *below-bounded* if  $L(\cdot, \cdot) \geq B$  holds for some constant  $B$ ;  $L$  is said to be *interval* if it holds, for some constant  $\gamma > 0$ , that either

$$L(\mathbf{f}(X), Y) = L(\mathbf{f}(X'), Y') \quad \text{or} \quad |L(\mathbf{f}(X), Y) - L(\mathbf{f}(X'), Y')| \geq \gamma,$$

for every  $X, X' \in \mathcal{X}$  and  $Y, Y' \in \mathcal{Y}$ .

There are many multi-label loss functions (also called *evaluation criteria*), e.g., *ranking loss*, *hamming loss*, *one-error*, *coverage* and *average precision* (Schapire and Singer, 2000; Zhang and Zhou, 2006); *accuracy*, *precision*, *recall* and  $F_1$  (Godbole and Sarawagi, 2004; Qi et al., 2007); *subset accuracy* (Ghamrawi and McCallum, 2005); etc. In this paper, we focus on two well-known losses, i.e., ranking loss and hamming loss, and leave the discussion on other losses to future work.

Notice that all the above multi-label losses are non-convex and discontinuous. It is difficult, or even impossible, to optimize these losses directly. A feasible method in practice is to consider instead a surrogate loss function which can be optimized by efficient algorithms. Actually, most existing multi-label learning approaches, such as the boosting algorithm AdaBoost.MH (Schapire and Singer, 2000), neural network algorithm BP-MIL (Zhang and Zhou, 2006), SVM-style algorithms (Elisseeff and Weston, 2002; Taskar et al., 2004; Hariharan et al., 2010), etc., in essence, try to optimize some surrogate losses such as the exponential loss and hinge loss.

There are many definitions of consistency, e.g., the Fisher consistency (Lin, 2002), infinite-sample consistency (Zhang, 2004a), classification calibration (Bartlett et al., 2006; Tewari and Bartlett, 2007), edge-consistency (Duchi et al., 2010), etc., and the consistency of learning algorithms based on optimizing a surrogate loss function has been well-studied for binary classification (Zhang, 2004b; Steinwart, 2005; Bartlett et al., 2006), multi-class classification (Zhang, 2004a; Tewari and Bartlett, 2007), learning to rank (Cossack and Zhang, 2008; Xia et al., 2008; Duchi et al., 2010), etc. The consistency of multi-label learning, however, remains untouched, and to the best of our knowledge, this paper presents the first theoretical analysis on the consistency of multi-label learning.

### 3. Multi-Label Consistency

Given an instance  $X \in \mathcal{X}$ , we denote by  $\mathbf{p}(X)$  a vector of conditional probability for  $Y \in \mathcal{Y}$ , i.e.,

$$\mathbf{p}(X) = (p_Y(X))_{Y \in \mathcal{Y}} = (\Pr(Y|X))_{Y \in \mathcal{Y}},$$

for some  $\mathbf{p}(X) \in \Lambda$ , where  $\Lambda$  is the set of all possible conditional probability distribution vectors, i.e.,

$$\Lambda = \left\{ \mathbf{p} \in \mathbb{R}^{|\mathcal{Y}|} : \sum_{Y \in \mathcal{Y}} p_Y = 1 \text{ and } p_Y \geq 0 \right\}.$$

In the following, for notational simplicity, we will suppress dependence of  $\mathbf{p}(X)$  and  $\mathbf{f}(X)$  on the instance  $X$  as  $\mathbf{p}$  and  $\mathbf{f}$ , respectively, when it is clear from the context.

For an instance  $X \in \mathcal{X}$ , we define the conditional risk of  $\mathbf{f}$  as

$$l(\mathbf{p}, \mathbf{f}) = \sum_{Y \in \mathcal{Y}} p_Y L(\mathbf{f}, Y) = \sum_{Y \in \mathcal{Y}} \Pr(Y|X) L(\mathbf{f}(X), Y). \quad (2)$$

It is easy to get the expected risk and the minimal risk, respectively, as

$$R(\mathbf{f}) = \mathbb{E}_X[l(\mathbf{p}, \mathbf{f})] \quad \text{and} \quad R^* = \mathbb{E}_X \left[ \inf_{\mathbf{f}} l(\mathbf{p}, \mathbf{f}) \right].$$

We further define the *set of Bayes predictors* as

$$A(\mathbf{p}) = \{\mathbf{f}: l(\mathbf{p}, \mathbf{f}) = \inf_{\mathbf{f}'} l(\mathbf{p}, \mathbf{f}')\}.$$

Notice that  $A(\mathbf{p}) \neq \emptyset$  since  $L$  is interval and below-bounded.

As mentioned above, the loss  $L$  in multi-label learning is generally non-convex and discontinuous, and it is difficult, even impossible, to minimize the risk given by Eq.(1) directly. In practice, a surrogate loss function  $\Psi$  is usually considered in place of  $L$ . We define the  $\Psi$ -risk and Bayes  $\Psi$ -risk of  $\mathbf{f}$ , respectively, as

$$R_\Psi(\mathbf{f}) = \mathbb{E}_{X,Y}[\Psi(\mathbf{f}(X), Y)] \quad \text{and} \quad R_\Psi^* = \inf_{\mathbf{f}} R_\Psi(\mathbf{f}).$$

Similarly, we define the conditional surrogate risk and the conditional Bayes surrogate risk of  $\mathbf{f}$ , respectively, as

$$W(\mathbf{p}, \mathbf{f}) = \sum_{Y \in \mathcal{Y}} p_Y \Psi(\mathbf{f}, Y) \quad \text{and} \quad W^*(\mathbf{p}) = \inf_{\mathbf{f}} W(\mathbf{p}, \mathbf{f}).$$

It is obvious that  $R_\Psi(\mathbf{f}) = \mathbb{E}_X[W(\mathbf{p}, \mathbf{f})]$  and  $R_\Psi^* = \mathbb{E}_X[W^*(\mathbf{p})]$ .

We now define the *multi-label consistency* as follows:

**Definition 2** *Given a below-bounded surrogate loss function  $\Psi$  where  $\Psi(\cdot, Y)$  is continuous for every  $Y \in \mathcal{Y}$ ,  $\Psi$  is said to be multi-label consistent w.r.t. the loss  $L$  if it holds, for every  $\mathbf{p} \in \Lambda$ , that*

$$W^*(\mathbf{p}) < \inf_{\mathbf{f}} \{W(\mathbf{p}, \mathbf{f}): \mathbf{f} \notin A(\mathbf{p})\}.$$

The following theorem states that the multi-label consistency is a necessary and sufficient condition for the convergence of  $\Psi$ -risk to the Bayes  $\Psi$ -risk, implying  $R(\mathbf{f}) \rightarrow R^*$ .

**Theorem 3** *The surrogate loss  $\Psi$  is multi-label consistent w.r.t. the loss  $L$  if and only if it holds for any sequence  $\mathbf{f}_n$  that*

$$R_\Psi(\mathbf{f}_n) \rightarrow R_\Psi^* \quad \text{then} \quad R(\mathbf{f}_n) \rightarrow R^*.$$

We defer the proof of this theorem to Section 6.1, which is inspired by the technique of (Zhang, 2004a) and (Tewari and Bartlett, 2007).

#### 4. Consistency w.r.t. Ranking Loss

The ranking loss concerns about the label pairs which are ordered reversely for an instance. For a real-valued vector function  $\mathbf{f} = (f_1, f_2, \dots, f_Q)$ , the ranking loss is given by

$$L_{\text{rankloss}}(\mathbf{f}, (X, Y)) = \sum_{\substack{y_i=-1 \\ y_j=+1}} a_Y I[f_i(X) \geq f_j(X)] = \sum_{y_i < y_j} a_Y I[f_i(X) \geq f_j(X)], \quad (3)$$

where  $a_Y$  is a non-negative penalty and  $I$  is the indicator function, i.e.,  $I[\pi]$  equals 1 if  $\pi$  holds and 0 otherwise. In multi-label learning, the most common penalty is

$$a_Y = |\{y_i : y_i = -1\}|^{-1} \times |\{y_j : y_j = +1\}|^{-1}.$$

In this paper we consider the more general penalty, i.e., any non-negative penalty. It is easy to see that the ranking loss is below-bounded and interval since  $L_{\text{rankloss}}(\mathbf{f}, (X, Y)) \geq 0$ , and for every  $X, X' \in \mathcal{X}$  and  $Y, Y' \in \mathcal{Y}$ , it holds that either

$$L_{\text{rankloss}}(\mathbf{f}, (X, Y)) = L_{\text{rankloss}}(\mathbf{f}, (X', Y')) \text{ or } |L_{\text{rankloss}}(\mathbf{f}, (X, Y)) - L_{\text{rankloss}}(\mathbf{f}, (X', Y'))| \geq \gamma,$$

where  $\gamma = \min \{ |ia_Y - ja_{Y'}| > 0 : i, j \in [(Q - \lfloor \frac{Q}{2} \rfloor) \cdot \lfloor \frac{Q}{2} \rfloor]\}$ . Considering Eq.(3), a natural choice for the surrogate loss is

$$\Psi(\mathbf{f}(X), Y) = \sum_{\substack{y_i=-1 \\ y_j=+1}} a_Y \phi(f_j(X) - f_i(X)) = \sum_{y_i < y_j} a_Y \phi(f_j(X) - f_i(X)), \quad (4)$$

where  $\phi$  is a convex and non-increasing real-valued function, which was chosen as hinge loss  $\phi(x) = (1 - x)_+$  in (Elisseeff and Weston, 2002) and exponential loss  $\phi(x) = \exp(-x)$  in (Schapire and Singer, 2000; Dekel et al., 2004; Zhang and Zhou, 2006).

Before our discussion, it is necessary to introduce the notations

$$\Delta_{i,j} = \sum_{Y: y_i < y_j} p_Y a_Y \quad \text{and} \quad \delta(i, j; k, l) = \sum_{Y: y_i < y_j, y_k < y_l} p_Y a_Y,$$

for a given vector  $\mathbf{p} \in \Lambda$  and a non-negative vector  $(a_Y)_{Y \in \mathcal{Y}}$ . It is easy to get the following:

**Lemma 4** *For a vector  $\mathbf{p} \in \Lambda$  and a non-negative vector  $(a_Y)_{Y \in \mathcal{Y}}$ , the following properties hold:*

1.  $\Delta_{i,i} = 0$ ;
2.  $\delta(i, j; k, l) = \delta(k, l; i, j)$ ;
3.  $\Delta_{i,j} = \delta(i, j; i, k) + \delta(i, j; k, j)$  for every  $k \neq i, j$ ;
4.  $\Delta_{i,k} + \Delta_{k,j} + \Delta_{j,i} = \Delta_{k,i} + \Delta_{i,j} + \Delta_{j,k}$ ;
5.  $\Delta_{i,k} \leq \Delta_{k,i}$  if  $\Delta_{i,j} \leq \Delta_{j,i}$  and  $\Delta_{j,k} \leq \Delta_{k,j}$ .

**Proof** The Properties 1 and 2 are immediate from the definitions. For Property 3, we have  $y_i = -1$  and  $y_j = +1$  for every  $Y \in \mathcal{Y}$  satisfying  $y_i < y_j$ . For  $y_k$  ( $k \neq i, j$ ), there are only two choices:  $y_k = +1$  or  $y_k = -1$ , and thus  $\Delta_{i,j} = \delta(i, j; i, k) + \delta(i, j; k, j)$ . For Property 4, we have, from Property 3:

$$\begin{aligned}\Delta_{i,k} &= \delta(i, k; j, k) + \delta(i, k; i, j), \quad \Delta_{k,i} = \delta(k, i; j, i) + \delta(k, i; k, j), \\ \Delta_{j,i} &= \delta(j, i; k, i) + \delta(j, i; j, k), \quad \Delta_{i,j} = \delta(i, j; k, j) + \delta(i, j; i, k), \\ \Delta_{k,j} &= \delta(k, j; k, i) + \delta(k, j; i, j), \quad \Delta_{j,k} = \delta(j, k; j, i) + \delta(j, k; i, k).\end{aligned}$$

Thus, it holds by combining with Property 2. Property 5 follows from Property 4. ■

**Lemma 5** For every  $\mathbf{p} \in \Lambda$  and non-negative vector  $(a_Y)_{Y \in \mathcal{Y}}$ , the set of Bayes predictors for ranking loss is given by

$$A(\mathbf{p}) = \{\mathbf{f}: \text{for all } i < j, f_i > f_j \text{ if } \Delta_{i,j} < \Delta_{j,i}; f_i \neq f_j \text{ if } \Delta_{i,j} = \Delta_{j,i}; \text{ and } f_i < f_j \text{ otherwise}\}.$$

**Proof** From the definition of the conditional risk given by Eq.(2), we have

$$\begin{aligned}l(\mathbf{p}, \mathbf{f}) &= \sum_{Y \in \mathcal{Y}} p_Y L_{\text{rankloss}}(\mathbf{f}, (X, Y)) = \sum_{Y \in \mathcal{Y}} p_Y \sum_{y_i < y_j} a_Y I[f_i \geq f_j] \\ &= \sum_{Y \in \mathcal{Y}} p_Y \sum_{1 \leq i, j \leq Q} a_Y I[f_i \geq f_j] \cdot I[y_i < y_j].\end{aligned}$$

By swapping the two sums, we get

$$\begin{aligned}l(\mathbf{p}, \mathbf{f}) &= \sum_{1 \leq i, j \leq Q} I[f_i \geq f_j] \sum_{Y: y_i < y_j} p_Y a_Y = \sum_{1 \leq i, j \leq Q} I[f_i \geq f_j] \Delta_{i,j} \\ &= \sum_{1 \leq i < j \leq Q} I[f_i \geq f_j] \Delta_{i,j} + I[f_i \leq f_j] \Delta_{j,i}.\end{aligned}$$

Hence we complete the proof by combining with Property 5 in Lemma 4. ■

The following theorem discloses that none convex surrogate loss is consistent with ranking loss, and the proof is deferred to Section 6.2.

**Theorem 6** For any convex function  $\phi$ , the surrogate loss

$$\Psi(\mathbf{f}(X), Y) = \sum_{y_i < y_j} a_Y \phi(f_j(X) - f_i(X))$$

is not multi-label consistent w.r.t. ranking loss.

Intuitively, Property 5 of Lemma 4 implies that  $\{\Delta_{i,j}\}$  defines an order for the label set  $\mathcal{L} = \{1, 2, \dots, Q\}$  by  $i \succeq j$  if  $\Delta_{i,j} \leq \Delta_{j,i}$ . Notice that, for  $i \succeq j$ , there is possible that  $\Delta_{i,j} = \Delta_{j,i}$ . The set of Bayes predictors of a reasonable loss function should include all functions which are compatible with this order, i.e.,  $\mathbf{f}$ 's which enable  $f_i \geq f_j$  if  $i \succeq j$ . In the definition of ranking loss given by Eq.(3), the same penalty term is applied to  $f_i < f_j$  and  $f_i = f_j$ ; thus, the set of Bayes predictors with respect to ranking loss does not include some functions which are compatible with the above order since what enforces by ranking

loss is  $i \succ j$  or  $j \succ i$  if  $\Delta_{i,j} = \Delta_{j,i}$  for the label set  $\mathcal{L}$ . For an extreme example, i.e., when all  $\Delta_{i,j}$ 's are equal for all  $i \neq j$ , minimizing the convex surrogate loss function  $\Psi$  leads to the optimal solution  $\mathbf{f}^* \in \{\mathbf{f}: f_1 = f_2 = \dots = f_Q\}$  but  $\mathbf{f}^* \notin A(\mathbf{p})$  (from Lemma 5). So, the same penalty on  $f_i < f_j$  and  $f_i = f_j$  encumbers the multi-label consistency.

To overcome the deficiency, we instead introduce the *partial ranking loss*

$$L_{\text{p-rankloss}}(\mathbf{f}, (X, Y)) = \sum_{y_i < y_j} a_Y \left( I[f_i(X) > f_j(X)] + \frac{1}{2} I[f_i(X) = f_j(X)] \right), \quad (5)$$

which has been commonly used for ranking problems. The only difference from ranking loss lies in the use of different penalties for  $\sum_{y_i < y_j} I[f_i = f_j]$ , where the ranking loss uses  $a_Y$  while the partial ranking loss uses  $a_Y/2$ . With a proof similar to that of Lemma 5, we can get the set of Bayes predictors with respect to the partial ranking loss:

$$A(\mathbf{p}) = \{\mathbf{f}: \text{for all } i < j, f_i > f_j \text{ if } \Delta_{i,j} < \Delta_{j,i}; \text{ and } f_i < f_j \text{ if } \Delta_{i,j} > \Delta_{j,i}\}. \quad (6)$$

Now, consider the above extreme example, i.e.,  $\Delta_{i,j}$ 's are equal for all  $i \neq j$ , again. It is easy to see that by minimizing the surrogate loss function  $\Psi$ , the optimal solution  $\mathbf{f}^* \in \{\mathbf{f}: f_1 = f_2 = \dots = f_Q\} \subseteq A(\mathbf{p})$ , which exhibits multi-label consistency.

We further study more general cases, and the following theorem provides a sufficient condition for multi-label consistency w.r.t. partial ranking loss:

**Theorem 7** *The surrogate loss  $\Psi$  given by Eq.(4) is multi-label consistent w.r.t. partial ranking loss if  $\phi: \mathbb{R} \rightarrow \mathbb{R}$  is a differential and non-increasing function, and it holds that*

$$\phi'(0) < 0 \text{ and } \phi(x) + \phi(-x) \equiv 2\phi(0), \quad (7)$$

i.e.,  $\phi(x) + \phi(-x) = 2\phi(0)$  for every  $x \in \mathbb{R}$ .

The proof of Theorem 7 can be found in Section 6.3, and from this theorem we can get:

**Corollary 8** *The surrogate loss  $\Psi$  given by Eq.(4) is multi-label consistent w.r.t partial ranking loss if*

$$\phi(x) = -\arctan(x) \text{ or } \phi(x) = \frac{1 - e^{2x}}{1 + e^{2x}}.$$

Notice that Theorem 7 could not be applied directly to  $\phi(x) = -cx^{2k+1}$  for some constant  $c > 0$  and integer  $k \geq 0$ , since such setting yields that the surrogate loss  $\Psi$  is not below-bounded. This problem, however, can be solved by introducing a regularization term. With a proof similar to that of Theorem 7, we get:

**Theorem 9** *The following surrogate loss is multi-label consistent w.r.t. partial ranking loss:*

$$\Psi(\mathbf{f}(X), Y) = \sum_{y_i < y_j} a_Y \phi(f_j(X) - f_i(X)) + \tau \Upsilon(\mathbf{f}(X)),$$

where  $\tau > 0$ ,  $\phi(x) = -cx^{2k+1}$  for some constant  $c > 0$  and integer  $k \geq 0$ , and  $\Upsilon$  is symmetric, that is,  $\Upsilon(\dots, f_i(X), \dots, f_j(X), \dots) = \Upsilon(\dots, f_j(X), \dots, f_i(X), \dots)$ .

For example, we can easily construct the following convex surrogate loss

$$\Psi(\mathbf{f}(X), Y) = \sum_{y_i < y_j} -a_Y(f_j(X) - f_i(X)) + \tau \sum_{i=1}^Q f_i^2(X),$$

which is multi-label consistent w.r.t. partial ranking loss.

It is worth noting that it does not mean any convex surrogate loss  $\Psi$  given by Eq.(4) is consistent w.r.t. partial ranking loss. In fact, the following theorem proves that, many non-linear surrogate losses are inconsistent w.r.t. partial ranking loss.

**Theorem 10** *If  $\phi: \mathbb{R} \rightarrow \mathbb{R}$  is a convex, differential, non-linear and non-increasing function, the surrogate loss  $\Psi$  given by Eq.(4) is not multi-label consistent w.r.t. partial ranking loss.*

The proof is deferred to Section 6.4. The following corollary shows that some state-of-the-art multi-label learning approaches (Schapire and Singer, 2000; Dekel et al., 2004; Zhang and Zhou, 2006) are even not multi-label consistent w.r.t. partial ranking loss.

**Corollary 11** *If  $\phi(x) = e^{-x}$  or  $\phi(x) = \ln(1 + \exp(-x))$ , the surrogate loss  $\Psi$  given by Eq.(4) is not multi-label consistent w.r.t. partial ranking loss.*

In summary, the ranking loss has been suggested as a standard from early work by Schapire and Singer (2000), and many state-of-the-art multi-label learning approaches work with the formulation given by Eq.(4). However, our analysis shows that none convex surrogate loss functions are consistent w.r.t. ranking loss. Thus, ranking loss might not be a good loss function and evaluation criterion for multi-label learning. The partial ranking loss is more reasonable than ranking loss, since it enables many, though not all, convex surrogate loss functions to have consistency. In future work it would be interesting to develop some new multi-label learning approaches based on minimizing the partial ranking loss, and design other loss functions with better consistency.

## 5. Consistency w.r.t. Hamming Loss

The hamming loss concerns about how many instance-label pairs are misclassified. For a given vector  $\mathbf{f}$  and prediction function  $F$ , the hamming loss is given by

$$L_{\text{hamloss}}(F(\mathbf{f}(X)), Y) = \frac{1}{Q} \sum_{i=1}^Q I[\hat{y}_i \neq y_i],$$

where  $\hat{Y} = F(\mathbf{f}(X)) = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_Q)$ .

Hamming loss is obviously below-bounded and interval since  $0 \leq L_{\text{hamloss}}(F(\mathbf{f}(X)), Y) \leq 1$ , and for every  $X, X' \in \mathcal{X}$  and  $Y, Y' \in \mathcal{Y}$ , it holds either  $L_{\text{hamloss}}(F(\mathbf{f}(X)), Y) = L_{\text{hamloss}}(F(\mathbf{f}(X')), Y')$  or  $|L_{\text{hamloss}}(F(\mathbf{f}(X)), Y) - L_{\text{hamloss}}(F(\mathbf{f}(X')), Y')| \geq 1/Q$ . We have the conditional risk:

$$\begin{aligned} l(\mathbf{p}, F(\mathbf{f}(X))) &= \sum_{Y \in \mathcal{Y}} p_Y L_{\text{hamloss}}(\hat{Y}, Y) = \frac{1}{Q} \sum_{Y \in \mathcal{Y}} p_Y \sum_{i=1}^Q I[\hat{y}_i \neq y_i] \\ &= \frac{1}{Q} \sum_{i=1}^Q \left( \sum_{Y \in \mathcal{Y}, y_i=+1} p_Y I[\hat{y}_i \neq +1] + \sum_{Y \in \mathcal{Y}, y_i=-1} p_Y I[\hat{y}_i \neq -1] \right), \end{aligned}$$

and the set of Bayes predictors with regard to hamming loss:

$$A(\mathbf{p}) = \left\{ \mathbf{f} = \mathbf{f}(X) : \hat{Y} = F(\mathbf{f}) \text{ with } \hat{y}_i = \operatorname{sgn} \left( \sum_{Y \in \mathcal{Y}, y_i=+1} p_Y - \frac{1}{2} \right) \right\}. \quad (8)$$

A straightforward multi-label learning approach is to regard each subset of labels as a new class and then try to learn  $2^Q$  functions, i.e.,  $\mathbf{f} = (f_Y)_{Y \in \mathcal{Y}}$ . Such a prediction function is given by

$$F(\mathbf{f}(X)) = \max_{Y \in \mathcal{Y}} f_Y(X). \quad (9)$$

This approach can be viewed as a direct extension of the one-vs-all strategy for multi-class learning. We consider the following formulation:

$$\Psi(\mathbf{f}(X), Y) = \max_{\hat{Y} \neq Y} \phi(\delta(\hat{Y}, Y) + f_{\hat{Y}}(X) - f_Y(X)), \quad (10)$$

where  $\phi(x)$  and  $\delta(Y, \hat{Y})$  are set as  $\phi(x) = \max(0, x)$  and  $\delta(Y, \hat{Y}) = \sum_{i=1}^Q I[y_i \neq \hat{y}_i]$ , respectively, by Taskar et al. (2004); Hariharan et al. (2010).

Before a further discussion, we divide multi-label classification tasks into two categories, i.e., deterministic and non-deterministic, as follows:

**Definition 12** *In multi-label classification, if for every instance  $X \in \mathcal{X}$  there exists a label  $Y \in \mathcal{Y}$  such that  $P(Y|X) > 0.5$ , the task is deterministic, and non-deterministic otherwise.*

Consistency for the deterministic case is easier than non-deterministic case. For example, many formulations of SVMs are inconsistent for non-deterministic multi-class classification (Zhang, 2004a; Tewari and Bartlett, 2007), but consistent for deterministic case as indicated by Zhang (2004a). The following lemma shows that the approaches of (Taskar et al., 2004; Hariharan et al., 2010) are not multi-label consistent w.r.t. hamming loss, even for the deterministic case.

**Lemma 13** *For deterministic multi-label classification, the surrogate loss  $\Psi$  given by Eq.(10) with  $\delta(\hat{Y}, Y) = \sum_{i=1}^Q I[y_i \neq \hat{y}_i]$  is not multi-label consistent w.r.t. hamming loss.*

However, if we instead choose  $\delta(\hat{Y}, Y) = I[\hat{Y} \neq Y]$ , the following theorem guarantees that the surrogate loss is multi-label consistent w.r.t. hamming loss, at least for the deterministic case.

**Theorem 14** *For deterministic multi-label classification, the surrogate loss  $\Psi$  given by Eq.(10) with  $\delta(\hat{Y}, Y) = I[Y \neq \hat{Y}]$  is multi-label consistent w.r.t. hamming loss.*

The proofs of Lemma 13 and Theorem 14 can be found in Sections 6.5 and 6.6, respectively.

Alternatively, it is possible to transform a multi-label learning task into  $Q$  independent binary classification tasks (Boutell et al., 2004). Now the goal is to learn  $Q$  functions,  $\mathbf{f} = (f_1, f_2, \dots, f_Q)$ , and the prediction function is given by

$$F(\mathbf{f}(X)) = (\operatorname{sgn}[f_1(X)], \operatorname{sgn}[f_2(X)], \dots, \operatorname{sgn}[f_Q(X)]).$$

A common choice for the surrogate loss is

$$\Psi(\mathbf{f}(X), Y) = \sum_{i=1}^Q \phi(y_i f_i(X)), \quad (11)$$

where  $\phi$  is a convex function. For example, it was chosen as hinge loss  $\phi(t) = (1 - t)_+$  in (Elisseeff and Weston, 2002) and exponential loss  $\phi(t) = \exp(-t)$  in (Schapire and Singer, 2000). We have the conditional surrogate loss

$$\begin{aligned} W(\mathbf{p}, \mathbf{f}) &= \sum_{Y \in \mathcal{Y}} p_Y \Psi(\mathbf{f}(X), Y) = \sum_{i=1}^Q \sum_{Y \in \mathcal{Y}} p_Y \phi(y_i f_i(X)) \\ &= \sum_{i=1}^Q p_i^+ \phi(f_i(X)) + (1 - p_i^+) \phi(-f_i(X)), \end{aligned}$$

where  $p_i^+ = \sum_{Y: y_i=+1} p_Y$  and  $1 - p_i^+ = \sum_{Y: y_i=-1} p_Y$ . For simplicity, we denote by

$$W_i(p_i^+, f_i) = p_i^+ \phi(f_i) + (1 - p_i^+) \phi(-f_i).$$

This yields that minimizing  $W(\mathbf{p}, \mathbf{f})$  is equivalent to minimizing  $W_i(p_i^+, f_i)$  for all  $1 \leq i \leq Q$ , that is

$$W^*(\mathbf{p}) = \inf_{\mathbf{f}} W(\mathbf{p}, \mathbf{f}) = \sum_{i=1}^Q \inf_{f_i} W_i(p_i^+, f_i).$$

The consistency for binary classification has been well-studied from (Zhang, 2004b; Bartlett et al., 2006), and based on the work of Bartlett et al. (2006), we can easily get:

**Theorem 15** *The surrogate loss  $\Psi$  given by Eq.(11) is consistent w.r.t. hamming loss for convex function  $\phi$  with  $\phi'(0) < 0$ .*

It is evident from this theorem that the surrogate loss  $\Psi$  given by Eq.(11) is consistent w.r.t. hamming loss if  $\phi$  is any of the following:

- Exponential:  $\phi(x) = e^{-x}$ ;
- Hinge:  $\phi(x) = \max(0, 1 - x)$ ;
- Least squares:  $\phi(x) = (1 - x)^2$ ;
- Logistic regression:  $\phi(x) = \ln(1 + \exp(-x))$ .

Notice that in this paper, we do not consider how to learn the real-valued functions  $\mathbf{f}$  for small data or large number of labels, where many labels or label subsets lack enough training examples, while this is a challenging task and the exploitation of label correlation is therefore needed. In our analysis, we assume sufficient training data and ignore the label correlation. Moreover, we do not consider how to decide the number of relevant labels for ranking loss and partial ranking loss, while in practice this is very challenging. These are important future issues.

## 6. Proofs

### 6.1. Proof of Theorem 3

We first introduce some useful lemmas:

**Lemma 16**  $W^*(\mathbf{p})$  is continuous on  $\Lambda$ .

**Proof** From the Heine definition of continuity, we need to show that  $W^*(\mathbf{p}^{(n)}) \rightarrow W^*(\mathbf{p})$  for any sequence  $\mathbf{p}^{(n)} \rightarrow \mathbf{p}$ .

Let  $B_r$  be a closed ball with radius  $r$  in  $\mathbb{R}^K$ . Since  $|\mathcal{Y}|$  is finite, we have

$$\sum_{Y \in \mathcal{Y}} p_Y^{(n)} \Psi(\mathbf{f}, Y) \rightarrow \sum_{Y \in \mathcal{Y}} p_Y \Psi(\mathbf{f}, Y)$$

uniformly for every  $\mathbf{f} \in B_r$  and every sequence  $\mathbf{p}^{(n)} \rightarrow \mathbf{p}$ , leading to

$$\inf_{\mathbf{f} \in B_r} \sum_{Y \in \mathcal{Y}} p_Y^{(n)} \Psi(\mathbf{f}, Y) \rightarrow \inf_{\mathbf{f} \in B_r} \sum_{Y \in \mathcal{Y}} p_Y \Psi(\mathbf{f}, Y).$$

From the fact  $W^*(\mathbf{p}^{(n)}) \leq \inf_{\mathbf{f} \in B_r} \sum_{Y \in \mathcal{Y}} p_Y^{(n)} \Psi(\mathbf{f}, Y)$ , by letting  $r \rightarrow \infty$ , we have:

$$\limsup_{n \rightarrow \infty} W^*(\mathbf{p}^{(n)}) \leq W^*(\mathbf{p}). \quad (12)$$

Denote by  $\mathcal{Y}' = \{Y | p_Y > 0 \text{ for } Y \in \mathcal{Y}\}$  and assume  $\Psi(\cdot, \cdot) \geq C$  for some constant  $C$  (since  $\Psi$  is below-bounded). We have  $W^*(\mathbf{p}^{(n)}) \geq \inf_{\mathbf{f}} \sum_{Y \in \mathcal{Y}'} p_Y^{(n)} \Psi(\mathbf{f}, Y) + C \sum_{Y \in \mathcal{Y}/\mathcal{Y}'} p_Y^{(n)}$ , which yields

$$\liminf_{n \rightarrow \infty} W^*(\mathbf{p}^{(n)}) \geq \liminf_{n \rightarrow \infty} \left( \inf_{\mathbf{f}} \sum_{Y \in \mathcal{Y}'} p_Y^{(n)} \Psi(\mathbf{f}, Y) + C \sum_{Y \in \mathcal{Y}/\mathcal{Y}'} p_Y^{(n)} \right) = W^*(\mathbf{p}),$$

which completes the proof by combining Eq.(12). ■

**Lemma 17** If the surrogate loss function  $\Psi$  is multi-label consistent w.r.t. loss function  $L$ , then for any  $\epsilon > 0$  there exists  $\delta > 0$  such that, for every  $\mathbf{p} \in \Lambda$ ,  $l(\mathbf{p}, \mathbf{f}) - \inf_{\mathbf{f}'} l(\mathbf{p}, \mathbf{f}') \geq \epsilon$  implies  $W(\mathbf{p}, \mathbf{f}) - W^*(\mathbf{p}) \geq \delta$ .

**Proof** We proceed by contradiction. Suppose  $\Psi$  is multi-label consistent and there exists  $\epsilon > 0$  and a sequence  $(\mathbf{p}^{(n)}, \mathbf{f}^{(n)})$  such that  $l(\mathbf{p}^{(n)}, \mathbf{f}^{(n)}) - \inf_{\mathbf{f}'} l(\mathbf{p}^{(n)}, \mathbf{f}') \geq \epsilon$  and  $W(\mathbf{p}^{(n)}, \mathbf{f}^{(n)}) \rightarrow W^*(\mathbf{p}^{(n)})$ . From the compactness of  $\Lambda$ , there exists a convergence sequence  $n_k$  such that  $\mathbf{p}^{(n_k)} \rightarrow \mathbf{p}$  for some  $\mathbf{p} \in \Lambda$ . From Lemma 16, we have

$$W(\mathbf{p}^{(n_k)}, \mathbf{f}^{(n_k)}) \rightarrow W^*(\mathbf{p}).$$

Similar to the proof of Lemma 16, denoted by  $\mathcal{Y}' = \{Y | p_Y > 0 \text{ for } Y \in \mathcal{Y}\}$ , we have

$$\begin{aligned} \limsup_{n_k} W(\mathbf{p}, \mathbf{f}^{(n_k)}) &= \limsup_{n_k} \left( C \sum_{Y \in \mathcal{Y}/\mathcal{Y}'} p_Y^{(n_k)} + \inf_{\mathbf{f}} \sum_{Y \in \mathcal{Y}'} p_Y^{(n_k)} \Psi(\mathbf{f}^{(n_k)}, Y) \right) \\ &\leq \lim_{n_k} W(\mathbf{p}^{(n_k)}, \mathbf{f}^{(n_k)}) = W^*(\mathbf{p}). \end{aligned}$$

This gives  $W(\mathbf{p}, \mathbf{f}^{(n_k)}) \rightarrow W^*(\mathbf{p})$  from the definition of  $W^*(\mathbf{p})$ . Since  $\Psi$  is multi-label consistent, there exists a sequence  $\mathbf{f}^{(n_{k_i})}$  satisfying  $l(\mathbf{p}, \mathbf{f}^{(n_{k_i})}) \rightarrow \inf_{\mathbf{f}'} l(\mathbf{p}, \mathbf{f}')$ , which contradicts the assumption  $l(\mathbf{p}^{(n)}, \mathbf{f}^{(n)}) - \inf_{\mathbf{f}'} l(\mathbf{p}^{(n)}, \mathbf{f}') \geq \epsilon$ , and thus the lemma holds. ■

### Proof of Theorem 3:

(“ $\Rightarrow$ ”) We first introduce a new notation

$$H(\epsilon) = \inf_{\mathbf{p} \in \Lambda, \mathbf{f}} \{W(\mathbf{p}, \mathbf{f}) - \inf_{\mathbf{f}'} W(\mathbf{p}, \mathbf{f}') : l(\mathbf{p}, \mathbf{f}) - \inf_{\mathbf{f}'} l(\mathbf{p}, \mathbf{f}') \geq \epsilon\}.$$

It is obvious that  $H(0) = 0$  and  $H(\epsilon) > 0$  for  $\epsilon > 0$  from Lemma 17. Corollary 26 of (Zhang, 2004a) guarantees there exists a concave function  $\eta$  on  $[0, \infty]$  such that  $\eta(0) = 0$  and  $\eta(\epsilon) \rightarrow 0$  as  $\epsilon \rightarrow 0$  and

$$R(f) - R^* \leq \eta(R_\Psi(f) - R_\Psi^*).$$

Thus, if  $R_\Psi(f) \rightarrow R_\Psi^*$  then  $R(f) \rightarrow R^*$ .

(“ $\Leftarrow$ ”) We proceed by contradiction. Suppose  $\Psi$  is not multi-label consistent, and thus there exists some  $\mathbf{p}$  such that  $W^*(\mathbf{p}) = \inf_{\mathbf{f}} \{W(\mathbf{p}, \mathbf{f}) : \mathbf{f} \notin A(\mathbf{p})\}$ . Let  $\mathbf{f}^{(n)} \notin A(\mathbf{p})$  be a sequence such that  $W(\mathbf{p}, \mathbf{f}^{(n)}) \rightarrow W^*(\mathbf{p})$ . For simplicity, we consider  $\mathcal{X} = \{x\}$ , i.e., only one instance, and set  $\mathbf{f}_n(x) = \mathbf{f}^{(n)}$ . Then

$$R_\Psi(\mathbf{f}_n) = W(\mathbf{p}, \mathbf{f}^{(n)}) \rightarrow W^*(\mathbf{p}) = R_\Psi^*,$$

yielding  $l(\mathbf{p}, \mathbf{f}^{(n)}) \rightarrow \inf_{\mathbf{f}} l(\mathbf{p}, \mathbf{f})$  which is contrary to

$$l(\mathbf{p}, \mathbf{f}^{(n)}) \geq \inf_{\mathbf{f}} l(\mathbf{p}, \mathbf{f}) + \gamma(\mathbf{p})$$

where  $\gamma(\mathbf{p}) = \inf_{\mathbf{f} \notin A(\mathbf{p})} l(\mathbf{p}, \mathbf{f}) - \inf_{\mathbf{f}} l(\mathbf{p}, \mathbf{f}) > 0$ , since  $\mathbf{f}^{(n)} \notin A(\mathbf{p})$  and  $L$  is interval. Thus we complete the proof. ■

### 6.2. Proof of Theorem 6

We proceed by contradiction. Suppose that the surrogate loss  $\Psi$  is multi-label consistent with ranking loss. We have

$$\begin{aligned} W(\mathbf{p}, \mathbf{f}) &= \sum_{Y \in \mathcal{Y}} p_Y \Psi(\mathbf{f}, Y) = \sum_{Y \in \mathcal{Y}} p_Y a_Y \sum_{y_i < y_j} \phi(f_j - f_i) \\ &= \sum_{1 \leq i < j \leq Q} \phi(f_j - f_i) \Delta_{i,j} + \phi(f_i - f_j) \Delta_{j,i}. \end{aligned}$$

Consider the probability vector  $\mathbf{p} = (p_Y)_{Y \in \mathcal{Y}}$  and penalty vector  $(a_Y)_{Y \in \mathcal{Y}}$  s.t.  $P_{Y_1} = P_{Y_2}$  and  $a_{Y_1} = a_{Y_2}$  for every  $Y_1 \neq Y_2$ ,  $Y_1, Y_2 \in \mathcal{Y}$ . This yields that  $\Delta_{i,j} = \Delta_{m,n}$  for every  $1 \leq i \neq j, m \neq n \leq Q$ , and thus we get

$$W(\mathbf{p}, \mathbf{f}) = \Delta_{1,2} \sum_{1 \leq i < j \leq Q} \phi(f_j - f_i) + \phi(f_i - f_j).$$

From the convexity of  $\phi$ , minimizing  $W(\mathbf{p}, \mathbf{f})$  gives

$$W^*(\mathbf{p}) = W(\mathbf{p}, \hat{\mathbf{f}}) = \Delta_{1,2} \sum_{1 \leq i < j \leq Q} 2\phi(0) = Q(Q-1)\phi(0)\Delta_{1,2},$$

where  $\hat{\mathbf{f}} = \{\hat{\mathbf{f}}: \hat{f}_1 = \hat{f}_2 = \dots = \hat{f}_Q\}$ . Notice that  $\hat{\mathbf{f}} \notin A(\mathbf{p})$  from Lemma 5, and we have

$$W^*(\mathbf{p}) = \inf_{\mathbf{f}} \{W(\mathbf{p}, \mathbf{f}): \mathbf{f} \notin A(\mathbf{p})\},$$

which completes the proof.  $\blacksquare$

### 6.3. Proof of Theorem 7

For any  $\mathbf{p} \in \Lambda$  and non-negative vector  $(ay)_{Y \in \mathcal{Y}}$ , we will prove that,  $f_i > f_j$  if  $\Delta_{i,j} < \Delta_{j,i}$  and  $f_i < f_j$  if  $\Delta_{i,j} > \Delta_{j,i}$ , for all  $\mathbf{f}$  satisfying  $W^*(\mathbf{p}) = W(\mathbf{p}, \mathbf{f})$ . Without loss of generality, it is sufficient to prove that  $f_1 > f_2$  if  $\Delta_{1,2} < \Delta_{2,1}$ .

We proceed by contradiction, and assume that there exists a vector  $\mathbf{f}$  such that  $f_1 \leq f_2$  and  $W^*(\mathbf{p}) = W(\mathbf{p}, \mathbf{f})$ .

For the case  $f_1 < f_2$ , we can introduce another vector  $\mathbf{f}'$  with  $f'_1 = f_2$ ,  $f'_2 = f_1$  and  $f'_k = f_k$  for  $k \neq 1, 2$ . From the definition of conditional surrogate risk, we have

$$W(\mathbf{p}, \mathbf{f}) = \sum_{1 \leq i < j \leq Q} \phi(f_j - f_i) \Delta_{i,j} + \phi(f_i - f_j) \Delta_{j,i},$$

which yields that

$$\begin{aligned} W(\mathbf{p}, \mathbf{f}) - W(\mathbf{p}, \mathbf{f}') &= (\Delta_{1,2} - \Delta_{2,1})(\phi(f_2 - f_1) - \phi(f_1 - f_2)) \\ &\quad + \sum_{i=3}^Q (\Delta_{1,i} - \Delta_{2,i})(\phi(f_i - f_1) - \phi(f_i - f_2)) + \sum_{i=3}^Q (\Delta_{i,1} - \Delta_{i,2})(\phi(f_1 - f_i) - \phi(f_2 - f_i)). \end{aligned}$$

From Property 4 of Lemma 4, we have

$$\Delta_{1,i} - \Delta_{2,i} - \Delta_{i,1} + \Delta_{i,2} = \Delta_{1,2} - \Delta_{2,1}. \quad (13)$$

It follows that

$$\begin{aligned} W(\mathbf{p}, \mathbf{f}) - W(\mathbf{p}, \mathbf{f}') &= (\Delta_{1,2} - \Delta_{2,1}) \left( \phi(f_2 - f_1) - \phi(f_1 - f_2) + \sum_{i=3}^Q (\phi(f_i - f_1) - \phi(f_i - f_2)) \right) \\ &\quad + \sum_{i=3}^Q (\Delta_{i,1} - \Delta_{i,2})(\phi(f_1 - f_i) + \phi(f_i - f_1) - \phi(f_i - f_2) - \phi(f_2 - f_i)) \\ &= (\Delta_{1,2} - \Delta_{2,1})(\phi(f_2 - f_1) - \phi(f_1 - f_2)) + (\Delta_{1,2} - \Delta_{2,1}) \sum_{i=3}^Q (\phi(f_i - f_1) - \phi(f_i - f_2)) \end{aligned}$$

where the last equality holds by the condition  $\phi(x) + \phi(-x) \equiv 2\phi(0)$ . For non-increasing function  $\phi$  with  $\phi'(0) < 0$ , we have  $\phi(z) < \phi(-z)$  for all  $z > 0$ , and  $\phi(f_i - f_1) \leq \phi(f_i - f_2)$ . From  $\Delta_{1,2} < \Delta_{2,1}$  we get  $W(\mathbf{p}, \mathbf{f}) > W(\mathbf{p}, \mathbf{f}')$ , which is contrary to  $W(\mathbf{p}, \mathbf{f}) = W^*(\mathbf{p})$ .

We now consider the case  $f_1 = f_2$ . For the optimal solution, we have the first-order condition  $\frac{\partial}{\partial f_i} W(\mathbf{p}, \mathbf{f}) = 0$  for  $i = 1, 2$ :

$$\begin{aligned} \sum_{i \neq 1} \phi'(f_1 - f_i) \Delta_{i,1} &= \sum_{i \neq 1} \phi'(f_1 - f_i) \Delta_{1,i}, \\ \sum_{i \neq 2} \phi'(f_1 - f_i) \Delta_{2,i} &= \sum_{i \neq 2} \phi'(f_1 - f_i) \Delta_{i,2}. \end{aligned}$$

By combining Eq.(13),  $f_1 = f_2$  and  $\phi'(x) = \phi'(-x)$  from Eq.(7), we have:

$$(\Delta_{2,1} - \Delta_{1,2}) \left( 2\phi'(0) + \sum_{i \neq 1,2} \phi'(f_1 - f_i) \right) = 0,$$

which is impossible since  $\Delta_{1,2} < \Delta_{2,1}$  and  $2\phi'(0) + \sum_{i \neq 1,2} \phi'(f_1 - f_i) \leq 2\phi'(0) < 0$ . Thus, we complete the proof.  $\blacksquare$

#### 6.4. Proof of Theorem 10

For convex function  $\phi$  we have  $\phi'(x) \leq \phi'(y)$  for every  $x \leq y$  from (Rockafellar, 1997), and the derivative function  $\phi'(x)$  is continuous for  $x \in \mathbb{R}$  if  $\phi$  is differential and convex. Since  $\phi$  is non-increasing, we have  $\phi'(x) \leq 0$  for all  $x \in \mathbb{R}$ , and without loss of generality, we assume  $\phi'(x) < 0$ .

We proceed by contradiction. Assume the surrogate loss  $\Psi$  is multi-label consistent with partial ranking loss for some non-linear function  $\phi$ . Then, from the continuity of  $\phi'(x)$ , there exists an interval  $(c, d)$  for  $c < d < 0$  or  $0 < c < d$ , such that

$$\phi'(x) < \phi'(y) \text{ for every } x < y, x, y \in (c, d).$$

In the following, we focus on the case  $0 < c < d$ , and similar consideration could be made for the case  $c < d < 0$ .

We first fix  $a \in (c, d)$  and introduce a new function

$$G(x) = (\phi'(x - a) - \phi'(a - x))(\phi'(a) + \phi'(x)) + \phi'(x)(\phi'(-a) - \phi'(a)).$$

It is easy to find that  $G(a) = \phi'(a)(\phi'(-a) - \phi'(a)) > 0$  and  $G(x)$  is continuous. Thus there exists  $b > a$  and  $b \in (c, d)$  such that

$$G(b) = (\phi'(b - a) - \phi'(a - b))(\phi'(a) + \phi'(b)) + \phi'(b)(\phi'(-a) - \phi'(a)) > 0,$$

which gives

$$\frac{\phi'(b - a)(\phi'(a) + \phi'(b)) + \phi'(b)\phi'(-a)}{\phi'(a - b)(\phi'(a) + \phi'(b)) + \phi'(b)\phi'(a)} > 1. \quad (14)$$

Moreover, from  $\phi'(a) < \phi'(b) < 0$  and  $\phi'(-b) \leq \phi'(-a) < 0$ , we have

$$0 < \frac{\phi'(-a)}{\phi'(a)} \leq \frac{\phi'(-b)}{\phi'(a)} < \frac{\phi'(-b)}{\phi'(b)}. \quad (15)$$

We consider the following multi-label classification with  $Q = 3$  labels:

$$Y_1 = (-1, +1, +1), Y_2 = (+1, -1, -1), Y_3 = (+1, +1, -1), Y_4 = (-1, -1, +1).$$

Let  $\mathbf{f} = (f_1, f_2, f_3)$  such that  $a = f_3 - f_1$  and  $b = f_3 - f_2$ , and thus  $f_1 > f_2$ . For every probability vector  $\mathbf{p} = (p_{Y_1}, p_{Y_2}, p_{Y_3}, p_{Y_4}) \in \Lambda$  and every penalty vector  $(a_{Y_1}, a_{Y_2}, a_{Y_3}, a_{Y_4})$ , we have

$$\begin{aligned} W(\mathbf{p}, \mathbf{f}) = & \Delta_{1,2}\phi(f_2 - f_1) + \Delta_{2,1}\phi(f_1 - f_2) + \Delta_{1,3}\phi(f_3 - f_1) + \Delta_{3,1}\phi(f_1 - f_3) \\ & + \Delta_{2,3}\phi(f_3 - f_2) + \Delta_{3,2}\phi(f_2 - f_3), \end{aligned}$$

where  $\Delta_{1,2} = P_1$ ,  $\Delta_{2,1} = P_2$ ,  $\Delta_{1,3} = P_1 + P_4$ ,  $\Delta_{3,1} = P_2 + P_3$ ,  $\Delta_{2,3} = P_4$  and  $\Delta_{3,2} = P_3$  with  $P_i = p_{Y_1}a_{Y_1}$  for  $i = 1, 2, 3, 4$ . In the following, we will construct some  $\bar{\mathbf{p}}$  and  $\bar{\mathbf{a}}$  such that  $W^*(\bar{\mathbf{p}}) = W(\bar{\mathbf{p}}, \mathbf{f})$  and  $\Delta_{1,2} > \Delta_{2,1}$ .

The subgradient condition for optimality of  $W(\mathbf{p}, \mathbf{f})$  gives that

$$\partial W(\mathbf{p}, \mathbf{f}) / \partial f_1 = -P_1\phi'(a - b) + P_2\phi'(b - a) - (P_1 + P_4)\phi'(a) + (P_2 + P_3)\phi'(-a) = 0,$$

$$\partial W(\mathbf{p}, \mathbf{f}) / \partial f_2 = P_1\phi'(a - b) - P_2\phi'(b - a) - P_4\phi'(b) + P_3\phi'(-b) = 0,$$

$$\partial W(\mathbf{p}, \mathbf{f}) / \partial f_3 = (P_1 + P_4)\phi'(a) - (P_2 + P_3)\phi'(-a) + P_4\phi'(b) - P_3\phi'(-b) = 0,$$

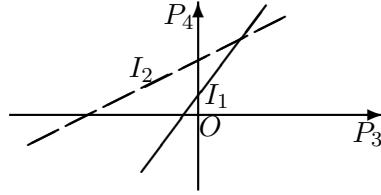


Figure 1: Lines  $I_1$  (solid) and  $I_2$  (dash) corresponding to Eqs. (16) and (17), respectively.

equivalent to

$$P_1\phi'(a-b) - P_2\phi'(b-a) = P_4\phi'(b) - P_3\phi'(-b), \quad (16)$$

$$-P_1\phi'(a) + P_2\phi'(-a) = P_4(\phi'(a) + \phi'(b)) - P_3(\phi'(-a) + \phi'(-b)). \quad (17)$$

From Lemma 18, there exists  $\bar{\mathbf{p}} = (\bar{p}_{Y_1}, \bar{p}_{Y_2}, \bar{p}_{Y_3}, \bar{p}_{Y_4})$  and  $(\bar{a}_{Y_1}, \bar{a}_{Y_2}, \bar{a}_{Y_3}, \bar{a}_{Y_4})$  satisfying Eqs. (16), (17) and  $P_1 > P_2$ . Hence this yields  $W(\bar{\mathbf{p}}, \mathbf{f}) = W^*(\bar{\mathbf{p}})$ . Notice that  $\mathbf{f} \notin A(\bar{\mathbf{p}})$  from Eq.(6), since  $\Delta_{1,2} = P_1 > P_2 = \Delta_{2,1}$  and  $f_1 > f_2$ , we have

$$W^*(\bar{\mathbf{p}}) = \inf_{\mathbf{f}} \{W(\bar{\mathbf{p}}, \mathbf{f}) : \mathbf{f} \notin A(\bar{\mathbf{p}})\},$$

which completes the proof. ■

**Lemma 18** *There exist some  $P_1 > P_2 > 0$ ,  $P_3 > 0$  and  $P_4 > 0$  satisfying Eqs. (16) and (17), if Eqs. (14) and (15) hold for some  $0 < a < b$ .*

**Proof** From Eq.(14), we set

$$1 < \frac{P_1}{P_2} < \frac{\phi'(b-a)(\phi'(a) + \phi'(b)) + \phi'(b)\phi'(-a)}{\phi'(a-b)(\phi'(a) + \phi'(b)) + \phi'(b)\phi'(a)}. \quad (18)$$

For  $a < b$ , we have  $\phi'(a-b) \leq \phi'(b-a) < 0$ , yielding

$$\frac{P_1}{P_2} > 1 \geq \frac{\phi'(b-a)}{\phi'(a-b)},$$

which gives  $P_1\phi'(a-b) - P_2\phi'(b-a) < 0$ . Thus, Eq.(16) corresponds to the Line  $I_1$  in Figure 1. From Eq.(15), we further obtain

$$0 < \frac{\phi'(-a) + \phi'(-b)}{\phi'(a) + \phi'(b)} < \frac{\phi'(-b)}{\phi'(b)}.$$

To guarantee  $P_3 > 0$  and  $P_4 > 0$  satisfying Eqs. (16) and (17), as shown in Figure 1, we need:

$$\frac{P_1\phi'(a-b) - P_2\phi'(b-a)}{\phi'(b)} < \frac{-P_1\phi'(a) + P_2\phi'(-a)}{\phi'(a) + \phi'(b)}.$$

The above holds obviously from Eq.(18). Thus, we complete the proof. ■

### 6.5. Proof of Lemma 13

We consider a multi-label classification task with  $Q = 2$  labels, i.e.,  $\mathcal{L} = \{1, 2\}$  and let  $Y_1 = (-1, -1)$ ,  $Y_2 = (-1, 1)$ ,  $Y_3 = (1, 1)$  and  $Y_4 = (1, -1)$ . Suppose  $p_{Y_4} = 0$  and  $p_{Y_2} + p_{Y_3} < p_{Y_1} < 2p_{Y_2} + p_{Y_3}$ . It is evident that  $p_{Y_1} > 0.5$  and thus, this multi-label classification task is deterministic. It is necessary to consider  $\mathbf{f} = (f_{Y_1}, f_{Y_2}, f_{Y_3})$ . By combining Eqs. (8) and (10), we get

$$A(\mathbf{p}) = \{\mathbf{f}: f_{Y_1} \geq f_{Y_2} \text{ and } f_{Y_1} \geq f_{Y_3}\}.$$

We also have

$$\begin{aligned} W(\mathbf{p}, \mathbf{f}) &= p_{Y_1} \max\{\phi(1 + f_{Y_2} - f_{Y_1}), \phi(2 + f_{Y_3} - f_{Y_1})\} + p_{Y_2} \max\{\phi(1 + f_{Y_1} - f_{Y_2}), \\ &\quad \phi(1 + f_{Y_3} - f_{Y_2})\} + p_{Y_3} \max\{\phi(2 + f_{Y_1} - f_{Y_3}), \phi(1 + f_{Y_2} - f_{Y_3})\} \end{aligned}$$

and

$$W^*(\mathbf{p}) = W(\mathbf{p}, \mathbf{f}^*) = \inf_{\mathbf{f}} W(\mathbf{p}, \mathbf{f}) = 2p_{Y_1} + 2p_{Y_3},$$

where  $\mathbf{f}^* = (f_{Y_1}^*, f_{Y_2}^*, f_{Y_3}^*)$  such that  $f_{Y_1}^* = f_{Y_3}^*$  and  $f_{Y_1}^* = f_{Y_2}^* - 1$ . Hence  $\mathbf{f} \notin A(\mathbf{p})$  and it is not multi-label consistent.  $\blacksquare$

### 6.6. Proof of Theorem 14

We first introduce a useful lemma:

**Lemma 19** *For the surrogate loss  $\Psi$  given by Eq.(10) with  $\delta(\hat{Y}, Y) = I[\hat{Y} \neq Y]$ , and for every  $\mathbf{p} \in \Lambda$  and  $\mathbf{f}$  such that  $W(\mathbf{p}, \mathbf{f}) = W^*(\mathbf{p})$ , we have  $f_{Y_1} \geq f_{Y_2}$  if  $p_{Y_1} > p_{Y_2}$ .*

**Proof** We proceed by contradiction. Assume there exist some  $\mathbf{p} \in \Lambda$  and  $\mathbf{f}$  such that  $f_{Y_1} < f_{Y_2}$ ,  $p_{Y_1} > p_{Y_2}$  and  $W^*(\mathbf{p}) = W(\mathbf{p}, \mathbf{f})$ . We define a new real-valued vector  $\mathbf{f}'$  such that  $f'_{Y_1} = f_{Y_2}$ ,  $f'_{Y_2} = f_{Y_1}$  and  $f'_{Y_i} = f_{Y_i}$  for  $i \neq 1, 2$ . This follows that

$$W(\mathbf{p}, \mathbf{f}) - W(\mathbf{p}, \mathbf{f}') = (p_{Y_1} - p_{Y_2})(\max_{Y \neq Y_1} \phi(1 + f_Y(X) - f_{Y_1}(X)) - \max_{Y \neq Y_2} \phi(1 + f_Y(X) - f_{Y_2}(X))).$$

From the assumption  $f_{Y_1} < f_{Y_2}$ , we have

$$\max_{Y \neq Y_1} \phi(1 + f_Y(X) - f_{Y_1}(X)) > \max_{Y \neq Y_2} \phi(1 + f_Y(X) - f_{Y_2}(X)),$$

leading to  $W(\mathbf{p}, \mathbf{f}) - W(\mathbf{p}, \mathbf{f}') > 0$ , which is contrary to the assumption  $W^*(\mathbf{p}) = W(\mathbf{p}, \mathbf{f})$ . Thus, we complete the proof.  $\blacksquare$

**Proof of Theorem 14:** Without loss of generality, we consider a probability vector  $\mathbf{p} = (p_Y)_{Y \in \mathcal{Y}}$  satisfying  $p_{Y_1} > p_{Y_2} \geq p_{Y_3} \dots \geq p_{Y_{2Q}}$  and  $p_{Y_1} > 0.5$ . It is easy to get

$$A(\mathbf{p}) = \{\mathbf{f}: f_{Y_1} > f_Y \text{ for } Y \neq Y_1\},$$

from Eqs. (8) and (9). On the other hand, for each  $\mathbf{f}$  satisfying  $W(\mathbf{p}, \mathbf{f}) = W^*(\mathbf{p})$ , we have

$$f_{Y_1} \geq f_{Y_2} \geq \dots \geq f_{Y_{2Q}}$$

from Lemma 19. Thus, from the fact that

$$W(\mathbf{p}, \mathbf{f}) = p_{Y_1} \phi(1 + f_{Y_2} - f_{Y_1}) + \sum_{Y \neq Y_1} p_Y \phi(1 + f_{Y_1} - f_Y).$$

and  $p_{Y_1} > \sum_{Y \neq Y_1} p_Y$ , we complete the proof.  $\blacksquare$

## Acknowledgments

The authors want to thank the anonymous reviewers for their helpful comments and suggestions, and Tong Zhang for reading a preliminary version. This research was supported by the National Fundamental Research Program of China (2010CB327903), the National Science Foundation of China (61073097) and the Jiangsu Science Foundation (BK2008018).

## References

- P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.
- M. R. Boutell, J. Luo, X. Shen, and C.M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.
- D. Cossack and T. Zhang. Statistical analysis of bayes optimal subset ranking. *IEEE Transactions on Information Theory*, 54(11):5140–5154, 2008.
- O. Dekel, C. Manning, and Y. Singer. Log-linear models for label ranking. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
- K. Dembczyński, W. W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on Machine Learning*, pages 279–286, Haifa, Israel, 2010.
- J. C. Duchi, L. W. Mackey, and M. I. Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th International Conference on Machine Learning*, pages 327–334, Haifa, Israel, 2010.
- A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 681–687. MIT Press, Cambridge, MA, 2002.
- N. Ghamrawi and A. McCallum. Collective multi-label classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 195–200, Bremen, Germany, 2005.
- S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 22–30, Sydney, Australia, 2004.
- B. Hariharan, L. Zelnik-Manor, S. Vishwanathan, and M. Varma. Large scale max-margin multi-label classification with priors. In *Proceedings of the 27th International Conference on Machine Learning*, pages 423–430, Haifa, Israel, 2010.
- D. Hsu, S. M. Kakade, J. Langford, and T. Zhang. Multi-label prediction via compressed sensing. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 772–780. MIT Press, Cambridge, MA, 2009.

- Y. Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):259–275, 2002.
- J. Petterson and T. Caetano. Reverse multi-label learning. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1912–1920. MIT Press, Cambridge, MA, 2010.
- G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *Proceedings of the 15th ACM International Conference on Multimedia*, pages 17–26, Augsburg, Germany, 2007.
- R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1997.
- R. E. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2):135–168, 2000.
- I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.
- B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
- A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.
- F. Xia, T. Y. Liu, J. Wang, W. Zhang, and H. Li. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1192–1199, Helsinki, Finland, 2008.
- M.-L. Zhang and Z.-H. Zhou. Multi-label neural network with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.
- M.-L. Zhang and Z.-H. Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.
- T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004a.
- T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85, 2004b.
- Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19*, pages 1609–1616. MIT Press, Cambridge, MA, 2007.

# The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond

**Aurélien Garivier** and **Olivier Cappé**  
*LTCI, CNRS & Telecom ParisTech, Paris, France*

GARIVIER,CAPPE@TELECOM-PARISTECH.FR

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

This paper presents a finite-time analysis of the KL-UCB algorithm, an online, horizon-free index policy for stochastic bandit problems. We prove two distinct results: first, for arbitrary bounded rewards, the KL-UCB algorithm satisfies a uniformly better regret bound than UCB and its variants; second, in the special case of Bernoulli rewards, it reaches the lower bound of Lai and Robbins. Furthermore, we show that simple adaptations of the KL-UCB algorithm are also optimal for specific classes of (possibly unbounded) rewards, including those generated from exponential families of distributions. A large-scale numerical study comparing KL-UCB with its main competitors (UCB, MOSS, UCB-Tuned, UCB-V, DMED) shows that KL-UCB is remarkably efficient and stable, including for short time horizons. KL-UCB is also the only method that always performs better than the basic UCB policy. Our regret bounds rely on deviations results of independent interest which are stated and proved in the Appendix. As a by-product, we also obtain an improved regret bound for the standard UCB algorithm.

**Keywords:** List of keywords

## 1. Introduction

The multi-armed bandit problem is a simple, archetypal setting of reinforcement learning, where an agent facing a slot machine with several arms tries to maximize her profit by a proper choice of arm draws. In the stochastic<sup>1</sup> bandit problem, the agent sequentially chooses, for  $t = 1, 2, \dots, n$ , an arm  $A_t \in \{1, \dots, K\}$ , and receives a reward  $X_t$  such that, conditionally on the arm choices  $A_1, A_2, \dots$ , the rewards are independent and identically distributed, with expectation  $\mu_{A_1}, \mu_{A_2}, \dots$ . Her *policy* is the (possibly randomized) decision rule that, to every past observations  $(A_1, X_1, \dots, A_{t-1}, X_{t-1})$ , associates her next choice  $A_t$ . The best choice is any arm  $a^*$  with maximal expected reward  $\mu_{a^*}$ . The performance of her policy can be measured by the *regret*  $R_n$ , defined as the difference between the rewards she accumulates up to the horizon  $t = n$ , and the rewards that she would have accumulated during the same period, had she known from the beginning which arm had the highest expected reward.

The agent typically faces an “exploration versus exploitation dilemma” : at time  $t$ , she can take advantage of the information she has gathered, by choosing the so-far best performing arm, but she has to consider the possibility that the other arms are actually

---

1. Another interesting setting is the *adversarial* bandit problem, where the rewards are not stochastic but chosen by an opponent - this setting is not the subject of this paper.

under-rated and she must play sufficiently often all of them. Since the works of Gittins (1979) in the 1970s, this problem raised much interest and several variants, solutions and extensions have been proposed, see Even-Dar et al. (2002) and references therein.

Two families of bandit settings can be distinguished: in the first family, the distribution of  $X_t$  given  $A_t = a$  is assumed to belong to a family  $\{p_\theta, \theta \in \Theta_a\}$  of probability distributions. In a particular parametric framework, Lai and Robbins (1985) proved a lower-bound on the performance of any policy, and determined optimal policies. This framework was extended to multi-parameter models by Burnetas and Katehakis (1997) who showed that the number of draws up to time  $n$ ,  $N_a(n)$ , of any sub-optimal arm  $a$  is lower-bounded by

$$N_a(n) \geq \left( \inf_{\theta \in \Theta_a : E[p_\theta] > \mu_{a^*}} \frac{1}{KL(p_{\theta_a}, p_\theta)} + o(1) \right) \log(n), \quad (1)$$

where  $KL$  denotes the Kullback-Leibler divergence and  $E(p_\theta)$  is the expectation under  $p_\theta$ ; hence, the regret is lower-bounded as follows:

$$\liminf_{n \rightarrow \infty} \frac{\mathbb{E}[R_n]}{\log(n)} \geq \sum_{a: \mu_a < \mu_{a^*}} \inf_{\theta \in \Theta_a : E[p_\theta] > \mu_{a^*}} \frac{\mu_{a^*} - \mu_a}{KL(p_{\theta_a}, p_\theta)}. \quad (2)$$

Recently, Honda and Takemura (2010) proposed an algorithm called *Deterministic Minimum Empirical Divergence (DMED)* which they proved to be first order optimal. This algorithm, which maintains a list of arms that are close enough to the best one (and which thus must be played), is inspired by large deviations ideas and relies on the availability of the rate function associated to the reward distribution.

In the second family of bandit problems, the rewards are only assumed to be bounded (say, between 0 and 1), and policies rely directly on the estimates of the expected rewards for each arm. The KL-UCB algorithm considered in this paper is primarily meant to address this second, non-parametric, setting. We will nonetheless show that KL-UCB also matches the lower bound of Burnetas and Katehakis (1997) in the binary case and that it can be extended to a larger class of parametric bandit problems.

Among all the bandit policies that have been proposed, a particular family has raised a strong interest, after Gittins (1979) proved that (in the Bayesian setting he considered) optimal policies could be chosen in the following very special form: compute for each arm a *dynamic allocation index* (which only depends on the draws of this arm), and simply choose an arm with maximal index. These policies not only compute an estimate of the expected rewards, but rather an *upper-confidence bound* (UCB), and the agent's choice is an arm with highest UCB. This approach is sometimes called "optimistic", as the agent acts as if, at each instant, the expected rewards were equal to the highest possible values that are compatible with her past observations. Auer et al. (2002), following Agrawal (1995), proposed and analyzed two variants, UCB1 (usually called simply UCB in latter works) and UCB2, for which they provided regret bounds. UCB is an online, horizon-free procedure for which (Auer et al., 2002) proves that there exists a constant  $C$  such that

$$\mathbb{E}[R_n] \leq \sum_{a: \mu_a < \mu_{a^*}} \frac{8 \log(n)}{(\mu_{a^*} - \mu_a)} + C. \quad (3)$$

The UCB2 variant relies on a parameter  $\alpha$  that needs to be tuned, depending in particular on the horizon, and satisfies the tighter regret bound

$$\mathbb{E}[R_n] \leq \sum_{a:\mu_a < \mu_{a^*}} \frac{(1 + \epsilon(\alpha)) \log(n)}{2(\mu_{a^*} - \mu_a)} + C(\alpha),$$

where  $\epsilon(\alpha) > 0$  is a constant that can get arbitrary small when  $\alpha$  is small, at the expense of an increased value of the constant  $C(\alpha)$ . The constant  $1/2$  in front of the factor  $\log(n)/(\mu_{a^*} - \mu_a)$  cannot be improved. We show in Proposition 4, as a by-product of our analysis, that a correctly tuned UCB algorithm satisfies a similar bound. However, Auer et al. (2002) found in numerical experiments that UCB and UCB2 were outperformed by a third heuristic variant called UCB-Tuned, which includes estimates of the variance, but no theoretical guarantee was given. In a latter work, Audibert et al. (2009) proposed a related policy, called UCB-V, which uses an empirical version of the Bernstein bound to obtain refined upper confidence bounds. Recently, Audibert and Bubeck (2010) introduced an improved UCB algorithm, termed MOSS, which achieves the distribution-free optimal rate.

In this contribution, we first consider the stochastic, non-parametric, bounded bandit problem. We consider an online index policy, called KL-UCB (for Kullback-Leibler UCB), that requires no problem- or horizon-dependent tuning. This algorithm was recently advocated by Filippi (2010), together with a similar procedure for Markov Decision Processes (Filippi et al., 2010), and we learnt since our initial submission that an analysis of the Bernoulli case can also be found in Maillard et al. (2011), together with other results. We prove in Theorem 1 below that the regret of KL-UCB satisfies

$$\limsup_{n \rightarrow \infty} \frac{\mathbb{E}[R_n]}{\log(n)} \leq \sum_{a:\mu_a < \mu_{a^*}} \frac{\mu_{a^*} - \mu_a}{d(\mu_a, \mu_{a^*})},$$

where  $d(p, q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q))$  denotes the Kullback-Leibler divergence between Bernoulli distributions of parameters  $p$  and  $q$ , respectively. This comes as a consequence of Theorem 2, a non-asymptotic upper-bound on the number of draws of a sub-optimal arm  $a$ : for all  $\epsilon > 0$  there exist  $C_1, C_2(\epsilon)$  and  $\beta(\epsilon)$  such that

$$\mathbb{E}[N_n(a)] \leq \frac{\log(n)}{d(\mu_a, \mu_{a^*})} (1 + \epsilon) + C_1 \log(\log(n)) + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}}.$$

We insist on the fact that, despite the presence of divergence  $d$ , this bound is not specific to the Bernoulli case and applies to all reward distributions bounded in  $[0, 1]$  (and thus, by rescaling, to all bounded reward distributions). By Pinsker's inequality,  $d(\mu_a, \mu_{a^*}) > 2(\mu_a - \mu_{a^*})^2$ , and thus KL-UCB has strictly better theoretical guarantees than UCB, while it has the same range of application. The improvement appears to be significant in simulations. Moreover, KL-UCB is the first index policy that reaches the lower-bound of Lai and Robbins (1985) for binary rewards; it does also achieve lower regret than UCB-V in that case. Hence, KL-UCB is both a general-purpose procedure for bounded bandits, and an optimal solution for the binary case.

Furthermore, it is easy to adapt KL-UCB to particular (possibly non-bounded) bandit settings, when the distribution of reward is known to belong to some family of probability

laws. By simply changing the definition of the divergence  $d$ , optimal algorithms may be built in a great variety of situations.

The proofs we give for these results are short and elementary. They rely on deviation results for bounded variables that are of independent interest : Lemma 9 shows that Bernoulli variable are, in a sense, the “least concentrated” bounded variables with a given expectation (as is well-known for variance), and Theorem 10 shows an efficient way to build confidence intervals for sums of bounded variables with an unknown number of summands.

In practice, numerical experiments confirm the significant advantage of KL-UCB over existing procedures; not only does this method outperform UCB, MOSS, UCB-V and even UCB-tuned in various scenarios, but it also compares well to DMED in the Bernoulli case, especially for small or moderate horizons.

The paper is organized as follows: in Section 2, we introduce some notation and present the KL-UCB algorithm. Section 3 contains the main results of the paper, namely the regret bound for KL-UCB and the optimality in the Bernoulli case. In Section 4, we show how to adapt the KL-UCB algorithm to address general families of reward distributions, and we provide finite-time regret bounds showing asymptotic optimality. Section 5 reports the results of extensive numerical experiments, showing the practical superiority of KL-UCB. Section 6 is devoted to an elementary proof of the main theorem. Finally, the Appendix gathers some deviation results that are useful in the proofs of our regret bounds, but which are also of independent interest.

## 2. The KL-UCB Algorithm

We consider the following bandit problem: the set of actions is  $\{1, \dots, K\}$ , where  $K$  denotes a finite integer. For each  $a \in \{1, \dots, K\}$ , the rewards  $(X_{a,t})_{t \geq 1}$  are independent and bounded<sup>2</sup> in  $\Theta = [0, 1]$  with common expectation  $\mu_a$ . The sequences  $(X_{a,\cdot})_a$  are independent. At each time step  $t = 1, 2, \dots$ , the agent chooses an action  $A_t$  according to his past observations (possibly using some independent randomization) and gets the reward  $X_t = X_{A_t, N_{A_t}(t)}$ , where  $N_a(t) = \sum_{s=1}^t \mathbb{1}\{A_s = a\}$  denotes the number of times action  $a$  was chosen up to time  $t$ . The sum of rewards she has obtained when choosing action  $a$  is denoted by  $S_a(t) = \sum_{s \leq t} \mathbb{1}\{A_s = a\} X_s = \sum_{s=1}^{N_a(t)} X_{a,s}$ . For  $(p, q) \in \Theta^2$  denote the Bernoulli Kullback-Leibler divergence by

$$d(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q},$$

with, by convention,  $0 \log 0 = 0 \log 0/0 = 0$  and  $x \log x/0 = +\infty$  for  $x > 0$ .

Algorithm 1 provides the pseudo-code for KL-UCB. On line 6,  $c$  is a parameter that, in the regret bound stated below in Theorem 1 is chosen equal to 3; in practice, however, we recommend to take  $c = 0$  for optimal performance. For each arm  $a$  the upper-confidence bound

$$\max \left\{ q \in \Theta : N[a] d \left( \frac{S[a]}{N[a]}, q \right) \leq \log(t) + c \log(\log(t)) \right\}$$

can be efficiently computed using Newton iterations, as for any  $p \in [0, 1]$  the function  $q \mapsto d(p, q)$  is strictly convex and increasing on the interval  $[p, 1]$ . In case of ties between

---

2. if the rewards are bounded in another interval  $[a, b]$ , they should first be rescaled to  $[0, 1]$ .

**Algorithm 1** KL-UCB

---

**Require:**  $n$  (horizon),  $K$  (number of arms), REWARD (reward function, bounded in  $[0, 1]$ )

---

```

1: for  $t = 1$  to  $K$  do
2:    $N[t] \leftarrow 1$ 
3:    $S[t] \leftarrow \text{REWARD}(\text{arm} = t)$ 
4: end for
5: for  $t = K + 1$  to  $n$  do
6:    $a \leftarrow \arg \max_{1 \leq a \leq K} \max \left\{ q \in \Theta : N[a] d \left( \frac{S[a]}{N[a]}, q \right) \leq \log(t) + c \log(\log(t)) \right\}$ 
7:    $r \leftarrow \text{REWARD}(\text{arm} = a)$ 
8:    $N[a] \leftarrow N[a] + 1$ 
9:    $S[a] \leftarrow S[a] + r$ 
10: end for

```

---

several arms, any maximizer can be chosen (for instance, at random). The KL-UCB elaborates on ideas suggested in Sections 3 and 4 of Lai and Robbins (1985).

### 3. Regret bounds and optimality

We first state the main result of this paper. It is a direct consequence of the non-asymptotic bound in Theorem 2 stated below.

**Theorem 1** Consider a bandit problem with  $K$  arms and independent rewards bounded in  $[0, 1]$ , and denote by  $a^*$  an optimal arm. Choosing  $c = 3$ , the regret of the KL-UCB algorithm satisfies:

$$\limsup_{n \rightarrow \infty} \frac{\mathbb{E}[R_n]}{\log(n)} \leq \sum_{a: \mu_a < \mu_{a^*}} \frac{\mu_{a^*} - \mu_a}{d(\mu_a, \mu_{a^*})}.$$

**Theorem 2** Consider a bandit problem with  $K$  arms and independent rewards bounded in  $[0, 1]$ . Let  $\epsilon > 0$ , and take  $c = 3$  in Algorithm 1. Let  $a^*$  denote an arm with maximal expected reward  $\mu_{a^*}$ , and let  $a$  be an arm such that  $\mu_a < \mu_{a^*}$ . For any positive integer  $n$ , the number of times algorithm KL-UCB chooses arm  $a$  is upper-bounded by

$$\mathbb{E}[N_n(a)] \leq \frac{\log(n)}{d(\mu_a, \mu_{a^*})} (1 + \epsilon) + C_1 \log(\log(n)) + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}},$$

where  $C_1$  denotes a positive constant and where  $C_2(\epsilon)$  and  $\beta(\epsilon)$  denote positive functions of  $\epsilon$ . Hence,

$$\limsup_{n \rightarrow \infty} \frac{\mathbb{E}[N_n(a)]}{\log(n)} \leq \frac{1}{d(\mu_a, \mu_{a^*})}.$$

**Corollary 3** If the reward distributions are Bernoulli, the KL-UCB algorithm is asymptotically optimal, in the sense that the regret of KL-UCB matches the lower-bound proved by Lai and Robbins (1985) and generalized by Burnetas and Katehakis (1997):

$$N_n(a) \geq \left( \frac{1}{d(\mu_a, \mu_{a^*})} + o(1) \right) \log(n)$$

with a probability tending to 1.

The KL-UCB algorithm thus appears to be (asymptotically) optimal for Bernoulli rewards. However, Lemma 9 shows that the Chernoff bounds obtained for Bernoulli variables actually apply to any variable with range  $[0, 1]$ . This is why KL-UCB is not only efficient in the binary case, but also for general bounded rewards.

As a by-product, we obtain an improved upper-bound for the regret of the UCB algorithm:

**Proposition 4** *Consider the UCB algorithm tuned as follows: at step  $t > K$ , the arm that maximizes the upper-bound  $S[a]/N[a] + \sqrt{(\log(t) + c \log \log(t))/(2N[a])}$  is chosen. Then, for the choice  $c = 3$ , the number of draws of a sub-optimal arm  $a$  is upper-bounded as:*

$$\mathbb{E}[N_n(a)] \leq \frac{\log(n)}{2(\mu_a - \mu_{a^*})^2} (1 + \epsilon) + C_1 \log(\log(n)) + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}}. \quad (4)$$

This bound is “optimal”, in the sense that the constant  $1/2$  in the logarithmic term cannot be improved. The proof of this proposition just mimics that of Section 6 (which concerns KL-UCB), using the quadratic divergence  $d(p, q) := 2(p-q)^2$  instead of the Kullback-Leibler divergence; it is thus omitted. In contrast, Pinsker’s inequality  $d(\mu_a, \mu_{a^*}) \geq 2(\mu_a - \mu_{a^*})^2$  shows that KL-UCB dominates UCB, and we will see in the simulation study that the difference is significant, even for smaller values of the horizon.

**Remark 5** *At line 6, Algorithm 1 computes for each arm  $a \in \{1, \dots, K\}$  the upper-confidence bound*

$$\max \left\{ q \in \Theta : N[a] d \left( \frac{S[a]}{N[a]}, q \right) \leq \log(t) + c \log(\log(t)) \right\}.$$

*The level of this confidence bound is parameterized by the exploration function  $\log(t) + c \log(\log(t))$ , and the results of Theorems 1 and 2 are true as soon as  $c \geq 3$ . However, similar results can be proved with an exploration function equal to  $(1 + \epsilon) \log(t)$  (instead of  $\log(t) + c \log(\log(t))$ ) for every  $\epsilon > 0$ ; this is no surprise, as  $(1 + \epsilon) \log(t) \geq \log(t) + c \log(\log(t))$  when  $t$  is large enough. But “large enough”, in that case, can be quite large : for  $\epsilon = 0.1$ , this holds true only for  $t > 2.10^{51}$ . This is why, in practice (and for the simulations presented in Section 5), we rather suggest to choose  $c = 0$ .*

#### 4. KL-UCB for parametric families of reward distributions

The KL-UCB algorithm makes no assumption on the distribution of the rewards, except that they are bounded. Actually, the definition of the divergence function  $d$  in KL-UCB is dictated by the rate function of the Large Deviations Principle satisfied by Bernoulli random variables: the proof of Theorem 10 relies on the control of the Fenchel-Legendre transform of the Bernoulli distribution. Thanks to Lemma 9, this choice also makes sense for all bounded variables.

But the method presented here is not limited to the Bernoulli case: KL-UCB can very easily be adapted to other settings by choosing an appropriate divergence function  $d$ . As an

illustration, we will assume in this section that, for each arm  $a$ , the distribution of rewards belongs to a *canonical exponential family*, i.e., that its density with respect to some reference measure can be written as  $p_{\theta_a}(x)$  for some real parameter  $\theta_a$ , with

$$p_{\theta}(x) = \exp(x\theta - b(\theta) + c(x)) , \quad (5)$$

where  $\theta$  is a real parameter,  $c$  is a real function and the log-partition function  $b(\cdot)$  is assumed to be twice differentiable. This family contains for instance the Exponential, Poisson, Gaussian (with fixed variance), Gamma (with fixed shape parameter) distributions (as well as, of course, the Bernoulli distribution). For a random variable  $X$  with density defined in (5), it is easily checked that  $\mu(\theta) \triangleq \mathbb{E}_{\theta}[X] = b'(\theta)$ ; moreover, as  $b''(\theta) = \text{Var}(X) > 0$ , the function  $\theta \mapsto \mu(\theta)$  is one-to-one. Theorem 11 (in the Appendix) states that the probability of under-estimating the performance of the best arm can be upper-bounded just as in the Bernoulli case by replacing the divergence  $d(\cdot, \cdot)$  in line 6 of the KL-UCB algorithm by

$$d(x, \mu(\theta)) = \sup_{\lambda} \{\lambda x - \log \mathbb{E}_{\theta} [\exp(\lambda X)]\} .$$

For example, in the case of exponential rewards, one should choose  $d(x, y) = x/y - 1 - \log(x/y)$ . Or, for Poisson rewards, the right choice is  $d(x, y) = y - x + x \log(x/y)$ . Then, all the results stated above apply (as the proofs do not involve the particular form of the function  $d$ ), and in particular :

$$\limsup_{n \rightarrow \infty} \frac{\mathbb{E}[R_n]}{\log(n)} \leq \sum_{a: \mu_a < \mu_{a^*}} \frac{\mu_{a^*} - \mu_a}{d(\mu_a, \mu_{a^*})} .$$

In order to prove that, for those families of rewards, this version of the KL-UCB algorithm matches the bound of Lai and Robbins (1985), it remains only to show that  $d(x, y) = KL(p_{\mu^{-1}(x)}, p_{\mu^{-1}(y)})$ . This is the object of Lemma 6. Generalizations to other families of reward distributions (possibly different from arm to arm) are conceivable, but require more technical, topological discussions, as in Burnetas and Katehakis (1997) and Honda and Takemura (2010).

To conclude, observe that it is not required to work with the divergence function  $d$  corresponding exactly to the family of reward distributions: using an upper-bound instead often leads to more simple and versatile policies at the price of only a slight loss in performance. This is illustrated in the third scenario of the simulation study presented in Section 5, but also by Theorems 1 and 2 for bounded variables.

**Lemma 6** *Let  $(\beta, \theta)$  be two real numbers, let  $p_{\beta}$  and  $p_{\theta}$  be two probability densities of the canonical exponential family defined in (5), and let  $X$  have density  $p_{\theta}$ . Then Kullback-Leibler divergence  $KL(p_{\beta}, p_{\theta})$  is equal to  $d(\mu(\beta), \mu(\theta))$ . More precisely,*

$$KL(p_{\beta}, p_{\theta}) = d(\mu(\beta), \mu(\theta)) = \mu(\beta)(\beta - \theta) - b(\beta) + b(\theta) .$$

**Proof** First, it holds that

$$\begin{aligned} KL(p_{\beta}, p_{\theta}) &= \int \exp(x\beta - b(\beta) + c(x)) \{x(\beta - \theta) - b(\beta) + b(\theta)\} dx \\ &= \mu(\beta)(\beta - \theta) - b(\beta) + b(\theta) . \end{aligned}$$

Then, observe that  $\mathbb{E} [\exp(\lambda X)] = \int \exp(x(\beta + \lambda) - b(\beta) + c(x)) dx = \exp(b(\beta + \lambda) - b(\beta))$ . Thus, for every  $x$  the maximum of the (smooth, concave) function

$$\lambda \mapsto \lambda x - \log \mathbb{E} [\exp(\lambda X)] = \lambda x - b(\theta + \lambda) + b(\theta)$$

is reached for  $\lambda = \lambda^*$  such that  $x = \dot{b}(\theta + \lambda^*) = \mu(\theta + \lambda^*)$ . Thus, if  $x = \mu(\beta)$ , the fact that  $\mu$  is one-to-one implies that  $\theta + \lambda^* = \beta$  and thus that:

$$d(\mu(\beta), \mu(\theta)) = (\beta - \theta) \mu(\beta) - b(\beta) + b(\theta).$$

■

## 5. Numerical experiments and comparisons of the policies

Simulations studies require particular attention in the case of bandit algorithms. As pointed out by Audibert et al. (2009), for a fixed horizon  $n$  the distribution of the regret is very poorly concentrated around its expectation. This can be explained as follows: most of the time, the estimates of all arms remain correctly ordered for almost all instants  $t = 1, \dots, n$  and the regret is of order  $\log(n)$ . But sometimes, at the beginning, the best arm is under-estimated while one of the sub-optimal arms is over-estimated, so that the agent keeps playing the latter; and as she neglects the best arm, she has hardly an occasion to realize her mistake, and the error perpetuates for a very long time. This happens with a small, but not negligible probability, because the regret is very important (of order  $n$ ) on these occasions. Bandit algorithms are typically designed to control the probability of such adverse events but usually at a rate which only decays slightly faster than  $1/n$ , which results in very skewed regret distributions with slowly vanishing upper tails.

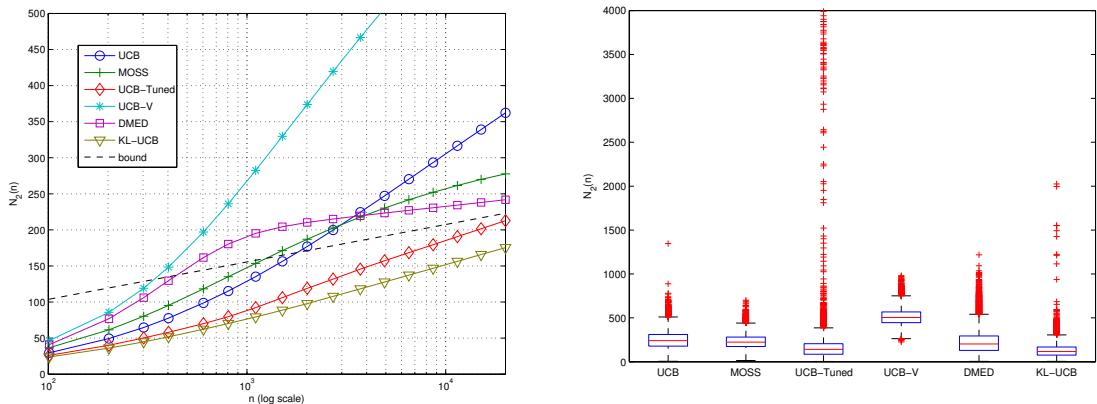


Figure 1: Performance of the various algorithms in the simple two arms, scenario. Left, mean number of draws of the suboptimal arm as a function of time; right, box-plots showing the distributions of the number of draws of the suboptimal arm at time  $n = 5,000$ . Results based on 50,000 independent runs.

### 5.1. Scenario 1: two arms

We first consider the basic two arm scenario with Bernoulli rewards of expectations  $\mu_1 = 0.9$  and  $\mu_2 = 0.8$ , respectively. The left panel of Figure 1 shows the average number of draws of the suboptimal arm as a function of time (on a logarithmic scale) for KL-UCB compared to five benchmark algorithms (UCB, MOSS, UCB-Tuned, UCB-V and DMED). The right panel of Figure 1 shows the empirical distributions of suboptimal draws, represented as box-and-whiskers plots, at a particular time ( $t = 5,000$ ) for all six algorithms. These plot are obtained from  $N = 50,000$  independent runs of the algorithms and the right panel of Figure 1 clearly highlight the tail effect mentioned above. On this very simple example, we observed that results obtained from  $N = 1,000$  or less simulations were not reliable, typically resulting in a significant over-estimation<sup>3</sup> of the performance of “risky” algorithms, in particular of UCB-Tuned. More generally, results obtained in configurations where  $N$  is much smaller than  $n$  are likely to be unreliable. For this reason, we limit our investigations to a final instant of  $n = 20,000$ . Note however that the average number of suboptimal draws of most algorithms at  $n = 20,000$  is only of the order of 300, showing that there is no point in considering larger horizons for such a simple problem.

MOSS, UCB-Tuned and UCB-V are run exactly as described by Audibert and Bubeck (2010), Auer et al. (2002) and Audibert et al. (2009), respectively. For UCB, we use an upper confidence bound  $S[a]/N[a] + \sqrt{\log(t)/(2N[a])}$ , as in Proposition 4, again with  $c = 0$ . Note that in our two arm scenario,  $\{2(\mu_1 - \mu_2)^2\}^{-1} = 50$  while  $d^{-1}(\mu_2, \mu_1) = 22.5$ . Hence, the performance of DMED and KL-UCB should be about two times better than that of UCB. The left panel of Figure 1 does show the expected behavior but with a difference of lesser magnitude. Indeed, one can observe that the bound  $d^{-1}(\mu_2, \mu_1) \log(n)$  (shown in dashed line) is quite pessimistic for the values of the horizon  $n$  considered here as the actual performance of KL-UCB is significantly below the bound. For DMED, we follow Honda and Takemura (2010) but using

$$N[a] d \left( \frac{S[a]}{N[a]}, \max_b \frac{S[b]}{N[b]} \right) < \log t \quad (6)$$

as the criterion to decide whether arm  $a$  should be included in the list of arms to be played. This criterion is clearly related to the decision rule used by KL-UCB when  $c = 0$  (see line 6 of Algorithm 1) except for the fact that in KL-UCB the estimate  $S[a]/N[a]$  is not compared to that of the current best arm  $\max_b S[b]/N[b]$  but to the corresponding upper confidence bound. As a consequence, any arm that is not included in the list of arms to be played by DMED would not be played by KL-UCB either (assuming that both algorithms share a common history). As one can observe on the left panel of Figure 1, this results in a degraded performance for DMED. We also observed this effect on UCB, for instance, and it seems that index algorithms are generally preferable to their “arm elimination” variant.

The original proposal of Honda and Takemura (2010) consists in using in the exploration function a factor  $\log(t/N[a])$  instead of  $\log(t)$ , as in the MOSS algorithm. As will be seen below on Figure 2, this variant (which we refer to as DMED+) indeed outperforms DMED. But our previous conjecture appears to hold also in this case as the heuristic variant of KL-

---

3. Incidentally, Theorem 10 could be used to construct sharp confidence bounds for the regret.

UCB (termed KL-UCB+) in which  $\log(t)$  in line 6 of Algorithm 1 is replaced by  $\log(t/N[a])$  remains preferable to DMED+.

As final comments on Figure 1, first note that UCB-Tuned performs as expected — though slightly worse than KL-UCB— but is a very risky algorithm: the right panel of Figure 1 casts some doubts on the fact that the tails of  $N_a(n)$  are indeed controlled uniformly in  $n$  for UCB-Tuned. Second, the performance of UCB-V is somewhat disappointing. Indeed, the upper-confidence bounds of UCB-V differ from those of UCB-Tuned simply by the non-asymptotic correction term  $3\log(t)/N[a]$  required by Bennett’s and Bernstein’s inequalities (Audibert et al., 2009). This correction term appears to have a significant impact on moderate time horizons: for a sub-optimal arm  $a$ , the number of draws  $N[a]$  does not grow faster than the  $\log(t)$  exploration function, and  $\log(t)/N[a]$  does not vanish.

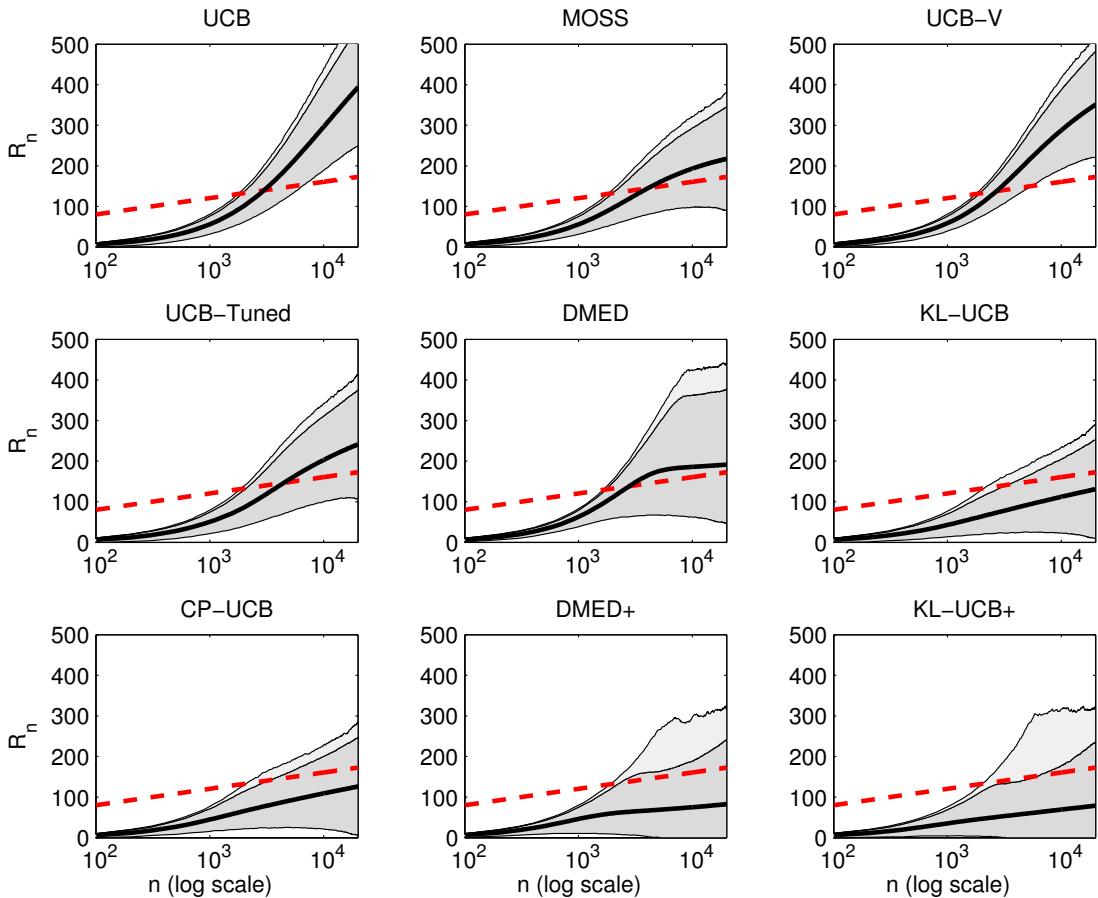


Figure 2: Regret of the various algorithms as a function of time (on a log scale) in the ten arm scenario. On each graph, the red dashed line shows the lower bound, the solid bold curve corresponds to the mean regret while the dark and light shaded regions show respectively the central 99% region and the upper 0.05% quantile, respectively.

## 5.2. Scenario 2: low rewards

In Figure 2 we consider a significantly more difficult scenario, again with Bernoulli rewards, inspired by a situation (frequent in applications like marketing or Internet advertising) where the mean reward of each arm is very low. In this scenario, there are ten arms: the optimal arm has expected reward 0.1, and the nine suboptimal arms consist of three different groups of three (stochastically) identical arms each with expected rewards 0.05, 0.02 and 0.01, respectively. We again used  $N = 50,000$  simulations to obtain the regret plots of Figure 2. These plots show, for each algorithm, the average cumulated regret together with quantiles of the cumulated regret distribution as a function of time (again on a logarithmic scale).

In this scenario, the difference is more pronounced between UCB and KL-UCB. The performance gain of UCB-Tuned is also much less significant. KL-UCB and DMED reach a performance that is on par with the lower bound of Burnetas and Katehakis (1997) in (2), although the performance of KL-UCB is here again significantly better. Using KL-UCB+ and DMED+ results in significant mean improvements, although there are hints that those algorithms might indeed be too risky with occasional very large deviations from the mean regret curve.

The final algorithm included in this roundup, called CP-UCB, is in some sense a further adaptation of KL-UCB to the specific case of Bernoulli rewards. For  $n \in \mathbb{N}$  and  $p \in [0, 1]$ , denote by  $P_{n,p}$  the binomial distribution with parameters  $n$  and  $p$ . For a random variable  $X$  with distribution  $P_{n,p}$ , the *Clopper-Pearson* (see Clopper and Pearson (1934)) or “exact” upper-confidence bound of risk  $\alpha \in ]0, 1[$  for  $p$  is

$$u^{CP}(X, n, \alpha) = \max \{q \in [0, 1] : P_{n,q}([0, X]) \geq \alpha\} .$$

It is easily verified that  $P_{n,p}(\mu \leq u^{CP}(X)) \geq 1 - \alpha$ , and that  $u^{CP}(X)$  is the smallest quantity satisfying this property:  $u^{CP}(X) \leq \tilde{u}(X)$  for any other upper-confidence bound  $\tilde{u}(X)$  of risk at most  $\alpha$ .

The Clopper-Pearson Upper-Confidence Bound algorithm (CP-UCB) differs from KL-UCB only in the way the upper-confidence bound on the performance of each arm is computed, replacing line 6 of Algorithm 1 by

$$a \leftarrow \arg \max_{1 \leq a \leq K} u^{CP} \left( S[a], N[a], \frac{1}{t \log(t)^c} \right) .$$

As the Clopper-Wilson confidence intervals are always sharper than the Kullback-Leibler intervals, one can very easily adapt the proof of Section 6 to show that the regret bounds proved for the KL-UCB algorithm also hold for CP-UCB in the case of Bernoulli rewards. However, the improvement over KL-UCB is very limited (often, the two algorithms actually take exactly the same decisions). In terms of results, one can observe on Figure 2 that CP-UCB only achieves a performance that is marginally better than that of KL-UCB. Besides, there is no guarantee that the CP-UCB algorithm is also efficient on arbitrary bounded distributions.

## 5.3. Scenario 3: bounded exponential rewards

In the third example, there are 5 arms: the rewards are exponential variables, with parameters  $1/5, 1/4, 1/3, 1/2$  and 1 respectively, truncated at  $x_{\max} = 10$  (thus, they are bounded

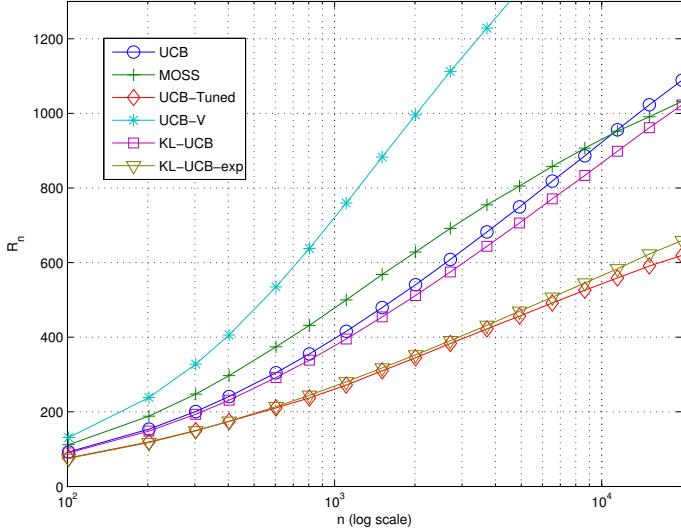


Figure 3: Regret of the various algorithms as a function of time in the bounded exponential scenario.

in  $[0, 10]$ ). The interest of this scenario is twofold: first, it shows the performance of KL-UCB for non-binary, non-discrete, non  $[0, 1]$ -valued rewards. Second, it illustrates that, as explained in Section 4, specific variants of the KL-UCB algorithm can reach an even better performance.

In this scenario, UCB and MOSS, but also KL-UCB are clearly sub-optimal. UCB-Tuned and UCB-V, by taking into account the variance of the reward distributions (much smaller than the variance of a  $\{0, 10\}$ -valued distribution with the same expectation), were expected to perform significantly better. For the reasons mentioned above this is not the case for UCB-V on a time horizon  $n = 20,000$ . Yet, UCB-Tuned is spectacularly more efficient, and is only caught up by KL-UCB-exp, the variant of KL-UCB designed for exponential rewards. Actually, the KL-UCB-exp algorithm ignores the fact that the rewards are truncated, and uses the divergence  $d(x, y) = x/y - 1 - \log(x/y)$  prescribed for genuine exponential distributions. One can easily show that this choice leads to slightly too large upper confidence bounds. Yet, the performance is still excellent, stable, and the algorithm is particularly simple.

## 6. Proof of Theorem 2

Consider a positive integer  $n$ , a small  $\epsilon > 0$ , an optimal arm  $a^*$  and a sub-optimal arm  $a$  such that  $\mu_a < \mu_{a^*}$ . Without loss of generality, we will assume that  $a^* = 1$ . For any arm  $b$ , the past average performance of arm  $b$  is denoted by  $\hat{\mu}_b(t) = S_b(t)/N_b(t)$ ; by convenience, for every positive integer  $s$  we will also denote  $\hat{\mu}_{b,s} = (X_{b,1} + \dots + X_{b,s})/s$ , so that  $\hat{\mu}_t(b) = \hat{\mu}_{b,N_b(t)}$ . KL-UCB relies on the following upper-confidence bound for  $\mu_b$ :

$$u_b(t) = \max \{q > \hat{\mu}_b(t) : N_b(t) d(\hat{\mu}_b(t), q) \leq \log(t) + 3 \log(\log(t))\} .$$

For  $x, y \in [0, 1]$ , define  $d^+(x, y) = d(x, y)\mathbb{1}_{x < y}$ . The expectation of  $N_n(a)$  is upper-bounded by using the following decomposition:

$$\begin{aligned}\mathbb{E}[N_n(a)] &= \mathbb{E}\left[\sum_{t=1}^n \mathbb{1}\{A_t = a\}\right] \leq \mathbb{E}\left[\sum_{t=1}^n \mathbb{1}\{\mu_1 > u_1(t)\}\right] + \mathbb{E}\left[\sum_{t=1}^n \mathbb{1}\{A_t = a, \mu_1 \leq u_1(t)\}\right] \\ &\leq \sum_{t=1}^n \mathbb{P}(\mu_1 > u_1(t)) + \mathbb{E}\left[\sum_{s=1}^n \mathbb{1}\{sd^+(\hat{\mu}_{a,s}, \mu_1) < \log(n) + 3\log(\log(n))\}\right],\end{aligned}$$

where the last inequality is a consequence of Lemma 7. The first summand is upper-bounded as follows: by Theorem 10 (proved in the Appendix),

$$\begin{aligned}P(\mu_1 > u_1(t)) &\leq e^{\lceil \log(t)(\log(t) + 3\log(\log(t))) \rceil} \exp(-\log(t) - 3\log(\log(t))) \\ &= \frac{e^{\lceil \log(t)^2 + 3\log(t)\log(\log(t)) \rceil}}{t\log(t)^3}.\end{aligned}$$

Hence,

$$\sum_{t=1}^n P(\mu_1 > u_1(t)) \leq \sum_{t=1}^n \frac{e^{\lceil \log(t)^2 + 3\log(t)\log(\log(t)) \rceil}}{t\log(t)^3} \leq C'_1 \log(\log(n))$$

for some positive constant  $C'_1$  ( $C'_1 \leq 7$  is sufficient). For the second summand, let

$$K_n = \left\lfloor \frac{1+\epsilon}{d^+(\mu_a, \mu_1)} (\log(n) + 3\log(\log(n))) \right\rfloor.$$

Then:

$$\begin{aligned}\sum_{s=1}^n \mathbb{P}(sd^+(\hat{\mu}_{a,s}, \mu_1) < \log(n) + 3\log(\log(n))) \\ &\leq K_n + \sum_{s=K_n+1}^{\infty} \mathbb{P}(sd^+(\hat{\mu}_{a,s}, \mu_1) < \log(n) + 3\log(\log(n))) \\ &\leq K_n + \sum_{s=K_n+1}^{\infty} \mathbb{P}(K_n d^+(\hat{\mu}_{a,s}, \mu_1) < \log(n) + 3\log(\log(n))) \\ &= K_n + \sum_{s=K_n+1}^{\infty} \mathbb{P}\left(d^+(\hat{\mu}_{a,s}, \mu_1) < \frac{d(\mu_a, \mu_1)}{1+\epsilon}\right) \\ &\leq \frac{1+\epsilon}{d^+(\mu_a, \mu_1)} (\log(n) + 3\log(\log(n))) + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}}\end{aligned}$$

according to Lemma 8. This will conclude the proof, provided that we prove the following two lemmas.

### Lemma 7

$$\sum_{t=1}^n \mathbb{1}\{A_t = a, \mu_1 \leq u_1(t)\} \leq \sum_{s=1}^n \mathbb{1}\{sd^+(\hat{\mu}_{a,s}, \mu_1) < \log(n) + 3\log(\log(n))\}.$$

**Proof** Observe that  $A_t = a$  and  $\mu_1 \leq u_1(t)$  together imply that  $u_a(t) \geq u_1(t) \geq \mu_1$ , and hence that

$$d^+(\hat{\mu}_a(t), \mu_1) \leq d(\hat{\mu}_a(t), u_a(t)) = \frac{\log(t) + 3\log(\log(t))}{N_a(t)}.$$

Thus,

$$\begin{aligned} \sum_{t=1}^n \mathbb{1}\{A_t = a, \mu_1 \leq u_1(t)\} &\leq \sum_{t=1}^n \mathbb{1}\{A_t = a, N_a(t)d^+(\hat{\mu}_a(t), \mu_1) \leq \log(t) + 3\log(\log(t))\} \\ &= \sum_{t=1}^n \sum_{s=1}^t \mathbb{1}\{N_t(a) = s, A_t = a, sd^+(\hat{\mu}_{a,s}, \mu_1) \leq \log(t) + 3\log(\log(t))\} \\ &\leq \sum_{t=1}^n \sum_{s=1}^t \mathbb{1}\{N_t(a) = s, A_t = a\} \mathbb{1}\{sd^+(\hat{\mu}_{a,s}, \mu_1) \leq \log(n) + 3\log(\log(n))\} \\ &= \sum_{s=1}^n \mathbb{1}\{sd^+(\hat{\mu}_{a,s}, \mu_1) \leq \log(n) + 3\log(\log(n))\} \sum_{t=s}^n \mathbb{1}\{N_t(a) = s, A_t = a\} \\ &= \sum_{s=1}^n \mathbb{1}\{sd^+(\hat{\mu}_{a,s}, \mu_1) \leq \log(n) + 3\log(\log(n))\}, \end{aligned}$$

as, for every  $s \in \{1, \dots, n\}$ ,  $\sum_{t=s}^n \mathbb{1}\{N_t(a) = s, A_t = a\} \leq 1$ . ■

**Lemma 8** For each  $\epsilon > 0$ , there exist  $C_2(\epsilon) > 0$  and  $\beta(\epsilon) > 0$  such that

$$\sum_{s=K_n+1}^{\infty} \mathbb{P}\left(d^+(\hat{\mu}_{a,s}, \mu_1) < \frac{d(\mu_a, \mu_1)}{1+\epsilon}\right) \leq \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}}.$$

**Proof** If  $d^+(\hat{\mu}_{a,s}, \mu_1) < d(\mu_a, \mu_1)/(1+\epsilon)$ , then  $\hat{\mu}_{a,s} > r(\epsilon)$ , where  $r(\epsilon) \in ]\mu_a, \mu_1[$  is such that  $d(r(\epsilon), \mu_1) = d(\mu_a, \mu_1)/(1+\epsilon)$ . Hence,

$$\begin{aligned} \mathbb{P}\left(d^+(\hat{\mu}_{a,s}, \mu_1) < \frac{d(\mu_a, \mu_1)}{1+\epsilon}\right) &\leq \mathbb{P}(d(\hat{\mu}_{a,s}, \mu_a) > d(r(\epsilon), \mu_a), \hat{\mu}_{a,s} > \mu_a) \\ &\leq \mathbb{P}(\hat{\mu}_{a,s} > r(\epsilon)) \leq \exp(-sd(r(\epsilon), \mu_a)), \end{aligned}$$

and

$$\sum_{s=K_n+1}^{\infty} \mathbb{P}\left(d^+(\hat{\mu}_{a,s}, \mu_1) < \frac{d(\mu_a, \mu_1)}{1+\epsilon}\right) \leq \frac{\exp(-d(r(\epsilon), \mu_a)K_n)}{1 - \exp(-d(r(\epsilon), \mu_a))} \leq \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}},$$

with  $C_2(\epsilon) = (1 - \exp(-d(r(\epsilon), \mu_a)))^{-1}$  and  $\beta(\epsilon) = (1+\epsilon)d(r(\epsilon), \mu_1)/d(\mu_a, \mu_1)$ . Easy computations show that  $r(\epsilon) = \mu_a + O(\epsilon)$ , so that  $C_2(\epsilon) = O(\epsilon^{-2})$  and  $\beta(\epsilon) = O(\epsilon^2)$ . ■

## 7. Conclusion

The self-normalized deviation bound of Theorems 10 and 11, together with the new analysis presented in Section 6, allowed us to design and analyze improved UCB algorithms. In this approach, only an upper-bound of the deviations (more precisely, of the exponential moments) of the rewards is required, which makes it possible to obtain versatile policies satisfying interesting regret bounds for large classes of reward distributions. The resulting index policies are simple, fast, and very efficient in practice, even for small time horizons.

## References

- R. Agrawal. Sample mean based index policies with  $O(\log n)$  regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.
- J-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.
- J-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19), 2009.
- P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.
- A.N. Burnetas and M.N. Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, pages 222–255, 1997.
- C.J. Clopper and E.S. Pearson. The use of confidence or fiducial limits illustration in the case of the binomial. *Biometrika*, 26:404–413, 1934.
- E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Conf. Comput. Learning Theory (Sydney, Australia, 2002)*, volume 2375 of *Lecture Notes in Comput. Sci.*, pages 255–270. Springer, Berlin, 2002.
- S. Filippi. *Optimistic strategies in Reinforcement Learning* (in French). PhD thesis, Telecom ParisTech, 2010. URL <http://tel.archives-ouvertes.fr/tel-00551401/>.
- S. Filippi, O. Cappé, and A. Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. In *Allerton Conf. on Communication, Control, and Computing*, Monticello, US, 2010.
- J.C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2):148–177, 1979.
- J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In T. Kalai and M. Mohri, editors, *Conf. Comput. Learning Theory*, Haifa, Israel, 2010.
- T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

O-A. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Conf. Comput. Learning Theory*, Budapest, Hungary, 2011.

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

## Appendix A. Kullback-Leibler deviations for bounded variables with a random number of summands

We start with a simple lemma justifying the focus on binary rewards.

**Lemma 9** *Let  $X$  be a random variable taking value in  $[0, 1]$ , and let  $\mu = \mathbb{E}[X]$ . Then, for all  $\lambda \in \mathbb{R}$ ,*

$$E[\exp(\lambda X)] \leq 1 - \mu + \mu \exp(\lambda),$$

**Proof** The function  $f : [0, 1] \rightarrow \mathbb{R}$  defined by  $f(x) = \exp(\lambda x) - x(\exp(\lambda) - 1) - 1$  is convex and such that  $f(0) = f(1) = 0$ , hence  $f(x) \leq 0$  for all  $x \in [0, 1]$ . Consequently,

$$\mathbb{E}[\exp(\lambda X)] \leq \mathbb{E}[X(\exp(\lambda) - 1) + 1] = \mu(\exp(\lambda) - 1) + 1.$$

■

**Theorem 10** *Let  $(X_t)_{t \geq 1}$  be a sequence of independent random variables bounded in  $[0, 1]$  defined on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  with common expectation  $\mu = \mathbb{E}[X_t]$ . Let  $\mathcal{F}_t$  be an increasing sequence of  $\sigma$ -fields of  $\mathcal{F}$  such that for each  $t$ ,  $\sigma(X_1, \dots, X_t) \subset \mathcal{F}_t$  and for  $s > t$ ,  $X_s$  is independent from  $\mathcal{F}_t$ . Consider a previsible sequence  $(\epsilon_t)_{t \geq 1}$  of Bernoulli variables (for all  $t > 0$ ,  $\epsilon_t$  is  $\mathcal{F}_{t-1}$ -measurable). Let  $\delta > 0$  and for every  $t \in \{1, \dots, n\}$  let*

$$S(t) = \sum_{s=1}^t \epsilon_s X_s, \quad N(t) = \sum_{s=1}^t \epsilon_s, \quad \hat{\mu}(t) = \frac{S(t)}{N(t)},$$

$$u(n) = \max \{q > \hat{\mu}_n : N(n)d(\hat{\mu}(n), q) \leq \delta\}.$$

Then

$$\mathbb{P}(u(n) < \mu) \leq e \lceil \delta \log(n) \rceil \exp(-\delta).$$

**Proof** For every  $\lambda \in \mathbb{R}$ , let  $\phi_\mu(\lambda) = \log \mathbb{E}[\exp(\lambda X_1)]$ . By Lemma 9, it holds that  $\phi_\mu(\lambda) \leq \log(1 - \mu + \mu \exp(\lambda))$ . Let  $W_0^\lambda = 1$  and for  $t \geq 1$ ,

$$W_t^\lambda = \exp(\lambda S(t) - N(t)\phi_\mu(\lambda)).$$

$(W_t^\lambda)_{t \geq 0}$  is a super-martingale relative to  $(\mathcal{F}_t)_{t \geq 0}$ . In fact,

$$\begin{aligned} \mathbb{E}[\exp(\lambda \{S(t+1) - S(t)\}) | \mathcal{F}_t] &= \mathbb{E}[\exp(\lambda \epsilon_{t+1} X_{t+1}) | \mathcal{F}_t] = \exp(\epsilon_{t+1} \log \mathbb{E}[\exp(\lambda X_1)]) \\ &\leq \exp(\epsilon_{t+1} \phi_\mu(\lambda)) = \exp(\{N(t+1) - N(t)\} \phi_\mu(\lambda)) \end{aligned}$$

which can be rewritten as

$$\mathbb{E}[\exp(\lambda S(t+1) - N(t+1)\phi_\mu(\lambda)) | \mathcal{F}_t] \leq \exp(\lambda S(t) - N(t)\phi_\mu(\lambda)).$$

To proceed, we make use of the so-called 'peeling trick' (see for instance Massart (2007)): we divide the interval  $\{1, \dots, n\}$  of possible values for  $N(n)$  into "slices"  $\{t_{k-1} + 1, \dots, t_k\}$  of geometrically increasing size, and treat the slices independently. We may assume that  $\delta > 1$ , since otherwise the bound is trivial. Take<sup>4</sup>  $\eta = 1/(\delta - 1)$ , let  $t_0 = 0$  and for  $k \in \mathbb{N}^*$ , let  $t_k = \lfloor (1 + \eta)^k \rfloor$ . Let  $D$  be the first integer such that  $t_D \geq n$ , that is  $D = \lceil \frac{\log n}{\log 1+\eta} \rceil$ . Let  $A_k = \{t_{k-1} < N(n) \leq t_k\} \cap \{u(n) < \mu\}$ . We have:

$$\mathbb{P}(u(n) < \mu) \leq \mathbb{P}\left(\bigcup_{k=1}^D A_k\right) \leq \sum_{k=1}^D \mathbb{P}(A_k). \quad (7)$$

Observe that  $u(n) < \mu$  if and only if  $\hat{\mu}(n) < \mu$  and  $N(n)d(\hat{\mu}(n), \mu) > \delta$ . Let  $s$  be the smallest integer such that  $\delta/(s+1) \leq d(0, \mu)$ ; if  $N(n) \leq s$ , then  $N(n)d(\hat{\mu}(n), \mu) \leq sd(\hat{\mu}(n), \mu) \leq sd(0, \mu) < \delta$  and  $\mathbb{P}(u(n) < \mu) = 0$ . Thus,  $\mathbb{P}(A_k) = 0$  for all  $k$  such that  $t_k \leq s$ .

For  $k$  such that  $t_k > s$ , let  $\tilde{t}_{k-1} = \max\{t_{k-1}, s\}$ . Let  $x \in ]0, \mu[$  be such that  $d(x, \mu) = \delta/N(n)$  and let  $\lambda(x) = \log(x(1-\mu)) - \log(\mu(1-x)) < 0$ , so that  $d(x, \mu) = \lambda(x)x - (1-\mu + \mu \exp(\lambda(x)))$ . Consider  $z$  such that  $z < \mu$  and  $d(z, \mu) = \delta/(1+\eta)^k$ . Observe that:

- if  $N(n) > \tilde{t}_{k-1}$ , then

$$d(z, \mu) = \frac{\delta}{(1+\eta)^k} \geq \frac{\delta}{(1+\eta)N(n)};$$

- if  $N(n) \leq \tilde{t}_{k-1}$ , then as

$$d(\hat{\mu}(n), \mu) > \frac{\delta}{N(n)} > \frac{\delta}{(1+\eta)^k} = d(z, \mu),$$

it holds that :

$$\hat{\mu}(n) < \mu \text{ and } d(\hat{\mu}(n), \mu) > \frac{\delta}{N(n)} \implies \hat{\mu}(n) \leq z.$$

Hence, on the event  $\{\tilde{t}_{k-1} < N(n) \leq t_k\} \cap \{\hat{\mu}(n) < \mu\} \cap \left\{d(\hat{\mu}(n), \mu) > \frac{\delta}{N(n)}\right\}$  it holds that

$$\lambda(z)\hat{\mu}(n) - \phi_\mu(\lambda(z)) \geq \lambda(z)z - \phi_\mu(\lambda(z)) = d(z, \mu) \geq \frac{\delta}{(1+\eta)N(n)}.$$

Putting everything together,

$$\begin{aligned} & \{\tilde{t}_{k-1} < N(n) \leq t_k\} \cap \{\hat{\mu}(n) < \mu\} \cap \left\{d(\hat{\mu}(n), \mu) \geq \frac{\delta}{N(n)}\right\} \\ & \subset \left\{\lambda(z)\hat{\mu}(n) - \phi_\mu(\lambda(z)) \geq \frac{\delta}{N(n)(1+\eta)}\right\} \\ & \subset \left\{\lambda(z)S_n - N(n)\phi_\mu(\lambda(z)) \geq \frac{\delta}{1+\eta}\right\} \\ & \subset \left\{W_n^{\lambda(z)} > \exp\left(\frac{\delta}{1+\eta}\right)\right\}. \end{aligned}$$

---

4. if  $\delta \leq 1$ , it is easy to check that the bound holds whatsoever.

As  $(W_t^\lambda)_{t \geq 0}$  is a supermartingale,  $\mathbb{E} [W_n^{\lambda(z)}] \leq \mathbb{E} [W_0^{\lambda(z)}] = 1$ , and the Markov inequality yields:

$$\begin{aligned} \mathbb{P} & \left( \{\tilde{t}_{k-1} < N(n) \leq t_k\} \cap \{\hat{\mu}(n) \geq \mu\} \cap \{N(n)d(\hat{\mu}(n), \mu) \geq \delta\} \right) \\ & \leq \mathbb{P} \left( W_n^{\lambda(z)} > \exp \left( \frac{\delta}{1+\eta} \right) \right) \leq \exp \left( -\frac{\delta}{1+\eta} \right). \end{aligned}$$

Finally, by Equation (7),

$$\mathbb{P} \left( \bigcup_{k=1}^D \{\tilde{t}_{k-1} < N(n) \leq t_k\} \cap \{u(n) < \mu\} \right) \leq D \exp \left( -\frac{\delta}{1+\eta} \right).$$

But as  $\eta = 1/(\delta - 1)$ ,  $D = \left\lceil \frac{\log n}{\log(1+1/(\delta-1))} \right\rceil$  and as  $\log(1+1/(\delta-1)) \geq 1/\delta$ , we obtain:

$$\mathbb{P}(u(n) < \mu) \leq \left\lceil \frac{\log n}{\log \left( 1 + \frac{1}{\delta-1} \right)} \right\rceil \exp(-\delta + 1) \leq e \lceil \delta \log(n) \rceil \exp(-\delta).$$

■

Of course, a symmetric bound for the probability of over-estimating  $\mu$  can be derived following the same lines. Together, they show that for all  $\delta > 0$ :

$$\mathbb{P}(N(n)d(\hat{\mu}(n), \mu) > \delta) \leq 2e \lceil \delta \log(n) \rceil \exp(-\delta).$$

Finally, we state a more general deviation bound for arbitrary reward distributions with finite exponential moments. The proof (very similar to that of Theorem 10) is omitted.

**Theorem 11** *Let  $(X_t)_{t \geq 1}$  be a sequence of i.i.d. random variables defined on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  with common expectation  $\mu$ . Assume that the cumulant-generating function*

$$\phi(\lambda) = \log \mathbb{E} [\exp(\lambda X_1)]$$

*is defined and finite on some open subset  $\lambda_1, \lambda_2$  of  $\mathbb{R}$  containing 0. Define  $d : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$  as follows: for all  $x \in \mathbb{R}$ ,*

$$d(x, \mu) = \sup_{\lambda \in ]\lambda_1, \lambda_2[} \{\lambda x - \phi(\lambda)\}.$$

*Let  $\mathcal{F}_t$  be an increasing sequence of  $\sigma$ -fields of  $\mathcal{F}$  such that for each  $t$ ,  $\sigma(X_1, \dots, X_t) \subset \mathcal{F}_t$  and for  $s > t$ ,  $X_s$  is independent from  $\mathcal{F}_t$ . Consider a previsible sequence  $(\epsilon_t)_{t \geq 1}$  of Bernoulli variables (for all  $t > 0$ ,  $\epsilon_t$  is  $\mathcal{F}_{t-1}$ -measurable). Let  $\delta > 0$  and for every  $t \in \{1, \dots, n\}$  let*

$$\begin{aligned} S(t) &= \sum_{s=1}^t \epsilon_s X_s, \quad N(t) = \sum_{s=1}^t \epsilon_s, \quad \hat{\mu}(t) = \frac{S(t)}{N(t)}, \\ u(n) &= \max \{q > \hat{\mu}_n : N(n)d(\hat{\mu}(n), q) \leq \delta\}. \end{aligned}$$

*Then*

$$\mathbb{P}(u(n) < \mu) \leq e \lceil \delta \log(n) \rceil \exp(-\delta).$$

# Sparsity Regret Bounds for Individual Sequences in Online Linear Regression

Sébastien Gerchinovitz

École Normale Supérieure\*

Paris, France

SEBASTIEN.GERCHINOVITZ@ENS.FR

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We consider the problem of online linear regression on arbitrary deterministic sequences when the ambient dimension  $d$  can be much larger than the number of time rounds  $T$ . We introduce the notion of sparsity regret bound, which is a deterministic online counterpart of recent risk bounds derived in the stochastic setting under a sparsity scenario. We prove such regret bounds for an online-learning algorithm called SeqSEW and based on exponential weighting and data-driven truncation. In a second part we apply a parameter-free version of this algorithm on i.i.d. data and derive risk bounds of the same flavor as in Dalalyan and Tsybakov (2008, 2011) but which solve two questions left open therein. In particular our risk bounds are adaptive (up to a logarithmic factor) to the unknown variance of the noise if the latter is Gaussian.

**Keywords:** Individual Sequences, Sparsity, Online Linear Regression, Regret Bounds

## 1. Introduction

We consider the problem of online linear regression on arbitrary deterministic sequences. A forecaster has to predict in a sequential fashion the values  $y_t \in \mathbb{R}$  of an unknown sequence of observations given some input data  $x_t \in \mathcal{X}$  and some base forecasters  $\varphi_j : \mathcal{X} \rightarrow \mathbb{R}$ ,  $1 \leq j \leq d$ , on the basis of which he outputs a prediction  $\hat{y}_t \in \mathbb{R}$ . The quality of the predictions is assessed by the square loss. The goal of the forecaster is to predict almost as well as the best linear forecaster  $\mathbf{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^d u_j \varphi_j$ , where  $\mathbf{u} \in \mathbb{R}^d$ , i.e., to satisfy, uniformly over all individual sequences  $(x_t, y_t)_{1 \leq t \leq T}$ , a regret bound of the form

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + \Delta_{T,d}(\mathbf{u}) \right\},$$

for some regret term  $\Delta_{T,d}(\mathbf{u})$  that should be as small as possible and, in particular, sublinear in  $T$ .

In this setting the variant of the sequential Ridge regression forecaster studied by Azoury and Warmuth (2001) and Vovk (2001) has a regret of order at most  $d \ln T$ . When the ambient dimension  $d$  is much larger than the number of time rounds  $T$ , the latter regret bound may unfortunately be larger than  $T$  and is thus somehow trivial. Since the regret bound

---

\* This research was carried out within the INRIA project CLASSIC hosted by École Normale Supérieure and CNRS.

$d \ln T$  is optimal in a certain sense (see Vovk 2001, Theorem 2), additional assumptions are needed to get interesting theoretical guarantees.

A natural assumption, which has already been extensively studied in the stochastic setting, is that there is a sparse linear combination  $\mathbf{u}^*$  (i.e., with  $s \ll T/(\ln T)$  non-zero coefficients) which has a small cumulative square loss. If the forecaster knew in advance the support  $J(\mathbf{u}^*) \triangleq \{j : u_j^* \neq 0\}$  of  $\mathbf{u}^*$ , he could apply the same forecaster as above but only to the  $s$ -dimensional linear subspace  $\{\mathbf{u} \in \mathbb{R}^d : \forall j \notin J(\mathbf{u}^*), u_j = 0\}$ . The regret bound of this ‘oracle’ would be roughly of order  $s \ln T$  and thus sublinear in  $T$ . Under this sparsity scenario, a sublinear regret thus seems possible, though, of course, the aforementioned regret bound  $s \ln T$  can only be used as an ideal benchmark (since the support of  $\mathbf{u}^*$  is unknown).

In this paper, we prove that a regret bound proportional to  $s$  is achievable (up to logarithmic factors). In Corollary 2 and its refinements (Proposition 4 combined with Remark 6, and Theorem 8) we indeed derive regret bounds of the form

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + (\|\mathbf{u}\|_0 + 1) g_{T,d}(\|\mathbf{u}\|_1, \|\boldsymbol{\varphi}\|_\infty) \right\}, \quad (1)$$

where  $\|\mathbf{u}\|_0$  denotes the number of non-zero coordinates of  $\mathbf{u}$  and where  $g$  is increasing but grows at most logarithmically in  $T$ ,  $d$ ,  $\|\mathbf{u}\|_1 \triangleq \sum_{j=1}^d |u_j|$ , and  $\|\boldsymbol{\varphi}\|_\infty \triangleq \sup_{x \in \mathcal{X}} \max_{1 \leq j \leq d} |\varphi_j(x)|$ .

We call regret bounds of the above form *sparsity regret bounds*.

This work is in connection with several papers that appeared at previous COLT conferences, either in the stochastic setting (Bunea et al., 2006; Dalalyan and Tsybakov, 2007, 2009) or in online convex optimization (Duchi et al., 2010). Next we discuss these papers and some related references.

### Related works in the stochastic setting

The above regret bound (1) can be seen as a deterministic online counterpart of the so-called *sparsity oracle inequalities* introduced in the stochastic setting in the past decade. The latter are risk bounds expressed in terms of the number of non-zero coefficients of the oracle vector. Such inequalities were introduced by Bunea et al. (2004, 2006) for the regression model with random design. The same authors prove similar results for the case of a fixed design in Bunea et al. (2007) through general model selection arguments of Birgé and Massart (2001). As we do not have the space to thoroughly review the extensive literature related to sparsity oracle inequalities, we refer the reader to the full version of this paper (Gerchinovitz, 2011) for further references.

We only mention that, recently, sparsity oracle inequalities with leading constant equal to 1 have been proved for procedures based on exponential weighting; see Dalalyan and Tsybakov (2007) and the other references given in Gerchinovitz (2011). These papers show that a trade-off can be reached between strong theoretical guarantees (as with  $\ell^0$ -regularization) and computational efficiency (as with  $\ell^1$ -regularization). They indeed propose aggregation algorithms which satisfy sparsity oracle inequalities under almost no assumption on the base forecasters  $(\varphi_j)_j$ , and which can be approximated numerically at a reasonable computational cost for large values of the ambient dimension  $d$ .

Our online-learning algorithm SeqSEW is inspired from Dalalyan and Tsybakov (2008, 2011). Following the same lines as in Dalalyan and Tsybakov (2009), it is possible to slightly adapt its statement to make it computationally tractable by means of Langevin Monte-Carlo approximation while not affecting its statistical properties. The technical details are however omitted in this paper, which only focuses on the theoretical guarantees of the algorithm SeqSEW.

### Previous works on sparsity in the framework of individual sequences

To the best of our knowledge, Corollary 2 and its refinements (Proposition 4 combined with Remark 6, and Theorem 8) provide the first examples of sparsity regret bounds in the sense of (1). To comment on the optimality of such regret bounds and compare them to related results in the framework of individual sequences, note that (1) can be rewritten in the equivalent form:

For all  $s \in \mathbb{N}$  and all  $U > 0$ ,

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 - \inf_{\substack{\|\mathbf{u}\|_0 \leq s \\ \|\mathbf{u}\|_1 \leq U}} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \varphi(x_t))^2 \leq (s+1) g_{T,d}(U, \|\varphi\|_\infty) ,$$

where  $g$  grows at most logarithmically in  $T$ ,  $d$ ,  $U$ , and  $\|\varphi\|_\infty$ . When  $s \ll T$ , this upper bound matches (up to logarithmic factors) the lower bound of order  $s \ln T$  that follows in a straightforward manner from Vovk (2001, Theorem 2) or Cesa-Bianchi and Lugosi (2006, Chapter 11). Indeed, if  $s \ll T$ ,  $\mathcal{X} = \mathbb{R}^d$ , and  $\varphi_j(x) = x_j$ , then for any forecaster, there is an individual sequence  $(x_t, y_t)_{1 \leq t \leq T}$  such that the regret of this forecaster on  $\{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\|_0 \leq s \text{ and } \|\mathbf{u}\|_1 \leq d\}$  is bounded from below by a quantity of order  $s \ln T$ . Therefore, up to logarithmic factors, any algorithm satisfying a sparsity regret bound of the form (1) is minimax optimal on intersections of  $\ell^0$ -balls (of radii  $s \ll T$ ) and  $\ell^1$ -balls. This is in particular the case for our algorithm SeqSEW, but this contrasts with related works discussed below.

Recent works in the field of online convex optimization addressed the sparsity issue in the online deterministic setting, but from a quite different angle. They focus on algorithms which output sparse linear combinations, while we are interested in algorithms whose regret is small under a sparsity scenario, i.e., on  $\ell^0$ -balls of small radii. See, e.g., Langford et al. (2009); Shalev-Shwartz and Tewari (2009); Xiao (2010); Duchi et al. (2010) and the references therein. All these articles focus on convex regularization. In the particular case of  $\ell^1$ -regularization under the square loss, the aforementioned works propose algorithms which predict as a sparse linear combination  $\hat{y}_t = \hat{\mathbf{u}}_t \cdot \varphi(x_t)$  of the base forecasts (i.e.,  $\|\hat{\mathbf{u}}_t\|_0$  is small), while no such guarantee can be proved for our algorithm SeqSEW. However they prove bounds on the  $\ell^1$ -regularized regret of the form

$$\sum_{t=1}^T ((y_t - \hat{\mathbf{u}}_t \cdot \mathbf{x}_t)^2 + \lambda \|\hat{\mathbf{u}}_t\|_1) \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T ((y_t - \mathbf{u} \cdot \mathbf{x}_t)^2 + \lambda \|\mathbf{u}\|_1) + \tilde{\Delta}_{T,d}(\mathbf{u}) \right\} , \quad (2)$$

for some regret term  $\tilde{\Delta}_{T,d}(\mathbf{u})$  which is suboptimal on intersections of  $\ell^0$ - and  $\ell^1$ -balls as explained below. The truncated gradient algorithm of Langford et al. (2009, Corollary 4.1)

satisfies<sup>1</sup> such a regret bound with  $\tilde{\Delta}_{T,d}(\mathbf{u})$  at least of order  $\|\varphi\|_\infty \sqrt{dT}$  when the base forecasts  $\varphi_j(x_t)$  are dense in the sense that  $\max_{1 \leq t \leq T} \sum_{j=1}^d \varphi_j^2(x_t) \approx d \|\varphi\|_\infty^2$ . This regret bound grows as a power of  $d$  and not logarithmically in  $d$  as is expected for sparsity regret bounds (recall that we are interested in the case when  $d \gg T$ ).

The three other papers mentioned above do prove (some) regret bounds with a logarithmic dependence in  $d$ , but these bounds do not have the dependence in  $\|\mathbf{u}\|_1$  and  $T$  we are looking for. For  $p-1 \approx 1/(\ln d)$ , the  $p$ -norm RDA method of Xiao (2010) and the algorithm SMIDAS of Shalev-Shwartz and Tewari (2009) – the latter being a particular case of the algorithm COMID of Duchi et al. (2010) specialized to the  $p$ -norm divergence – satisfy regret bounds of the above form (2) with  $\tilde{\Delta}_{T,d}(\mathbf{u}) \approx \mu \|\mathbf{u}\|_1 \sqrt{T \ln d}$ , for some gradient-based constant  $\mu$ . Therefore, in all three cases, the function  $\tilde{\Delta}$  grows at least linearly in  $\|\mathbf{u}\|_1$  and as  $\sqrt{T}$ . This is in contrast with the logarithmic dependence in  $\|\mathbf{u}\|_1$  and the fast rate  $\mathcal{O}(\ln T)$  we are looking for and prove, e.g., in Corollary 2.

Note that the suboptimality of the aforementioned algorithms is specific to the goal we are pursuing, i.e., prediction on  $\ell^0$ -balls (intersected with  $\ell^1$ -balls). On the contrary the rate  $\|\mathbf{u}\|_1 \sqrt{T \ln d}$  is more suited and actually optimal for learning on  $\ell^1$ -balls (see Raskutti et al. 2009). Moreover, the predictions output by our algorithm SeqSEW are not necessarily sparse linear combinations of the base forecasts. A question left open is thus whether it is possible to design an algorithm which both outputs sparse linear combinations (which is statistically useful and sometimes essential for computational issues) and satisfies a sparsity regret bound of the form (1).

### PAC-Bayesian analysis in the framework of individual sequences

To derive our sparsity regret bounds, we follow a PAC-Bayesian approach combined with the choice of a sparsity-favoring prior. We do not have the space to review the PAC-Bayesian literature in the stochastic setting and only refer the reader to Catoni (2004) for a thorough introduction to the subject. As for the online deterministic setting, PAC-Bayesian inequalities were proved in the framework of prediction with expert advice, e.g., in Freund et al. (1997) and Kivinen and Warmuth (1999), or in the same setting as ours with a Gaussian prior in Vovk (2001). More recently, Audibert (2009) proved a PAC-Bayesian result on individual sequences for general losses and prediction sets. The latter result relies on a unifying assumption called the online variance inequality, which holds true, e.g., when the loss function is exp-concave. In the present paper, we only focus on the particular case of the square loss. We first use Theorem 4.6 of Audibert (2009) to derive a non-adaptive sparsity regret bound. We then provide an adaptive online PAC-Bayesian inequality to automatically adapt to the unknown range of the observations  $\max_{1 \leq t \leq T} |y_t|$ .

---

1. The bound stated in Langford et al. (2009, Corollary 4.1) differs from (2) in that the constant before the infimum is equal to  $C = 1/(1 - 2c_d^2\eta)$ , where  $c_d^2 \approx \max_{1 \leq t \leq T} \sum_{j=1}^d \varphi_j^2(x_t) \leq d \|\varphi\|_\infty^2$ , and where a reasonable choice for  $\eta$  can easily be seen to be  $\eta \approx 1/\sqrt{2c_d^2 T}$ . If the base forecasts  $\varphi_j(x_t)$  are dense in the sense that  $c_d^2 \approx d \|\varphi\|_\infty^2$ , then we have  $C \approx 1 + \sqrt{2c_d^2/T}$ , which yields a regret bound with leading constant 1 as in (2) and with  $\tilde{\Delta}_{T,d}(\mathbf{u})$  at least of order  $\sqrt{c_d^2 T} \approx \|\varphi\|_\infty \sqrt{dT}$ .

### Open questions by Dalalyan and Tsybakov

In Section 4 we apply a parameter-free version of our algorithm SeqSEW on i.i.d. data and derive a risk bound of the same flavor as in Dalalyan and Tsybakov (2008, 2011). However, our risk bound holds on the whole  $\mathbb{R}^d$  space instead of  $\ell^1$ -balls of finite radii, which solves one question left open by Dalalyan and Tsybakov (2011, Section 4.2). Besides, our algorithm does not need the a priori knowledge of the variance factor of the noise when the latter is subgaussian, which solves a second question raised in Dalalyan and Tsybakov (2011, Section 5.1, Remark 6).

### Outline of the paper

This paper is organized as follows. In Section 2 we describe our deterministic setting and main notations. In Section 3 we prove the aforementioned sparsity regret bounds for our algorithm SeqSEW, first when the forecaster has access to some a priori knowledge on the observations (Sections 3.1 and 3.2), and then when no a priori information is available (Section 3.3), which yields a fully automatic algorithm. Finally, in Section 4, we apply one version of the algorithm SeqSEW on i.i.d. data and provide positive answers to two questions left open by Dalalyan and Tsybakov (2011).

## 2. Setting and notations

The main setting considered in this paper is an equivalent variant of an extension of the game of prediction with expert advice called *prediction with side information (under the square loss)* or, more simply, *online linear regression*; see Cesa-Bianchi and Lugosi (2006, Chapter 11) for references on this setting. We give in Figure 1 a detailed description of our repeated game.

We now define some notations. Vectors in  $\mathbb{R}^d$  will be denoted by bold letters. For all  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ , the standard inner product in  $\mathbb{R}^d$  between  $\mathbf{u} = (u_1, \dots, u_d)$  and  $\mathbf{v} = (v_1, \dots, v_d)$  will be denoted by  $\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^d u_i v_i$ ; the  $\ell^0$ -,  $\ell^1$ -, and  $\ell^2$ -norms of  $\mathbf{u} = (u_1, \dots, u_d)$  are respectively defined by

$$\|\mathbf{u}\|_0 \triangleq \sum_{j=1}^d \mathbb{I}_{\{u_j \neq 0\}} = |\{j : u_j \neq 0\}|, \quad \|\mathbf{u}\|_1 \triangleq \sum_{j=1}^d |u_j|, \quad \text{and} \quad \|\mathbf{u}\|_2 \triangleq \left( \sum_{j=1}^d u_j^2 \right)^{1/2}.$$

The set of all probability distributions on a set  $\Theta$  (endowed with some  $\sigma$ -algebra, e.g., the Borel  $\sigma$ -algebra when  $\Theta = \mathbb{R}^d$ ) will be denoted by  $\mathcal{M}_1^+(\Theta)$ . For all  $\rho, \pi \in \mathcal{M}_1^+(\Theta)$ , the Kullback-Leibler divergence between  $\rho$  and  $\pi$  is defined by

$$\mathcal{K}(\rho, \pi) \triangleq \begin{cases} \int_{\mathbb{R}^d} \ln \left( \frac{d\rho}{d\pi} \right) d\rho & \text{if } \rho \text{ is absolutely continuous with respect to } \pi; \\ +\infty & \text{otherwise,} \end{cases}$$

where  $\frac{d\rho}{d\pi}$  denotes the Radon-Nikodym derivative of  $\rho$  with respect to  $\pi$ .

**Parameters:** input data set  $\mathcal{X}$ , base forecasters  $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_d)$  with  $\varphi_j : \mathcal{X} \rightarrow \mathbb{R}$ ,  $1 \leq j \leq d$ .

**Initial step:** the environment chooses<sup>a</sup> a sequence of observations  $(y_t)_{t \geq 1}$  in  $\mathbb{R}$  and a sequence of input data  $(x_t)_{t \geq 1}$  in  $\mathcal{X}$  but the forecaster has not access to them.

**At each time round**  $t \in \mathbb{N}^*$ ,

1. The environment reveals the input data  $x_t \in \mathcal{X}$ .
2. The forecaster chooses a prediction  $\hat{y}_t \in \mathbb{R}$  (possibly as a linear combination of the  $\varphi_j(x_t)$ , but this is not necessary).
3. The environment reveals the observation  $y_t \in \mathbb{R}$ .
4. Each linear forecaster  $\mathbf{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^d u_j \varphi_j$ ,  $\mathbf{u} \in \mathbb{R}^d$ , incurs the loss  $(y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2$  and the forecaster incurs the loss  $(y_t - \hat{y}_t)^2$ .

*a.* The game is described as if the environment were oblivious to the forecaster's predictions. Actually, since we only consider deterministic forecasters, our results also hold when  $(x_t)_{t \geq 1}$  and  $(y_t)_{t \geq 1}$  are chosen by an adversarial environment.

Figure 1: Description of the repeated game of online linear regression.

For all  $x \in \mathbb{R}$  and  $B > 0$ , we denote by  $\lceil x \rceil$  the smallest integer larger than or equal to  $x$ , and by  $[x]_B$  its thresholded value:

$$[x]_B \triangleq \begin{cases} -B & \text{if } x < -B; \\ x & \text{if } -B \leq x \leq B; \\ B & \text{if } x > B. \end{cases}$$

Finally, we will use the (natural) convention  $0 \ln(1 + U/0) = 0$  for all  $U \geq 0$ .

### 3. Sparsity regret bounds for individual sequences

In this section we prove sparsity regret bounds for different variants of our algorithm SeqSEW. We first assume in Section 3.1 that the forecaster has access in advance to a bound  $B_y$  on the observations  $|y_t|$  and a bound  $B_\Phi$  on the trace of the empirical Gram matrix. We then remove these requirements one by one in Sections 3.2 and 3.3.

#### 3.1. Known bounds $B_y$ on the observations and $B_\Phi$ on the trace of the empirical Gram matrix

To simplify the analysis, we first assume that, at the beginning of the game, the number of rounds  $T$  is known to the forecaster and that he has access to a bound  $B_y$  on all the

observations  $y_1, \dots, y_T$  and to a bound  $B_\Phi$  on the trace of the empirical Gram matrix, i.e.,

$$y_1, \dots, y_T \in [-B_y, B_y] \quad \text{and} \quad \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leq B_\Phi .$$

The first version of the algorithm studied in this paper is defined in Figure 2 (adaptive variants will be introduced later). We name it *SeqSEW* for it is a variant of the Sparse Exponential Weighting algorithm introduced in the stochastic setting by Dalalyan and Tsybakov (2007, 2008) which is tailored for the prediction of individual sequences.

The choice of the heavy-tailed prior  $\pi_\tau$  is due to Dalalyan and Tsybakov (2007). The role of heavy-tailed priors to tackle the sparsity issue was already pointed out earlier; see, e.g., the discussion in Seeger (2008, Section 2.1). In high dimension, such heavy-tailed priors favor sparsity: sampling from these prior distributions (or posterior distributions based on them) typically results in approximately sparse vectors, i.e., vectors having most coordinates almost equal to zero and the few remaining ones with quite large values.

**Parameters:** threshold  $B > 0$ , inverse temperature  $\eta > 0$ , and prior scale  $\tau > 0$  with which we associate the *sparsity prior*  $\pi_\tau \in \mathcal{M}_1^+(\mathbb{R}^d)$  defined by

$$\pi_\tau(\mathrm{d}\mathbf{u}) \triangleq \prod_{j=1}^d \frac{(3/\tau) \mathrm{d}u_j}{2(1 + |u_j|/\tau)^4} . \quad (3)$$

**Initialization:**  $p_1 \triangleq \pi_\tau$ .

**At each time round**  $t \geq 1$ ,

1. Get the input data  $x_t$  and predict as  $\hat{y}_t \triangleq \int_{\mathbb{R}^d} [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_B p_t(\mathrm{d}\mathbf{u})$ ;
2. Get the observation  $y_t$  and compute the posterior distribution  $p_{t+1} \in \mathcal{M}_1^+(\mathbb{R}^d)$  as

$$p_{t+1}(\mathrm{d}\mathbf{u}) \triangleq \frac{\exp\left(-\eta \sum_{s=1}^t (y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_B)^2\right)}{W_{t+1}} \pi_\tau(\mathrm{d}\mathbf{u}) ,$$

where

$$W_{t+1} \triangleq \int_{\mathbb{R}^d} \exp\left(-\eta \sum_{s=1}^t (y_s - [\mathbf{v} \cdot \boldsymbol{\varphi}(x_s)]_B)^2\right) \pi_\tau(\mathrm{d}\mathbf{v}) .$$

Figure 2: Definition of the algorithm SeqSEW $_{\tau}^{B,\eta}$ .

**Proposition 1** Assume that, for a known constant  $B_y > 0$ , the  $(x_1, y_1), \dots, (x_T, y_T)$  are such that

$$y_1, \dots, y_T \in [-B_y, B_y].$$

Then, for all  $B \geq B_y$ , all  $\eta \leq 1/(8B^2)$ , and all  $\tau > 0$ , the algorithm SeqSEW $_{\tau}^{B,\eta}$  satisfies

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \varphi(x_t))^2 + \frac{4}{\eta} \|\mathbf{u}\|_0 \ln \left( 1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0 \tau} \right) \right\} + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t). \quad (4)$$

**Corollary 2** Assume that, for some known constants  $B_y > 0$  and  $B_\Phi > 0$ , the  $(x_1, y_1), \dots, (x_T, y_T)$  are such that  $y_1, \dots, y_T \in [-B_y, B_y]$  and  $\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leq B_\Phi$ .

Then, when used with  $B = B_y$ ,  $\eta = \frac{1}{8B_y^2}$ , and  $\tau = \sqrt{\frac{16B_y^2}{B_\Phi}}$ , the algorithm SeqSEW $_{\tau}^{B,\eta}$  satisfies

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \varphi(x_t))^2 + 32 B_y^2 \|\mathbf{u}\|_0 \ln \left( 1 + \frac{\sqrt{B_\Phi} \|\mathbf{u}\|_1}{4 B_y \|\mathbf{u}\|_0} \right) \right\} + 16 B_y^2. \quad (5)$$

To prove Proposition 1, we first need the following deterministic PAC-Bayesian inequality which is at the core of our analysis. It is a straightforward consequence of Theorem 4.6 of Audibert (2009) when applied to the square loss. An adaptive variant of this inequality will be provided in Section 3.2.

**Lemma 3** Assume that for some known constant  $B_y > 0$ , we have  $y_1, \dots, y_T \in [-B_y, B_y]$ . For all  $\tau > 0$ , if the algorithm SeqSEW $_{\tau}^{B,\eta}$  is used with  $B \geq B_y$  and  $\eta \leq 1/(8B^2)$ , then

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - [\mathbf{u} \cdot \varphi(x_t)]_B)^2 \rho(d\mathbf{u}) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta} \right\} \quad (6)$$

$$\leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \varphi(x_t))^2 \rho(d\mathbf{u}) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta} \right\}. \quad (7)$$

**Proof (of Lemma 3)** Inequality (6) is a straightforward consequence of Theorem 4.6 of Audibert (2009) when applied to the square loss, the set of prediction functions  $\mathcal{G} \triangleq \{x \mapsto [\mathbf{u} \cdot \varphi(x)]_B : \mathbf{u} \in \mathbb{R}^d\}$ , and the prior<sup>2</sup>  $\pi$  on  $\mathcal{G}$  induced by the prior  $\pi_\tau$  on  $\mathbb{R}^d$  via the mapping  $\mathbf{u} \in \mathbb{R}^d \mapsto [\mathbf{u} \cdot \varphi(\cdot)]_B \in \mathcal{G}$ .

To apply the aforementioned theorem, recall from Vovk (2001, Remark 3) that the square loss is  $1/(8B^2)$ -exp-concave on  $[-B, B]$  and thus  $\eta$ -exp-concave<sup>3</sup> (since  $\eta \leq 1/(8B^2)$ )

2. The set  $\mathcal{G}$  is endowed with the  $\sigma$ -algebra generated by all the coordinate mappings  $g \in \mathcal{G} \mapsto g(x) \in \mathbb{R}$ ,  $x \in \mathcal{X}$  (where  $\mathbb{R}$  is endowed with its Borel  $\sigma$ -algebra).

3. This means that for all  $y \in [-B, B]$ , the function  $x \mapsto \exp(-\eta(y - x)^2)$  is concave on  $[-B, B]$ .

by assumption). Therefore, by Theorem 4.6 of Audibert (2009) with the variance function  $\delta_\eta \equiv 0$  (see the comments following Remark 4.1 therein), we get

$$\begin{aligned} \sum_{t=1}^T (y_t - \hat{y}_t)^2 &\leq \inf_{\mu \in \mathcal{M}_1^+(\mathcal{G})} \left\{ \int_{\mathcal{G}} \sum_{t=1}^T (y_t - g(x_t))^2 \mu(dg) + \frac{\mathcal{K}(\mu, \pi)}{\eta} \right\} \\ &\leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - [\mathbf{u} \cdot \varphi(x_t)]_B)^2 \rho(d\mathbf{u}) + \frac{\mathcal{K}(\tilde{\rho}, \pi)}{\eta} \right\}, \end{aligned}$$

where the last inequality follows by restricting the infimum over  $\mathcal{M}_1^+(\mathcal{G})$  to the subset  $\{\tilde{\rho} : \rho \in \mathcal{M}_1^+(\mathbb{R}^d)\} \subset \mathcal{M}_1^+(\mathcal{G})$ , where  $\tilde{\rho} \in \mathcal{M}_1^+(\mathcal{G})$  denotes the probability distribution induced by  $\rho \in \mathcal{M}_1^+(\mathbb{R}^d)$  via the mapping  $\mathbf{u} \in \mathbb{R}^d \mapsto [\mathbf{u} \cdot \varphi(\cdot)]_B \in \mathcal{G}$ . Inequality (6) then follows from the fact that for all  $\rho \in \mathcal{M}_1^+(\mathbb{R}^d)$ , we have  $\mathcal{K}(\tilde{\rho}, \pi) \leq \mathcal{K}(\rho, \pi_\tau)$  by joint convexity of  $\mathcal{K}(\cdot, \cdot)$ .

As for Inequality (7), it follows from (6) by noting that

$$\forall y \in [-B, B], \quad \forall x \in \mathbb{R}, \quad |y - [x]_B| \leq |y - x|.$$

Therefore, truncation to  $[-B, B]$  can only improve prediction under the square loss if the observations are  $[-B, B]$ -valued, which is the case here since by assumption  $y_t \in [-B_y, B_y] \subset [-B, B]$  for all  $t = 1, \dots, T$ .  $\blacksquare$

**Proof (of Proposition 1)** Our proof mimics the proof of Theorem 5 in Dalalyan and Tsybakov (2008). We thus only write the outline of the proof and stress the minor changes that are needed to derive Inequality (4).

Let  $\mathbf{u}^* \in \mathbb{R}^d$ . Since  $B \geq B_y$  and  $\eta \leq 1/(8B^2)$ , we can apply Lemma 3 and get

$$\begin{aligned} \sum_{t=1}^T (y_t - \hat{y}_t)^2 &\leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \varphi(x_t))^2 \rho(d\mathbf{u}) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta} \right\} \\ &\leq \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \varphi(x_t))^2 \rho_{\mathbf{u}^*, \tau}(d\mathbf{u}) + \frac{\mathcal{K}(\rho_{\mathbf{u}^*, \tau}, \pi_\tau)}{\eta}. \end{aligned} \tag{8}$$

In the last inequality,  $\rho_{\mathbf{u}^*, \tau}$  is taken as the translated of  $\pi_\tau$  at  $\mathbf{u}^*$ , namely,

$$\rho_{\mathbf{u}^*, \tau}(d\mathbf{u}) \triangleq \frac{d\pi_\tau}{d\mathbf{u}}(\mathbf{u} - \mathbf{u}^*) d\mathbf{u} = \prod_{j=1}^d \frac{(3/\tau) du_j}{2(1 + |u_j - u_j^*|/\tau)^4}.$$

The two terms of the right-hand side of (8) can be upper bounded as in the proof of Theorem 5 in Dalalyan and Tsybakov (2008). It is proved therein that, by a symmetry argument,

$$\int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \varphi(x_t))^2 \rho_{\mathbf{u}^*, \tau}(d\mathbf{u}) = \sum_{t=1}^T (y_t - \mathbf{u}^* \cdot \varphi(x_t))^2 + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t),$$

and, by elementary calculations,

$$\frac{\mathcal{K}(\rho_{\mathbf{u}^*, \tau}, \pi_\tau)}{\eta} \leq \frac{4}{\eta} \|\mathbf{u}^*\|_0 \ln \left( 1 + \frac{\|\mathbf{u}^*\|_1}{\|\mathbf{u}^*\|_0 \tau} \right) .$$

Combining (8) with the last two equations, which all hold for all  $\mathbf{u}^* \in \mathbb{R}^d$ , we get Inequality (4).  $\blacksquare$

**Proof (of Corollary 2)** Applying Proposition 1, we have, since  $B \geq B_y$  and  $\eta \leq 1/(8B^2)$ ,

$$\begin{aligned} \sum_{t=1}^T (y_t - \hat{y}_t)^2 &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \varphi(x_t))^2 + \frac{4}{\eta} \|\mathbf{u}\|_0 \ln \left( 1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0 \tau} \right) \right\} + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \\ &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \varphi(x_t))^2 + \frac{4}{\eta} \|\mathbf{u}\|_0 \ln \left( 1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0 \tau} \right) \right\} + \tau^2 B_\Phi , \end{aligned} \quad (9)$$

since  $\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leq B_\Phi$  by assumption. The particular choices<sup>4</sup> for  $\eta$  and  $\tau$  given in the statement of the corollary then yield the desired inequality (5).  $\blacksquare$

### 3.2. Unknown bound $B_y$ on the observations but known bound $B_\Phi$ on the trace of the empirical Gram matrix

In the previous section, to prove the upper bounds stated in Lemma 3 and Proposition 1, we assumed that the forecaster had access to a bound  $B_y$  on the observations  $|y_t|$ . In this section, we remove this requirement and prove a sparsity regret bound for a variant of the algorithm SeqSEW $_{\tau}^{B,\eta}$  which is adaptive to the unknown bound  $B_y = \max_{1 \leq t \leq T} |y_t|$ ; see Proposition 4 and Remark 5 below.

For this purpose we consider the following algorithm called SeqSEW $_{\tau}^*$  thereafter. It differs from SeqSEW $_{\tau}^{B,\eta}$  defined in the previous section in that the threshold  $B$  and the inverse temperature  $\eta$  are now allowed to vary over time and are chosen at each time round as a function of the data available to the forecaster. More precisely, the algorithm SeqSEW $_{\tau}^*$  outputs at time  $t$  the prediction

$$\hat{y}_t \triangleq \int_{\mathbb{R}^d} [\mathbf{u} \cdot \varphi(x_t)]_{B_t} p_t(d\mathbf{u}) , \quad (10)$$

where

$$B_t \triangleq \left( 2^{\lceil \log_2 \max_{1 \leq s \leq t-1} y_s^2 \rceil} \right)^{1/2} , \quad \eta_t \triangleq \frac{1}{8B_t^2} ,$$

---

4. The best choice of  $(B, \eta)$  that satisfies the assumptions of Proposition 1 is  $B = B_y$  and  $\eta = 1/(8B_y^2)$ . As for the choice of  $\tau$ , it minimizes the function  $\tau \mapsto C_1 \ln(C_2/\tau) + C_3 \tau^2$  with  $C_1 = 4/\eta = 32B_y^2$  and  $C_3 = B_\Phi$ .

and where, for a normalizing constant  $W_t$ , the posterior distribution  $p_t \in \mathcal{M}_1^+(\mathbb{R}^d)$  is defined by

$$p_t(\mathrm{d}\mathbf{u}) \triangleq \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \left(y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_{B_s}\right)^2\right)}{W_t} \pi_\tau(\mathrm{d}\mathbf{u}).$$

Note that  $\max_{1 \leq s \leq t-1} |y_s| \leq B_t \leq \sqrt{2} \max_{1 \leq s \leq t-1} |y_s|$ .

The idea of truncating the base forecasts was already used in the past; see, e.g., Györfi et al. (2002) for the case of least squares regression and Györfi and Ottucsák (2007); Biau et al. (2010) for sequential prediction of unbounded time series under the square loss. A key ingredient in the present paper is to perform truncation with respect to a data-driven threshold. The online tuning of this threshold is based on a pseudo-doubling-trick technique provided in Cesa-Bianchi et al. (2007) (we use the prefix *pseudo* since the algorithm does not restart at the beginning of each new regime).

**Proposition 4** *For all  $\tau > 0$ , the algorithm SeqSEW $^*_\tau$  satisfies*

$$\begin{aligned} \sum_{t=1}^T (y_t - \hat{y}_t)^2 &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + 32B_{T+1}^2 \|\mathbf{u}\|_0 \ln \left( 1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0 \tau} \right) \right\} \\ &\quad + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) + 16B_{T+1}^2, \end{aligned} \quad (11)$$

where

$$B_{T+1}^2 \triangleq 2^{\lceil \log_2 \max_{1 \leq t \leq T} y_t^2 \rceil} \leq 2 \max_{1 \leq t \leq T} y_t^2.$$

**Remark 5** In view of Proposition 1, the algorithm SeqSEW $^*_\tau$  satisfies a sparsity regret bound which is adaptive to the unknown bound  $B_y = \max_{1 \leq t \leq T} |y_t|$ . The price for the automatic tuning with respect to  $B_y$  consists only of a multiplicative factor smaller than 2 and the additive factor  $16B_{T+1}^2$  which is smaller than  $32B_y^2$ .

**Remark 6** As in the previous section, several corollaries can be derived from Proposition 4. If the forecaster has access beforehand to a quantity  $B_\Phi > 0$  such that  $\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leq B_\Phi$ , then a suboptimal but reasonable choice of  $\tau$  is given by  $\tau = 1/\sqrt{B_\Phi}$ ; see the full version of this paper (Gerchinovitz, 2011, Corollary 3). We will also use the simpler choice  $\tau = 1/\sqrt{dT}$  for the stochastic setting in Section 4.

As in the previous section, to prove Proposition 4, we first need a key PAC-Bayesian inequality. The next lemma is an adaptive variant of Lemma 3.

**Lemma 7** For all  $\tau > 0$ , the algorithm SeqSEW $_{\tau}^*$  satisfies

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \left( y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t} \right)^2 \rho(d\mathbf{u}) + 8B_{T+1}^2 \mathcal{K}(\rho, \pi_{\tau}) \right\} + 8B_{T+1}^2 \quad (12)$$

$$\leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 \rho(d\mathbf{u}) + 8B_{T+1}^2 \mathcal{K}(\rho, \pi_{\tau}) \right\} + 16B_{T+1}^2, \quad (13)$$

where

$$B_{T+1}^2 \triangleq 2^{\lceil \log_2 \max_{1 \leq t \leq T} y_t^2 \rceil} \leq 2 \max_{1 \leq t \leq T} y_t^2.$$

**Proof (of Lemma 7)** The proof is based on similar arguments as for Lemma 3, except that we now need to deal with  $B$  and  $\eta$  changing over time. In the same spirit as in Auer et al. (2002); Cesa-Bianchi et al. (2007); Györfi and Ottucsák (2007), our analysis relies on the control of  $(\ln W_{t+1})/\eta_{t+1} - (\ln W_t)/\eta_t$  where  $W_1 \triangleq 1$  and, for all  $t \geq 2$ ,

$$W_t \triangleq \int_{\mathbb{R}^d} \exp \left( -\eta_t \sum_{s=1}^{t-1} \left( y_s - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_s)]_{B_s} \right)^2 \right) \pi_{\tau}(d\mathbf{u}).$$

On the one hand, we have

$$\begin{aligned} \frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} &= \frac{1}{\eta_{T+1}} \ln \int_{\mathbb{R}^d} \exp \left( -\eta_{T+1} \sum_{t=1}^T \left( y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t} \right)^2 \right) \pi_{\tau}(d\mathbf{u}) - \frac{1}{\eta_1} \ln 1 \\ &= \frac{1}{\eta_{T+1}} \sup_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \left( -\eta_{T+1} \sum_{t=1}^T \left( y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t} \right)^2 \right) \rho(d\mathbf{u}) - \mathcal{K}(\rho, \pi_{\tau}) \right\} \quad (14) \end{aligned}$$

$$= - \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \left( y_t - [\mathbf{u} \cdot \boldsymbol{\varphi}(x_t)]_{B_t} \right)^2 \rho(d\mathbf{u}) + \frac{\mathcal{K}(\rho, \pi_{\tau})}{\eta_{T+1}} \right\}, \quad (15)$$

where (14) follows from the fact that, for any measurable space  $(E, \mathcal{B})$ , any probability distribution  $\pi$  on  $(E, \mathcal{B})$ , and any non-positive measurable function  $h : E \rightarrow (-\infty, 0]$ , the Legendre transform of the Kullback-Leibler divergence can be expressed as

$$\ln \int_E e^h d\pi = \sup_{\rho \in \mathcal{M}_1^+(E)} \left\{ \int_E h d\rho - \mathcal{K}(\rho, \pi) \right\}.$$

This convex duality argument for the KL divergence is proved, e.g., in Catoni (2004, p. 159).

On the other hand, we can rewrite  $(\ln W_{T+1})/\eta_{T+1} - (\ln W_1)/\eta_1$  as a telescopic sum and get

$$\begin{aligned} \frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} &= \sum_{t=1}^T \left( \frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W_t}{\eta_t} \right) = \sum_{t=1}^T \underbrace{\left( \frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W'_{t+1}}{\eta_t} \right)}_{(1)} + \underbrace{\frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t}}_{(2)}, \end{aligned} \quad (16)$$

where  $W'_{t+1}$  is obtained from  $W_{t+1}$  by replacing  $\eta_{t+1}$  with  $\eta_t$ ; namely,

$$W'_{t+1} \triangleq \int_{\mathbb{R}^d} \exp \left( -\eta_t \sum_{s=1}^t \left( y_s - [\mathbf{u} \cdot \varphi(x_s)]_{B_s} \right)^2 \right) \pi_\tau(d\mathbf{u}) .$$

Let  $t \in \{1, \dots, T\}$ . The first term (1) is non-positive by Jensen's inequality (note that  $x \mapsto x^{\eta_{t+1}/\eta_t}$  is concave on  $\mathbb{R}_+^*$  since  $\eta_{t+1} \leq \eta_t$  by construction). As for the second term (2), by definition of  $W'_{t+1}$ ,

$$\begin{aligned} & \frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t} \\ &= \frac{1}{\eta_t} \ln \int_{\mathbb{R}^d} \frac{\exp \left( -\eta_t \left( y_t - [\mathbf{u} \cdot \varphi(x_t)]_{B_t} \right)^2 \right) \exp \left( -\eta_t \sum_{s=1}^{t-1} \left( y_s - [\mathbf{u} \cdot \varphi(x_s)]_{B_s} \right)^2 \right)}{W_t} \pi_\tau(d\mathbf{u}) \\ &= \frac{1}{\eta_t} \ln \int_{\mathbb{R}^d} \exp \left( -\eta_t \left( y_t - [\mathbf{u} \cdot \varphi(x_t)]_{B_t} \right)^2 \right) p_t(d\mathbf{u}) \end{aligned} \quad (17)$$

$$\leq \begin{cases} -(y_t - \hat{y}_t)^2 & \text{if } B_{t+1} = B_t; \\ -(y_t - \hat{y}_t)^2 + (2B_{t+1})^2 & \text{if } B_{t+1} > B_t; \end{cases} \quad (18)$$

where (17) follows by definition of  $p_t$ . To get Inequality (18) when  $B_{t+1} = B_t$ , or, equivalently,  $|y_t| \leq B_t$ , we used the fact that the square loss is  $1/(8B_t^2)$ -exp-concave on  $[-B_t, B_t]$  (as in Lemma 3). Indeed, by definition of  $\eta_t \triangleq 1/(8B_t^2)$  and by Jensen's inequality, we get

$$\int_{\mathbb{R}^d} e^{-\eta_t(y_t - [\mathbf{u} \cdot \varphi(x_t)]_{B_t})^2} p_t(d\mathbf{u}) \leq \exp \left( -\eta_t \left( y_t - \int_{\mathbb{R}^d} [\mathbf{u} \cdot \varphi(x_t)]_{B_t} p_t(d\mathbf{u}) \right)^2 \right) = e^{-\eta_t(y_t - \hat{y}_t)^2},$$

where the last equality follows by definition of  $\hat{y}_t$ . Taking the logarithms of both sides of the last inequality and dividing by  $\eta_t$ , we get (18) when  $B_{t+1} = B_t$ .

As for the rounds  $t$  such that  $B_{t+1} > B_t$ , the square loss  $x \mapsto (y_t - x)^2$  is no longer  $1/(8B_t^2)$ -exp-concave on  $[-B_t, B_t]$ . In this case (18) follows from the cruder upper bound  $(1/\eta_t) \ln(W'_{t+1}/W_t) \leq 0 \leq -(y_t - \hat{y}_t)^2 + (2B_{t+1})^2$  (since  $|y_t|, |\hat{y}_t| \leq B_{t+1}$ ). Summing (18) over  $t = 1, \dots, T$ , Equation (16) yields

$$\frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} \leq - \sum_{t=1}^T (y_t - \hat{y}_t)^2 + 4 \sum_{\substack{t=1 \\ t: B_{t+1} > B_t}}^T B_{t+1}^2 \leq - \sum_{t=1}^T (y_t - \hat{y}_t)^2 + 8B_{T+1}^2, \quad (19)$$

where, setting  $K \triangleq \lceil \log_2 \max_{1 \leq t \leq T} y_t^2 \rceil$ , we bounded the geometric sum  $\sum_{t: B_{t+1} > B_t} B_{t+1}^2$  from above by  $\sum_{k=-\infty}^K 2^k = 2^{K+1} \triangleq 2B_{T+1}^2$  in the same way as in Theorem 6 of Cesa-Bianchi et al. (2007).

Putting Equations (15) and (19) together, we get the PAC-Bayesian inequality

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \left( y_t - [\mathbf{u} \cdot \varphi(x_t)]_{B_t} \right)^2 \rho(d\mathbf{u}) + \frac{\mathcal{K}(\rho, \pi_\tau)}{\eta_{T+1}} \right\} + 8B_{T+1}^2,$$

which yields (12) by definition of  $\eta_{T+1} \triangleq 1/(8B_{T+1}^2)$ . The other PAC-Bayesian inequality (13), which is stated for non-truncated base forecasts, follows from (12) by the fact that truncation to  $B_t$  can only improve prediction if  $|y_t| \leq B_t$ . The remaining  $t$ 's such that  $|y_t| > B_t$  then just account for an overall additional term at most equal to  $\sum_{t:B_{t+1}>B_t}^T (2B_{t+1})^2 \leq 8B_{T+1}^2$ , which concludes the proof. ■

**Proof (of Proposition 4)** The proof follows the exact same lines as in Proposition 1 except that we apply Lemma 7 instead of Lemma 3. ■

### 3.3. A fully automatic algorithm

In the previous section, we proved that adaptation to  $B_y$  was possible. If we also no longer assume that a bound  $B_\Phi$  on the trace of the empirical Gram matrix is available to the forecaster, then one can use a doubling trick on the nondecreasing quantity

$$\gamma_t \triangleq \ln \left( 1 + \sqrt{\sum_{s=1}^t \sum_{j=1}^d \varphi_j^2(x_s)} \right)$$

and repeatedly run the algorithm SeqSEW $^*_\tau$  of the previous section for rapidly-decreasing values of  $\tau$ . This yields a sparsity regret bound with extra logarithmic multiplicative factors as compared to Proposition 4, but which holds for a fully automatic algorithm; see Theorem 8 below.

More formally, our algorithm SeqSEW $^*_*$  is defined as follows. The set of time rounds  $t = 1, 2, \dots$  is partitioned into regimes  $r = 0, 1, \dots$  whose final time instances  $t_r$  are data-driven. Let  $t_{-1} \triangleq 0$  by convention. We call regime  $r$ ,  $r = 0, 1, \dots$ , the sequence of time rounds  $(t_{r-1} + 1, \dots, t_r)$  where  $t_r$  is the first date  $t \geq t_{r-1} + 1$  such that  $\gamma_t > 2^r$ . At the beginning of regime  $r$ , we restart the algorithm SeqSEW $^*_\tau$  of the previous section with the parameter  $\tau = 1/(\exp(2^r) - 1)$ .

**Theorem 8** *Without requiring any preliminary knowledge at the beginning of the prediction game, the algorithm SeqSEW $^*_*$  satisfies, for all  $T \geq 1$  and all  $(x_1, y_1), \dots, (x_T, y_T) \in \mathcal{X} \times \mathbb{R}$ ,*

$$\begin{aligned} \sum_{t=1}^T (y_t - \hat{y}_t)^2 &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(x_t))^2 + 256 \left( \max_{1 \leq t \leq T} y_t^2 \right) \|\mathbf{u}\|_0 \ln \left( e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)} \right) \right. \\ &\quad \left. + 64 \left( \max_{1 \leq t \leq T} y_t^2 \right) A_T \|\mathbf{u}\|_0 \ln \left( 1 + \frac{\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} + \left( 1 + 38 \max_{1 \leq t \leq T} y_t^2 \right) A_T , \end{aligned} \tag{20}$$

where  $A_T \triangleq 2 + \log_2 \ln \left( e + \sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)} \right)$ .

**Proof** The proof relies on the application of Proposition 4 with  $\tau = 1/(\exp(2^r) - 1)$  on all regimes  $r$  visited up to time  $T$ . Summing the corresponding inequalities over  $r$  then concludes the proof. Due to lack of space, we refer the reader to the full version of this paper (Gerchinovitz, 2011, Theorem 1) for further details. ■

## 4. Adaptivity to the unknown variance in the stochastic setting

In this section we apply the algorithm SeqSEW to the regression model with random design. In this batch setting the forecaster is given at the beginning of the game  $T$  independent random copies  $(X_1, Y_1), \dots, (X_T, Y_T)$  of  $(X, Y) \in \mathcal{X} \times \mathbb{R}$  whose common distribution is unknown. We assume thereafter that  $\mathbb{E}[Y^2] < \infty$ ; the goal of the forecaster is to estimate the regression function  $f : \mathcal{X} \rightarrow \mathbb{R}$  defined by  $f(x) \triangleq \mathbb{E}[Y|X = x]$  for all  $x \in \mathcal{X}$ . We also set  $\|h\|_{L^2} \triangleq (\mathbb{E}[h(X)^2])^{1/2}$  for all measurable functions  $h : \mathcal{X} \rightarrow \mathbb{R}$  such that  $\mathbb{E}[h(X)^2] < \infty$ .

### 4.1. Algorithm and main result

Even if the whole sample  $(X_1, Y_1), \dots, (X_T, Y_T)$  is available at the beginning of the prediction game, we treat it in a sequential fashion. We run the algorithm SeqSEW $^*_\tau$  of Section 3.2 from time 1 to time  $T$  with  $\tau = 1/\sqrt{dT}$ . We then define our data-based regressor  $\hat{f}_T$  as the uniform average  $\hat{f}_T \triangleq \frac{1}{T} \sum_{t=1}^T f_t$  of the regressors  $f_t : \mathcal{X} \rightarrow \mathbb{R}$  sequentially built by the algorithm SeqSEW $^*_\tau$  as

$$\tilde{f}_t(x) \triangleq \int_{\mathbb{R}^d} [\mathbf{u} \cdot \boldsymbol{\varphi}(x)]_{B_t} p_t(d\mathbf{u}).$$

This technique is now quite standard in the machine learning community. Though we only state our risk bounds in expectation (which already improves on existing results in the stochastic setting), we refer to Kakade and Tewari (2009) to transform our results into risk bounds with large probability.

Note that, contrary to much prior work from the statistics community such as Catoni (2004) and Dalalyan and Tsybakov (2011), the regressors  $\tilde{f}_t : \mathcal{X} \rightarrow \mathbb{R}$  are tuned online. Therefore,  $\hat{f}_T$  does not depend on any prior knowledge on the unknown distribution of the  $(X_t, Y_t)$ ,  $1 \leq t \leq T$ , such as the unknown variance  $\mathbb{E}[(Y - f(X))^2]$  of the noise, the  $\|\varphi_j\|_\infty$ , or the  $\|f - \varphi_j\|_\infty$  (actually, the  $\varphi_j$  and the  $f - \varphi_j$  do not even need to be bounded in  $\ell^\infty$ -norm).

**Theorem 9** Assume that  $(X_1, Y_1), \dots, (X_T, Y_T) \in \mathcal{X} \times \mathbb{R}$  are independent random copies of  $(X, Y) \in \mathcal{X} \times \mathbb{R}$ , where  $\mathbb{E}[Y^2] < +\infty$  and  $\|\varphi_j\|_{L^2}^2 \triangleq \mathbb{E}[\varphi_j(X)^2] < +\infty$  for all  $j = 1, \dots, d$ . Then, the data-based regressor  $\hat{f}_T$  defined above satisfies

$$\begin{aligned} \mathbb{E}\left[\left\|f - \hat{f}_T\right\|_{L^2}^2\right] &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \|f - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 + 64 \frac{\mathbb{E}[\max_{1 \leq t \leq T} Y_t^2]}{T} \|\mathbf{u}\|_0 \ln \left(1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0}\right)\right\} \\ &\quad + \frac{1}{dT} \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + 32 \frac{\mathbb{E}[\max_{1 \leq t \leq T} Y_t^2]}{T}. \end{aligned}$$

**Proof** By Proposition 4 with  $\tau = 1/\sqrt{dT}$  and by definition of  $\tilde{f}_t$  above and  $\hat{y}_t \triangleq \tilde{f}_t(X_t)$  in Equation (10), we have, *almost surely*,

$$\begin{aligned} \sum_{t=1}^T (Y_t - \tilde{f}_t(X_t))^2 &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (Y_t - \mathbf{u} \cdot \boldsymbol{\varphi}(X_t))^2 + 64 \left( \max_{1 \leq t \leq T} Y_t^2 \right) \|\mathbf{u}\|_0 \ln \left( 1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{1}{dT} \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(X_t) + 32 \max_{1 \leq t \leq T} Y_t^2. \end{aligned}$$

Taking the expectations of both sides and applying Jensen's inequality straightforwardly concludes the proof (we refer the reader to the full version of this paper, Gerchinovitz 2011, Theorem 2 for more details).  $\blacksquare$

The above theorem can be used under several assumptions on the distribution of the output  $Y$ . We only discuss below its application to one important set of assumptions studied, e.g., in Dalalyan and Tsybakov (2011).

#### 4.2. Questions left open by Dalalyan and Tsybakov

Theorem 9 above provides answers to two questions left open in Dalalyan and Tsybakov (2011) when the regression function  $f$  is bounded and when the i.i.d. errors  $\varepsilon_t \triangleq Y_t - f(X_t)$  are subgaussian (conditionally on the  $X_t$ ) in the sense that, for some constant  $\sigma^2 > 0$ ,

$$\|f\|_\infty < +\infty \quad \text{and} \quad \mathbb{E}[e^{\lambda \varepsilon_1} \mid X_1] \leq e^{\lambda^2 \sigma^2 / 2} \quad \text{a.s.}, \quad \forall \lambda \in \mathbb{R}. \quad (21)$$

Under the above assumptions, we prove in Gerchinovitz (2011, Corollary 5 and Remark 8) that Theorem 9 above yields, for some universal constant  $C > 0$ , that for all  $T \geq 2$ ,

$$\begin{aligned} \mathbb{E} \left[ \left\| f - \hat{f}_T \right\|_{L^2}^2 \right] &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \|f - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 + 2C \left( \|f\|_\infty^2 + \sigma^2 \ln T \right) \frac{\|\mathbf{u}\|_0}{T} \ln \left( 1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{1}{dT} \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + \frac{C}{T} \left( \|f\|_\infty^2 + \sigma^2 \ln T \right). \end{aligned} \quad (22)$$

The above bound is of the same order (up to a  $\ln T$  factor) as the sparsity oracle inequality proved in Proposition 1 of Dalalyan and Tsybakov (2011). For the sake of comparison we state below with our notations (e.g.,  $\beta$  therein corresponds to  $1/\eta$  in this paper) a straightforward consequence of this proposition, which follows by Jensen's inequality and the particular<sup>5</sup> choice  $\tau = 1/\sqrt{dT}$ .

---

5. Proposition 1 of Dalalyan and Tsybakov (2011) may seem more general than Theorem 9 at first sight since it holds for all  $\tau > 0$ , but this is actually also the case for Theorem 9. The proof of the latter would indeed have remained true had we replaced  $\tau = 1/\sqrt{dT}$  with any value of  $\tau > 0$ . We however chose the reasonable value  $\tau = 1/\sqrt{dT}$  to make our algorithm parameter-free.

**Proposition 10 (A consequence of Prop. 1 of Dalalyan and Tsybakov 2011)**

Assume that  $\sup_{1 \leq j \leq d} \|\varphi_j\|_\infty < \infty$  and that the set of assumptions (21) above hold true. Then, for every  $R > 0$  and  $\eta \leq \left(2\sigma^2 + 2\sup_{\|\mathbf{u}\|_1 \leq R} \|\mathbf{u} \cdot \boldsymbol{\varphi} - f\|_\infty^2\right)^{-1}$ , the mirror averaging aggregate  $\hat{f}_T : \mathcal{X} \rightarrow \mathbb{R}$  defined in Dalalyan and Tsybakov (2011, Equations (1) and (3)) satisfies

$$\begin{aligned} \mathbb{E} \left[ \left\| f - \hat{f}_T \right\|_{L^2}^2 \right] &\leq \inf_{\|\mathbf{u}\|_1 \leq R-2d\tau} \left\{ \|f - \mathbf{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 + \frac{4\|\mathbf{u}\|_0}{\eta(T+1)} \ln \left( 1 + \frac{\sqrt{dT}\|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} \\ &\quad + \frac{4}{dT} \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + \frac{1}{\eta(T+1)}. \end{aligned}$$

We can now discuss the two questions left open by Dalalyan and Tsybakov (2011). Despite the similarity of the two bounds, the sparsity oracle inequality stated in Proposition 10 above only holds for vectors  $\mathbf{u}$  within  $\ell^1$ -balls of finite radii. The authors thus asked in Dalalyan and Tsybakov (2011, Section 4.2) whether it was possible to extend the infimum to the whole  $\mathbb{R}^d$  space. Our results show that, thanks to data-driven truncation, the answer is positive.

The second open question, which was raised in Dalalyan and Tsybakov (2011, Section 5.1, Remark 6), deals with the prior knowledge of the variance factor  $\sigma^2$  of the noise. The latter is indeed required by their algorithm for the choice of the inverse temperature parameter  $\eta$ . The authors thus asked whether adaptivity to  $\sigma^2$  was possible. Our sparsity oracle inequality (22) above provides a positive answer (up to a  $\ln T$  factor).

**Remark 11** Similar adaptivity results hold in the regression model with fixed design; see the full version of this paper (Gerchinovitz, 2011, Section 5.2). The framework of prediction of individual sequences thus seems to offer a unifying setting to address tuning issues both in the random and in the fixed design regression models.

## Acknowledgments

The author would like to thank Arnak Dalalyan, Gilles Stoltz, and Pascal Massart for their helpful comments and suggestions. The author acknowledges the support of the French Agence Nationale de la Recherche (ANR), under grant PARCIMONIE (<http://www.proba.jussieu.fr/ANR/Parcimonie>), and of the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886.

## References

- J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4):1591–1646, 2009.
- P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *J. Comp. Sys. Sci.*, 64:48–75, 2002.

- K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43(3):211–246, 2001.
- G. Biau, K. Bleakley, L. Györfi, and G. Ottucsák. Nonparametric sequential prediction of time series. *J. Nonparametr. Stat.*, 22(3–4):297–317, 2010.
- L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001.
- F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for regression learning. Technical report, 2004. Available at <http://arxiv.org/abs/math/0410214>.
- F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation and sparsity via  $\ell_1$  penalized least squares. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT'06)*, pages 379–391, 2006.
- F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.
- O. Catoni. *Statistical learning theory and stochastic optimization*. Springer, New York, 2004.
- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66(2/3):321–352, 2007.
- A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT'07)*, pages 97–111, 2007.
- A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2):39–61, 2008.
- A. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT'09)*, pages 83–92, 2009.
- A. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 2011. To appear. Available at <http://hal.archives-ouvertes.fr/hal-00461580/>.
- J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT'10)*, pages 14–26, 2010.
- Y. Freund, R. E. Schapire, Y. Singer, and M. K. Warmuth. Using and combining predictors that specialize. In *Proceedings of the 29th annual ACM Symposium on Theory of Computing (STOC'97)*, pages 334–343, 1997.
- S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. Technical report, 2011. Available at <http://arxiv.org/abs/1101.1057>.

- L. Györfi and G. Ottucsák. Sequential prediction of unbounded stationary time series. *IEEE Trans. Inform. Theory*, 53(5):1866–1872, 2007.
- L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.
- S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems 21 (NIPS'08)*, pages 801–808. 2009.
- J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *Proceedings of the 4th European Conference on Computational Learning Theory (EuroCOLT'99)*, pages 153–167, 1999.
- J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *J. Mach. Learn. Res.*, 10:777–801, 2009.
- G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of convergence for high-dimensional regression under  $\ell^q$ -ball sparsity. In *Proceedings of the 47th annual Allerton conference on communication, control, and computing (Allerton'09)*, pages 251–257, 2009.
- M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *J. Mach. Learn. Res.*, 9:759–813, 2008.
- S. Shalev-Shwartz and A. Tewari. Stochastic methods for  $\ell^1$ -regularized loss minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 929–936, 2009.
- V. Vovk. Competitive on-line statistics. *Internat. Statist. Rev.*, 69:213–248, 2001.
- L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, 2010.

GERCHINOVITZ

# Safe Learning: *bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity*

Peter Grünwald

PDG@CWI.NL

CWI Amsterdam and Mathematical Institute, Leiden University, The Netherlands

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We extend Bayesian MAP and Minimum Description Length (MDL) learning by testing whether the data can be substantially more compressed by a mixture of the MDL/MAP distribution with another element of the model, and adjusting the learning rate if this is the case. While standard Bayes and MDL can fail to converge if the model is wrong, the resulting ‘safe’ estimator continues to achieve good rates with wrong models. Moreover, when applied to classification and regression models as considered in statistical learning theory, the approach achieves optimal rates under, e.g., Tsybakov’s conditions, and reveals new situations in which we can penalize by  $(-\log \text{PRIOR})/n$  rather than  $\sqrt{(-\log \text{PRIOR})/n}$ .

## 1. Introduction

1. *Learning Theory; Predictor Models.* In much of statistical learning and machine learning theory, the goal is to learn, based on a set of observed data  $Z^n = (Z_1, \dots, Z_n)$ , a predictor  $\check{f}$  taken from some set of candidate prediction rules  $\mathcal{F}$ . Here each  $Z_i = (X_i, Y_i)$ , each  $X_i$  takes values in some set  $\mathcal{X}$ , each  $Y_i$  takes values in  $\mathcal{Y}$ , and  $\mathcal{F}$  is a set of functions  $f : \mathcal{X} \rightarrow \mathcal{Y}$ . The  $Z_i$  are assumed to be sampled i.i.d. according to some distribution  $P^*$  on  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ . The learned predictor  $\check{f}$  should have a small *generalization error* or *risk*, defined as  $\text{RISK}(f) := E^*[\text{LOSS}(Y, f(X))]$  where  $\text{LOSS}$  is some given loss function and here, as elsewhere in this paper,  $E^* = E_{(X,Y) \sim P^*}$  denotes joint expectation of  $(X, Y)$  over  $P^*$ . In a typical classification setting,  $\mathcal{Y} = \{0, 1\}$  and  $\text{LOSS}(y, \hat{y}) := |y - \hat{y}|$  is the 0/1-loss; in typical regression problems,  $\mathcal{Y} = \mathbb{R}$  and  $\text{LOSS}(y, \hat{y}) := (y - \hat{y})^2$  is the squared loss. Crucially, risk bounds are usually proved in worst-case settings, using only weak assumptions on  $P^*$ .

2. *Standard Statistics; Probability Models.* Here one models uncertainty by a statistical model, i.e. a set of probability distributions  $\mathcal{P}$ , and the goal is to learn a distribution  $\check{p}$  that is a good representation of the underlying distribution  $P^*$  from which the data  $Z^n$  are sampled. Here we focus on the case that  $Z^n$  are i.i.d.,  $Z_i = (X_i, Y_i)$  as above, and  $\mathcal{P}$  is a set of conditional distributions  $p(y | x)$ , identified by their mass functions (if  $\mathcal{Y}$  is finite/countable) or densities, and extended to  $n$  outcomes by independence. Witness papers such as *The Two Cultures* (Breiman, 2001), the difference between statistical/machine learning theory and standard statistics based on probability models is often regarded as fundamental. Here, I propose a first, preliminary, attempt at an overarching, single theory of learning, as embodied by a new ‘safe’ estimator. It is called ‘safe’ because, when applied to probability models  $\mathcal{P}$ , then, unlike standard Bayes and MDL, it is guaranteed to perform well in the often inevitable situation that ‘all models (elements of  $\mathcal{P}$ ) are wrong, yet some are useful’.

**Safe Estimation for Probability Models** For probability models  $\mathcal{P}$ , the safe estimator behaves similarly to the Bayesian MAP or two-part MDL estimator. Following Barron and Cover (1991), we define the  *$\kappa$ -two part estimator*, written as  $\ddot{p}_\kappa$ , as a generalization of the MAP/MDL estimator, as follows: fix some prior distribution  $w$  and some  $\kappa > 0$ . For each  $x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n$ ,  $\ddot{p}_\kappa$  is defined\* as the  $p \in \mathcal{P}$  achieving<sup>1</sup>

$$\min_{p \in \mathcal{P}} \{ -\kappa \log w(p) - \log p(y^n | x^n) \}. \quad (1)$$

When  $\kappa \geq 1$ , then, via the Kraft inequality, (1) can be thought of as the number of bits needed to encode  $Y^n$  given  $X^n$  in a two-stage code;  $-\kappa \log w(p)$  is the codelength needed to encode  $p$ , and acts as a complexity penalty.  $-\log p(y^n | x^n)$  is the codelength of the data  $y^n$  when encoded with the help of  $p$  and  $x^n$ . To get good convergence rates, one needs to set  $\kappa > 1$  (Zhang, 2006); while any fixed  $\kappa > 1$  will do, for ‘standard 2-part MDL’ one takes  $\kappa = 2$  which is mathematically convenient (Barron and Cover, 1991). In contrast, the safe estimator is defined (in Section 2, Eq. (9)) as  $\check{p}_{\text{SAFE}} = \ddot{p}_{2\check{\kappa}_{\text{SAFE}}}$  where  $\check{\kappa}_{\text{SAFE}}$  is not fixed but determined by the data.  $\check{\kappa}_{\text{SAFE}}$  will be a small constant  $\geq 2$ , unless the data indicate that the model is misspecified (wrong). Whereas ordinary Bayesian and MDL approaches can fail to converge in this case (Example 7 below), the safe estimator continues to perform well in the following sense: suppose that data  $Z^n$  are i.i.d.  $\sim P^*$ , as above, where for each  $x$ ,  $P^*(Y = \cdot | X = x)$  has conditional density  $p^*(\cdot | x)$ . Let  $q$  be the best approximation within  $\mathcal{P}$  of  $p^*$  in terms of Kullback-Leibler (KL) divergence. Then the KL divergence between  $\check{p}_{\text{SAFE}}$  and  $p^*$  converges to the KL divergence between  $q$  and  $p^*$  at fast rates. To express this formally, for any two conditional densities  $p$  and  $p'$ , we define\* the *generalized KL divergence* (Grünwald, 2007) relative to  $P^*$  as

$$D^*(p' \| p) := E^*[-\log p(Y | X) + \log p'(Y | X)].$$

Then, for  $q$  satisfying  $\inf_{p \in \mathcal{P}} D^*(p^* \| p) = D^*(p^* \| q)$  we prove, under suitable regularity conditions, in Theorem 1 in combination with Theorem 3 below that  $D(p^* \| \check{p}_{\text{SAFE}}) - D(p^* \| q) \rightarrow 0$ , or equivalently,  $D^*(q \| \check{p}_{\text{SAFE}}) \rightarrow 0$ , in probability at fast rates.

**Safe Estimation for Predictor Models** In our overarching approach, all models are formally defined as sets of probability distributions. Predictor models  $\mathcal{F}$  are “transformed” into corresponding probability models  $\mathcal{P}_{\mathcal{F}} := \{p_f \mid f \in \mathcal{F}\}$  by a standard transformation (called ‘entropification’ and extensively motivated from an MDL perspective by Grünwald (1999)): for each  $f \in \mathcal{F}$ ,

$$p_f(y | x) := \frac{1}{Z(\beta)} e^{-\beta \text{LOSS}(y, f(x))}, \quad p_f(y^n | x^n) := \prod_{i=1}^n p_f(y_i | x_i). \quad (2)$$

Here  $Z(\beta) = \int_{y \in \mathcal{Y}} \exp(\beta \text{LOSS}(y, f(x))) dy$  is a normalization factor (if  $\mathcal{Y}$  is finite/countable, then here, as everywhere else in this paper, the integral should be replaced by a sum). In this preliminary study, we set  $\beta$  to some fixed value, say, 1 (but see Section 6). For the squared loss,  $Z(\beta)$  does not depend on  $f(x)$ ; if we set  $\beta = 1/2\sigma^2$ , we see that (2) expresses that  $Y$  is Gaussian with mean  $f(X)$  and variance  $\sigma^2$ . For the 0/1-loss,  $Z(\beta)$  does not

---

1. Distracting aspects of proofs (such as showing that the minimum of a function exists) have been omitted in this paper, but will be provided in the journal version. Such details are marked by a \*, such as here\*.

depend on  $f(x)$  either; loss functions for which  $Z(\beta)$  depends on  $f$  are handled as described under Eq.(7) below. Taking logarithms in (2), we then get that the *excess risk* of any  $f$  as compared to any  $g$  is a linear function of the generalized KL divergence of the corresponding distributions:

$$\text{EXCESS-RISK}(g\|f) = \text{RISK}(f) - \text{RISK}(g) = E^*[\text{LOSS}(Y, f(X)) - \text{LOSS}(Y, g(X))] = \frac{1}{\beta} D^*(p_g\|p_f). \quad (3)$$

Now let  $g$  be such that  $\text{RISK}(g) = \inf_{f \in \mathcal{F}} \text{RISK}(f)$ . Even if  $\mathcal{F}$  is a good ‘model’, i.e.  $\text{RISK}(g)$  is small, the corresponding model  $\mathcal{P}_{\mathcal{F}}$  will typically be misspecified (e.g. in the squared loss case, the ‘true’ noise may not be Gaussian at all). Since the safe estimator is immune to this problem, we can safely apply it to the model  $\mathcal{P}_{\mathcal{F}}$ . Then Theorems 1 and 3 show that the excess risk  $\text{EXCESS-RISK}(g\|\check{f}_{\text{SAFE}})$  converges to 0 at rates that are in many cases optimal; here  $\check{f}_{\text{SAFE}} := f$  for the  $f$  with  $\check{p}_{\text{SAFE}} = p_f$ . Thus, by the construction (2), convergence in generalized KL-divergence becomes equivalent to convergence in the loss function of interest.

**The Role of Convexity** Our starting point is the known fact that ‘standard MDL still works’, i.e., (broadly speaking),  $D^*(q\|\check{p}_2) \rightarrow 0$  at the appropriate rate if the closure  $\langle \mathcal{P} \rangle$  (suitably defined as in (10) below) of the model  $\mathcal{P}$  is *convex* (Li, 1999, Theorem 5.5); see also (Kleijn and van der Vaart, 2006). Our first observation is that, even if  $\langle \mathcal{P} \rangle$  is not convex, then as long as we have the weaker condition

$$\inf_{p \in \mathcal{P}} D^*(p^*\|p) = \inf_{p \in \text{CONVEX-HULL}(\mathcal{P})} D^*(p^*\|p), \quad (4)$$

we still get that  $D^*(q\|\check{p}_2) \rightarrow 0$  at the right rates. Now define, for  $\eta \leq 1$ , the model  $\mathcal{P}^{(\eta)} := \{p^{(\eta)} \mid p \in \mathcal{P}\}$ , where  $p^{(\eta)}(y \mid x) \propto (p(y \mid x))^{\eta}$  (for predictor models  $\mathcal{P}_{\mathcal{F}}$ , this corresponds to replacing  $\beta$  in (2) by  $\eta \cdot \beta$ ; a precise definition is beneath (7) below). Our second insight is that, even if (4) does not hold for  $\mathcal{P}$ , then still, for all  $\eta$  no greater than some critical value  $\eta_{\text{CRIT}}$ , (4) will actually hold with  $\mathcal{P}$  replaced by  $\mathcal{P}^{(\eta)}$  and thus ‘standard MDL still works’ for  $\mathcal{P}^{(\eta)}$ . The third insight is that the MDL estimator  $\ddot{p}_{2\kappa}$  for model  $\mathcal{P}$  with  $\kappa = 1/\eta_{\text{CRIT}}$  is essentially equivalent to the standard MDL estimator  $\ddot{p}_2^{(\eta)}$  for model  $\mathcal{P}^{(\eta_{\text{CRIT}})}$ ; indeed, we will prove (implicitly in Theorem 3) that the MDL estimator  $\ddot{p}_{2\kappa}$  with  $\kappa = 1/\eta_{\text{CRIT}}$  leads to good results for the model  $\mathcal{P}$ . The fourth, and main, insight is that, for any given  $\eta$ , we can *test* whether  $\eta \leq \eta_{\text{CRIT}}$ , i.e. whether (4) is the case for  $\mathcal{P}^{(\eta)}$ , by looking at the observed data: *essentially, the likelihood of the data according to  $\ddot{p}_2^{(\eta)}$  will be significantly smaller than the likelihood according to a 2-component mixture of  $\ddot{p}_2^{(\eta)}$  and another, suitably chosen  $p \in \mathcal{P}$ , if and only if (4) does not hold.* The minus logarithm of this discrepancy is measured in terms of a function *CONV-LACK*, a key concept of this paper, defined formally in (8). The safe estimator is defined as  $\check{p}_{\text{SAFE}} = \ddot{p}_{2\check{\kappa}_{\text{SAFE}}}$ , i.e. it is the  $2\check{\kappa}_{\text{SAFE}}$ -two-part estimator where  $\check{\kappa}_{\text{SAFE}} = 1/\eta$ , and  $\eta$  is determined by the data: it is effectively set to the largest value for which *CONV-LACK* is small, i.e. for which we cannot fit the data better by a two-component mixture.

**Overview of Results** In Section 2 we formally define *CONV-LACK* and  $\check{p}_{\text{SAFE}}$ . In Section 3 and 4,  $\check{p}_{\text{SAFE}}$  will be shown to converge at optimal rates up to log factors in a variety of

situations, as illustrated by examples in Section 4. Convergence of  $\check{p}_{\text{SAFE}}$  is shown in two stages: Theorem 1 bounds  $D^*(q\|\check{p})$  for arbitrary estimators  $\check{p}$  in terms of a ‘redundancy’ term  $\text{RED}$  and the  $\text{CONV-LACK}$  term. The redundancy term also shows up in classical MDL analyses and tends to 0 if we set  $\check{p}$  to a two-part estimator. Theorem 3 shows (essentially) that if  $\check{p}$  is set to  $\check{p}_{2\kappa}$  with  $\kappa \geq 1/\eta_{\text{CRIT}}$ , then the  $\text{CONV-LACK}$ -term is small as well. Taken together, Theorem 1 and 3 imply that  $\check{p}_{\text{SAFE}}$  converges (a) at the right rates if the model is correct or convex; and (b) also if the model is “incorrect in the worst possible manner”, and finally, (c) also if the model is a classification model, i.e.  $\mathcal{P} = \mathcal{P}_{\mathcal{F}}$  for some set of classifiers  $\mathcal{F}$ , and a Tsybakov condition holds for  $\mathcal{F}$ .

In case (a),  $\eta_{\text{CRIT}} = 1$ . As shown in Example 4, if the model is in fact correct ( $p^* = q$ ) we get the same bound on  $D^*(p^*\|\check{p}_{\text{SAFE}})$  as the bounds obtained on  $D^*(p^*\|\check{p}_2)$  by Barron and Cover (1991), but with a larger constant factor — this is the price we have to pay for using a method that still works if the model is incorrect in a situation in which the model in fact, turns out to be correct. In the special case that  $w(p^*) = w(q) > 0$ ,  $D^*(q\|\check{p}_{\text{SAFE}})$  will tend to 0 as  $O((\log n)/n)$ . In case (b),  $\eta_{\text{CRIT}}$  may be as small as  $C/\sqrt{n}$  for some  $C > 0$ . Example 6 shows that  $D^*(q\|\check{p}_{\text{SAFE}})$  may then tend to 0 at rate as slow as  $(\log n)/\sqrt{n}$ , a worst-case bound familiar from the statistical learning literature. In case (c),  $\eta_{\text{CRIT}} \asymp n^{-(1-\nu)/(2-\nu)}$  for some  $0 \leq \nu \leq 1$  and the convergence rate depends on  $\nu$ . In the special case that  $w(q) > 0$ ,  $D^*(q\|\check{p}_{\text{SAFE}})$  will tend to 0 as  $O((\log n)n^{-1/(2-\nu)})$ ; see above Example 7. The examples illustrate the generality of  $\check{p}_{\text{SAFE}}$ , capturing both the common asymptotics for density estimation if the model is correct and for statistical classifier learning under the celebrated Tsybakov condition.  $\check{p}_{\text{SAFE}}$  also gives a new interpretation of the difference in complexity penalties prescribed by MDL/Bayes on the one hand and learning theory approaches (such as Structural Risk Minimization and PAC-Bayesian methods) on the other: since the  $\mathcal{P}_{\mathcal{F}}$  constructed from a predictor model  $\mathcal{F}$  is in general nonconvex, we may have  $\eta_{\text{CRIT}} \ll 1$ , and the standard MDL penalties become too small.

A third main result is Theorem 2, which gives a new PAC-Bayes style empirical generalization bound in which, if we are ‘lucky’ on the observed data, the codelength  $(-\log w(\check{p}_{\text{SAFE}}))$  only appears in an  $O(1/n)$  rather than an  $O(1/\sqrt{n})$  term, even if the empirical error of the learned classifier is not close to 0. As such it provides another step in the race to “root out the square root” that characterizes so much of the work on classification bounds in learning theory.

**Related Work — Learning the Learning Rate** The larger  $\eta = 1/\kappa$ , the more influence the data has on the chosen hypothesis  $\check{p}_{2\kappa}$ . For predictor models, the same holds for the  $\beta$  appearing in (2). Thus  $\eta \cdot \beta$  may be viewed as the learning rate. A straightforward approach to learn it from data is to fix  $\eta$  and instead pick the  $\beta$  in (2) that minimizes overall description length of the data, as suggested by Grünwald (1999). Soon after publishing that paper, it became clear to me that this does not work (this was shown formally in (Grünwald and Langford, 2007)), and I started looking for an estimator that performs as well as if the optimal learning rate  $\eta_{\text{CRIT}}$  had been known in advance. The safe estimator does achieve this goal, thus ending a twelve-year long search. In a sense,  $\check{p}_{\text{SAFE}}$  learns the optimal learning rate. Note however that we cannot prove that the  $\check{\kappa}_{\text{SAFE}}$  selected by the safe estimator is equal or close to  $1/\eta_{\text{CRIT}}$ ; we can only prove that it leads to the same asymptotic performance bounds.

A transformation similar to (2) is done in PAC-Bayesian methods (McAllester, 2003), where Bayesian averages of  $p_f$  are viewed as Gibbs distributions. Our approach is similar, yet closer in spirit to standard Bayes and MDL — There may be some relation with the advanced PAC-Bayesian analyses of Audibert (2004); Catoni (2007), who provide algorithms for predictor models that learn a learning rate (similar to  $\eta = 1/\kappa$ ) determined by the amount of ‘disagreement’ on the input data  $(X_1, \dots, X_n)$  between the chosen predictor  $\check{f}$  and other predictors in  $\mathcal{F}$ . It would be interesting to study the connections further.

Finally, our approach can (broadly) be seen as equipping (a form of) Bayesian inference with a frequentist test, and adjusting the priors if the test indicates that the model is misspecified. Such an idea was already suggested in broad terms by other researchers, e.g. Dawid (1982). It can also be viewed as equipping (a form of) MDL with a randomness test (can we compress the data more by stepping outside the model?), an idea that goes back to the Kolmogorov complexity roots of MDL.

## 2. The Safe Estimator

**Preliminaries** A *probabilistic model*  $\mathcal{P}$  is a *countable* set of conditional distributions on  $\mathcal{Y}$  given  $\mathcal{X}$ , identified with their mass functions (in case  $\mathcal{Y}$  is finite or countable) or otherwise their densities relative to Lebesgue measure, which we assume to exist;  $\mathcal{X}$  can be arbitrary. We allow the  $p \in \mathcal{P}$  to be defective (sum to less than one). That is,  $p(y | x)$  can be any function such that, for all  $x$ , for all  $y$ ,  $p(y | x) \geq 0$  and  $\sum_{y \in \mathcal{Y}} p(y | x) \leq 1$ . We extend  $p$  to  $n$  outcomes by independence, i.e.  $p(y^n | x^n) := \prod_{i=1}^n p(y_i | x_i)$ . For given  $z^n = (x_1, y_1), \dots, (x_n, y_n)$ ,  $z^n$  is shorthand for  $y^n | x^n$ , i.e.  $p(z^n) = p(y^n | x^n)$ . All logarithms are to base  $e$ .

Crucially, the models  $\mathcal{P}$  we are to consider, though countable, will usually represent very “large”, “complex” sets of distributions, which may be thought of as dense (in the information closure sense, see Section 3) subsets of an even larger, “nonparametric”  $\bar{\mathcal{P}}$  with  $\mathcal{P} \subset \bar{\mathcal{P}}$ : for example, we may consider the set  $\bar{\mathcal{P}}$  of all Gaussian mixtures with an arbitrary number of components, and then define  $\mathcal{P}$  as the subset of all  $p \in \bar{\mathcal{P}}$  with rational-valued means and mixture coefficients.

We may fix a probability mass function  $w$  on  $\mathcal{P}$ , which we shall think of as the *prior distribution* on  $\mathcal{P}$ . An *estimator* at sample size  $n$  is a function  $\check{p} : \mathcal{Z}^n \rightarrow \mathcal{P}$  that maps each possible sequence of observations  $z^n = (x_1, y_1), \dots, (x_n, y_n) \in \mathcal{Z}^n$  to some  $p \in \mathcal{P}$ . Following e.g. Barron and Cover (1991), the notation  $\check{p}(Z^n)$  denotes the density of the observed data  $Z^n$  according to the  $\check{p}$  that was selected (estimated) based on the same data  $Z^n$ .

**Conditions** We only consider combinations of  $(\mathcal{P}, P^*)$  and prior  $w$  for which **(A)** for each  $(x, y) \in \mathcal{Z}$ , there is a  $p \in \mathcal{P}$  such that  $w(p)p(y | x) > 0$ . We also assume **(B)** that for some finite integer  $L_{\max} > 0$ , for all  $n$ , all  $z^n \in \mathcal{Z}^n$ ,  $-\log w(\check{p}_2) \leq nL_{\max}$ . Hence the codelength of the 2-MDL estimator is of order no larger than  $n$ . This assumption is innocuous, since it can always be satisfied by adding one or a few distributions to  $\mathcal{P}$  (proof sketch in appendix). Finally, let

$$V = V(\mathcal{P}, P^*) := \text{ess sup}_{Z^n} \sup_{p, p' \in \mathcal{P}} \frac{p(Z^n)}{p'(Z^n)}. \quad (5)$$

We assume **(C)**  $1 < V$  and **(D)**  $V < \infty$ . We may think of  $V(\mathcal{P}, P^*)$  as the maximum ratio between the density of  $z_i = y | x$  assigned by different  $p \in \mathcal{P}$ , where the maximum is over

all  $(x, y)$  in the support of  $P^*$ . In case  $P^*$  has full support, the essential supremum can be replaced by the standard supremum. Assumptions (A), (B) and (C) are harmless; (D) is further discussed in Section 6.

**Mixing** The safe estimator is the  $\check{\kappa}_{\text{SAFE}}$ -two part estimator, with  $\check{\kappa}_{\text{SAFE}}$  determined by the data. To find  $\check{\kappa}_{\text{SAFE}}$ , we test, for each fixed  $\kappa$ , whether we can get a better fit of the data/additionally compress the data by a convex combination of  $\check{p}_\kappa$  with a single other distribution  $p' \in \mathcal{P}$ . Of course, since  $\mathcal{P}$  may be infinite and arbitrary, it may be the case that, no matter what data we observe, there is *always* some  $p \in \mathcal{P}$  such that a convex combination of  $\check{p}_\kappa$  and  $p$  gives a substantially better fit to the data. This problem, it turns out, can be addressed by only looking at distributions  $p$  with prior mass not much smaller than  $w(\check{p}_\kappa)$ : specifically, we require  $-\log w(p) \leq \lceil -\log w(\check{p}_\kappa) \rceil$ , where  $\lceil \cdot \rceil$  denotes rounding up.

To formalize this, for any  $p, p' \in \mathcal{P}$  and any  $\lambda \in [0, 1]$ , we define the mixture distribution  $\text{MIX}(p, p', \lambda)$  as  $(1 - \lambda)p + \lambda p'$ , so that for a single outcome  $z$ ,  $\text{MIX}(p, p', \lambda)(z) := (1 - \lambda)p(z) + \lambda p'(z)$  (note the somewhat special notation).  $\text{MIX}$  is extended to  $n$  outcomes by independence:

$$\text{MIX}(p, p', \lambda)(Z^n) := \prod_{i=1}^n ((1 - \lambda)p(z_i) + \lambda p'(z_i)). \quad (6)$$

Our test is defined in terms of how much better fit can be achieved by the best-fitting convex combination of this form. To this end, for an arbitrary estimator  $\check{p}$ , we let

$$\text{SUPMIX}(\check{p})(Z^n) := \sup_{p \in \mathcal{P}: -\log w(p) \leq \lceil -\log w(\check{p}) \rceil, \lambda \in [0, 1]} \text{MIX}(p, \check{p}, \lambda)(Z^n). \quad (7)$$

Let  $Z(\eta) := \sup_{x \in \mathcal{X}, p \in \mathcal{P}} \int_{y \in \mathcal{Y}} (p(y|x))^{\eta} dy$ . For each  $p \in \mathcal{P}$ , and each  $\eta \in \mathbb{R}$  with  $Z(\eta) < \infty$ , we define  $p^{(\eta)}(y|x) := (p(y|x))^{\eta}/Z(\eta)$ . Note that the  $p^{(\eta)}$  all represent distributions which, in general, may be defective, even if  $p$  is not. We define  $\mathcal{P}^{(\eta)} := \{p^{(\eta)} \mid p \in \mathcal{P}\}$ . Since, in all our equations, every occurrence of a density  $p^{(\eta)}$  will actually be as a *ratio* of two densities, i.e.  $p^{(\eta)}(Z^n)/q^{(\eta)}(Z^n)$ , we can safely write  $p^\eta$  instead of  $p^{(\eta)}$  everywhere without affecting the results. This is what we do below. However, for interpreting our results it is useful to think of the  $p^{(\eta)}$  as densities. For predictor models,  $\eta$  corresponds to  $\beta$  in (2); but see Section 6.

**Convexity Lack** We define the *convexity lack* of an arbitrary estimator  $\check{p}$  on data  $Z^n$  as

$$\text{CONV-LACK}(\eta, \check{p}) = -\frac{c_\eta}{\eta} \log \frac{\check{p}^\eta(Z^n)w^2(\check{p})}{\text{SUPMIX}(\check{p}^\eta)(Z^n)}, \quad (8)$$

where  $c_\eta = 1 + C_\eta C'_\eta$ , and  $C_\eta = 2 + 2\eta \log V$  and  $C'_\eta = 2V^{2\eta}$ . The rationale behind these values will become clear in Theorem 1 and Lemma 9 below. CONV-LACK is a measure of how many more bits are needed to encode the data using a two-part code (with  $\kappa = 2$ ) based on  $\check{p}^\eta$  (the numerator in the logarithm) as compared to the number of bits needed by the two-component mixture that provides the best fit (smallest codelength) with hindsight (the denominator). The larger this number, the more could have been gained by modelling the data with the convex hull of  $\mathcal{P}$  rather than just  $\mathcal{P}$ . Our main insight (see Theorem 3 below) is that, if  $\eta \leq \eta_{\text{CRIT}}$  ( $\eta_{\text{CRIT}}$ , introduced in Section 1, is formally defined below), then CONV-LACK is guaranteed to be small. This suggests to test various values of  $\kappa$ ,

and define the safe estimator as the  $\kappa$ -two part estimator for the smallest value of  $\kappa$  for which  $\text{CONV-LACK}(1/\kappa, \ddot{p}_\kappa)$  is below some fixed threshold. Here we opt for the essentially equivalent, but mathematically more convenient option to simply *add* CONV-LACK as an additional penalty to the codelength of the two-part estimator:

**Safe Estimation** Let  $\kappa_{\max} = \lceil \sqrt{n}/(2 \log V) \rceil$  (this value is motivated below Lemma 5). The *safe estimator*  $\check{p}_{\text{SAFE}}$  is defined as the  $2\kappa$ - two-part estimator for the  $\kappa \in \{1, 2, \dots, \kappa_{\max}\}$  that minimizes

$$-\log \ddot{p}_{2\kappa}(Z_1^n) - 2\kappa \log w(\ddot{p}_{2\kappa}) + \text{CONV-LACK}(\kappa, \ddot{p}_{2\kappa}). \quad (9)$$

This is just the formula for  $\ddot{p}_{2\kappa}$ , but with the term  $\text{CONV-LACK}(\kappa, \ddot{p}_{2\kappa})$  added. Here we establish that  $\check{p}_{\text{SAFE}}$  has good theoretical properties; whether it is useful in practice is discussed in Section 6.

### 3. Generalization Bounds for the Safe Estimator

**Preliminaries** We define the (generalized) *information closure* of  $\mathcal{P}$  (Barron and Cover, 1991) as

$$\langle \mathcal{P} \rangle := \{p' \mid \text{for some } P^*, \inf_{p \in \mathcal{P}} D^*(p^* \| p) = D^*(p^* \| p')\}, \quad (10)$$

where  $p'$  ranges over *all* conditional densities for  $Y$  given  $X$  (i.e.  $p'$  is not necessarily in  $\mathcal{P}$ ), and  $P^*$  ranges over *all* distributions on  $\mathcal{Z}$  that have some conditional density; we denote the density corresponding to  $P^*$  by  $p^*(y | x)$ . Henceforth we assume that data  $Z^n$  are sampled from a  $P^*$  that admits such a  $p^*$ . We also assume that  $P^*$  and  $\mathcal{P}$  are such that there exists a *best-approximating*  $q$ , defined as a  $q$  such that:

$$D^*(p^* \| q) = \inf_{p \in \mathcal{P}} D^*(p^* \| p) \text{ and } V(\mathcal{P} \cup \{q\}, P^*) = V(\mathcal{P}, P^*). \quad (11)$$

From now on, for given  $(\mathcal{P}, P^*)$ , we fix a particular best-approximating density once and for all and keep denoting it by  $q$ . We must have  $q \in \langle \mathcal{P} \rangle$ , and  $D^*(q \| p) \geq 0$  for all  $p \in \mathcal{P}$ . Our assumption that  $p^*$  and  $q$  exist simplifies the formulation of our theorems. Still, all our results continue to hold, with appropriately generalized definitions, if no such  $q$  or  $p^*$  exist\*. In the well-specified case, with  $q = p^* \in \langle \mathcal{P} \rangle$ , we trivially have that, for  $\eta = 1$ ,

$$\text{For all } p \in \mathcal{P}: E^* \left[ \left( \frac{p(Z_1)}{q(Z_1)} \right)^\eta \right] \leq 1, \text{ or equivalently, } d_\eta^*(q \| p) := -\frac{1}{\eta} \log E^* \left[ \left( \frac{p(Z_1)}{q(Z_1)} \right)^\eta \right] \geq 0, \quad (12)$$

as is seen by writing out the expectation in full and substituting  $q$  by  $p^*$ . Here  $d_\eta^*$  is the generalized Rényi divergence (Li, 1999); by a result of Li (1999), repeated as Proposition 15 in Section 5, if (12) holds then  $d_{\eta/2}^*$  may be viewed as a proxy for the generalized KL divergence, since then, for all  $p \in \mathcal{P}$ ,  $D^*(q \| p) \leq C_\eta d_{\eta/2}^*(q \| p) \leq C_\eta D^*(q \| p)$ , where  $C_\eta = 2 + 2\eta \log V$  is a constant. This is a key idea for our proofs. Classical theorems on two-part MDL inference for the well-specified case (Barron and Cover, 1991; Zhang, 2006; Grünwald, 2007) invariably make use of (12) at some point in the proofs; so do classical results on Bayesian consistency (Doob, 1949), in which (12) is used to establish that  $\{p(Z_1^n)/q(Z_1^n)\}_{n=1,2,\dots}$  forms a martingale. It can be shown (Li, 1999; Kleijn and van der Vaart, 2006) that (12) still holds for  $\eta = 1$  if  $\langle \mathcal{P} \rangle$  is convex, or, more generally, if (4) holds. This is the fundamental

reason why the MDL and Bayesian convergence bounds still hold in that setting. In fact, (4) with  $\mathcal{P}^{(\eta)}$  in the role of  $\mathcal{P}$  is equivalent to (12), as follows from Lemma 9 in Section 5 (proof sketch in Appendix). If (4) does not hold for  $\eta = 1$ , then (12) does not hold for  $\eta = 1$  and MDL and Bayes may not converge (Example 7 below). Luckily, for many types of  $\mathcal{P}$ , one can still show that (12) holds for some  $\eta < 1$ . Thus it makes sense to define the *critical exponent*  $\eta_{\text{CRIT}}$  as the largest value of  $\eta$  such that, for all  $p \in \mathcal{P}$ , (12) holds. It is useful to generalize the idea slightly. We define, for  $u \geq 0$ , the *u-critical exponent*  $\eta_{\text{CRIT}}(u)$  as

$$\eta_{\text{CRIT}}(u) := \sup \left\{ \eta \leq 1 : \text{for all } p \in \mathcal{P}, \quad \log E^* \left[ \left( \frac{p(Z_i)}{q(Z_i)} \right)^\eta \right] \leq \frac{u}{n} \right\}. \quad (13)$$

$\eta_{\text{CRIT}}(0)$  is just the critical value as defined before. In Section 1 we cheated a little, writing  $\eta_{\text{CRIT}}$  for  $\eta_{\text{CRIT}}(u)$  for the  $u$  which gives the best bounds; see below. Whenever we write ‘WHP’ (‘with high probability’), we really mean ‘for all  $K \geq 0$ , with  $P^*$ -probability at least  $1 - e^{-K}$ ,  $Z^n$  satisfies...’.

**Theorem 1 (Oracle Bound)** *Let  $(\mathcal{P}, P^*)$  and  $w$  satisfy conditions (A)-(D) of Section 2 with  $V$  as in (5),  $q$  as in (11),  $\eta_{\text{CRIT}}$  as in (13) and  $\kappa_{\max}$  as above (9). Let  $\check{p}$  be an arbitrary estimator. Let  $Z^n \sim P^*$ . WHP, uniformly for all  $\eta \in \{1, 1/2, 1/3, 1/4, \dots, 1/\kappa_{\max}\}$ , all  $u \in \{0, 1, \dots, nL_{\max}\}$ , we have*

$$D^*(q \|\check{p}) \leq \frac{C_\eta}{n} \left( \text{RED}(2/\eta, \check{p}) + \text{CONV-LACK}(\eta, \check{p}) + \frac{u}{\min\{\eta, \eta_{\text{CRIT}}(u)\}} + R \right), \quad (14)$$

where  $C_\eta = 2 + 2\eta \log V$ . The term CONV-LACK is given by (8). The term RED is given by

$$\text{RED}(2/\eta, \check{p}) = -\log \frac{w^{2/\eta}(\check{p})\check{p}(Z_i^n)}{q(Z_i^n)} = -\frac{2}{\eta} \log w(\check{p}) - \log \check{p}(Z_i^n) + \log q(Z_i^n). \quad (15)$$

The remainder term\* is  $R = O\left(\frac{K + \log n + \log(2 + [-\log w(\check{p})])}{\min\{\eta, \eta_{\text{CRIT}}(u)\}}\right)$ .

RED stands for ‘redundancy’. It can be interpreted as the extra number of nats needed to code the data using a two-part code based on  $\check{p}$  (with  $\kappa = 2/\eta$ ), as compared to the code based on the best-approximating  $q$ , and under mild conditions  $\text{RED}/n$  will tend to 0, WHP (Example 4). For predictor models as in (2),  $\check{p} = p_{\check{f}}$  and  $q = p_g$ , and then  $\text{RED}/n$  can be thought of as the difference in empirical risk between  $\check{f}$  and the optimal  $g$ , ‘penalized’ by  $-(2/\eta n) \log w(\check{p})$ . The RED and CONV-LACK terms depend on the data; the third term in (14) becomes 0 if we set  $u = 0$ ; for the possibility  $u > 0$  see Example 6. The fourth term is a remainder term which does not depend on the data except for the term  $\log(2 - \log w^{1/\eta}(\check{p}))$  which, by assumption (B), if  $\check{p}$  is a 2-part or the safe estimator, is bounded by  $O(\log n)$ .

We can now motivate the definition  $\check{p}_{\text{SAFE}}$ : among all two-part estimators, it is the one minimizing (14), ignoring the remainder term. To see this, note that the third term in (14) does not depend on the data, and the redundancy term can be written as the sum of two terms plus a term  $\log q(Z_i^n)$  that does not depend on our choice of estimator  $\check{p}$ . Thus,  $\check{p}_{\text{SAFE}}$  minimizes an upper bound on the KL divergence between  $\check{p}$  and the best-approximating  $q$ .

Let us compare (14) to the bounds on the standard two-part MDL estimator as derived by Barron and Cover (1991) under the assumption that  $q = p^*$ . One of their main results

(Theorem 4, as later strengthened by Zhang (2006) and (Grünwald, 2007, Section 15.3)) implies that, for all  $\kappa \geq 1$ ,

$$E_{Z^n \sim P^*} [ D^*(p^* \| \ddot{p}_{2\kappa}) ] \leq \frac{C}{n} E_{Z^n \sim P^*} [ \text{RED}(2\kappa, \ddot{p}_{2\kappa}) ], \quad (16)$$

where  $C \leq 2 + 2 \log V$  as above. This provides a frequentist justification for the 2-part MDL estimator  $\ddot{p}_{2\kappa}$  with  $\kappa \geq 1$ , since  $\ddot{p}_{2\kappa}$  is in fact defined as the  $p \in \mathcal{P}$  that minimizes  $\text{RED}(2\kappa, p)$ . Apart from the ‘in-expectation’ rather than ‘in-probability’ formulation, the “only” relevant difference to our result is the CONV-LACK term, which appears in (14) because we do not require a correct model. In Section 4 we show that for many combinations of  $P^*$  and  $\mathcal{P}$ , the CONV-LACK term will be small, WHP, for  $\check{p} = \check{p}_{\text{SAFE}}$ . In such cases Theorem 1 states something similar to Barron and Cover’s result ( $D^*(q \| \check{p}) \rightarrow 0$ ), but without the often unrealistic requirement that  $p^* \in \langle \mathcal{P} \rangle$ .

**Theorem 2 (Empirical Bound)** *Assume the notations and conditions of Theorem 1. WHP, uniformly for all  $\eta \in \{1, 1/2, 1/3, \dots, 1/\kappa_{\max}\}$ , we have,*

$$E^*[-\log \check{p}(Z_i)] \leq \frac{1}{n} (-\log \check{p}(Z_i^n) - \log w(\check{p})^{2/\eta} + 2\text{CONV-LACK}(\eta, \check{p}) + R), \quad (17)$$

with remainder term  $R = O\left(\sqrt{n(K + \log n + \log(2 - \log w(\check{p})))}\right)$ .

The proof of this result is similar to the proof of Theorem 1 and will be provided in the full paper. Note that the weights of the main terms on the right side in Theorem 2 and 1 are different. In Theorem 2, the left term’s weight is reduced from  $C_\eta$  to 1, and the weight of the right term (CONV-LACK) is reduced from  $C_\eta$  (which is always  $\geq 2$ ) to 2. Unlike the ‘oracle’ bound Theorem 1, the ‘empirical’ bound Theorem 2 gives useful information without knowledge of  $E^*[-\log q(Z_i)]$ . If  $\mathcal{P} = \mathcal{P}_F$  for a classification model  $F$ , then the first term on the right-hand side is the empirical risk  $\beta^{-1} n^{-1} \sum_{i=1}^n \text{LOSS}(y_i, \check{f}(x_i))$  and the bound becomes similar to the PAC-Bayesian and Occam’s Razor (OR) bounds (McAllester, 2003; Blumer et al., 1987). Yet, by Condition (B), for  $\check{p} = \check{p}_{\text{SAFE}}$ , the remaining error term  $R/n$  is of order  $O(\log n / \sqrt{n})$  rather than  $O\left(\sqrt{-\log w(\check{p})/n}\right)$  as it would be for PAC-Bayes and OR-bounds. In this sense, for data  $Z^n$  such that for some  $\gamma > 0$  (say,  $\gamma = 0.1$ ), for all  $f \in F$ , the empirical loss of  $f$  on  $Z^n$  is larger than  $\gamma$ , the bound of Theorem 2 will be stronger than the best PAC-Bayesian or OR bounds. In other cases Theorem 2 gives weaker bounds than PAC-Bayes, since unlike PAC-Bayes it does not improve if  $\check{f}$  has empirical error  $\approx 0$ ; also it is not suitable for randomized classifiers. Theorem 2 is thus a first step, to be improved in future work.

#### 4. What Actually Happens

**Theorem 3** *Assume the notations and conditions of Theorem 1. Let  $c_\eta$  be as in (8). Fix  $u \geq 0$  and let  $\eta \leq \eta_{\text{CRIT}}(u)$ . WHP, we have  $\text{CONV-LACK}(\eta, \check{p}) \leq c_\eta \left( 3\text{RED}(2/\eta, \check{p}) + \frac{u}{\eta} \right) + R$ , with remainder term  $R = O((\log n + K + u)/\eta)$*

Applying Theorem 1 to the safe estimator  $\check{p}_{\text{SAFE}}$  with  $\eta \leq \eta_{\text{CRIT}}(u)$ , and using Theorem 3 to rewrite CONV-LACK, and using the fact that, if two inequalities hold with high probability,

the combined inequality also holds with high probability (see Proposition 12 in Section 5), we see that for all  $\eta \in \{1, 1/2, 1/3, \dots, \kappa_{\max}\}$  with  $\eta \leq \eta_{\text{CRIT}}(u)$ , the safe estimator achieves, WHP,

$$\begin{aligned} D^*(q \parallel \check{p}_{\text{SAFE}}) &\leq \frac{C}{n} \left( \text{RED}(2/\eta, \check{p}_{\text{SAFE}}) + \frac{u}{\eta} + c_\eta 3\text{RED}(2/\eta, \check{p}_{\text{SAFE}}) + \frac{c_\eta}{\eta} u + R' \right) \\ &= \frac{C(1+3c_\eta)}{n} \left( \text{RED}(2/\eta, \ddot{p}_{2/\eta}) + R'' \right) + \frac{C(1+c_\eta)}{n} \cdot \frac{u}{\eta} \\ &\leq \frac{C''}{n} \left( \text{RED}(2/\eta, \ddot{p}_{2/\eta}) + \frac{u}{\eta} + R'' \right), \end{aligned} \quad (18)$$

with constant  $C'' = C(1+3c_\eta)$  and new remainder term  $R'' = O((\log n + K)/\eta)$ . As long as we use (18) with  $u = 0$  (directly below) or  $u = 1$  (in Lemma 5) and  $\eta \leq \eta_{\text{CRIT}}(u)$ , then the terms  $u/\eta$  and  $R''$  are at most of the same order as the first term  $(-2/\eta) \log w(\ddot{p}_{2\eta})$  in RED, and hence do not affect the obtained convergence rates of  $\check{p}_{\text{SAFE}}$ .

**Example 4 [Best-Case: Model  $\mathcal{P}$  correct or convex]** Suppose that  $P^*$  is in the information closure of  $\mathcal{P}$ , i.e.  $q = p^*$ . Then  $\eta_{\text{CRIT}}(0) = 1$ , and, using (18) with  $u = 0$  and  $\eta = \eta_{\text{CRIT}}(0)$ , WHP,

$$D^*(p^* \parallel \check{p}_{\text{SAFE}}) = D^*(q \parallel \check{p}_{\text{SAFE}}) \leq \frac{C''}{n} \left( \text{RED}(2, \ddot{p}_2) + R'' \right), \quad (19)$$

where by Barron and Cover's original analysis, we would get (16). Except for the in-probability rather than in-expectation formulation, the only real difference is that the KL divergence is bounded in terms of a larger constant factor. This is the price we pay for not knowing in advance that our model was, in fact, correct, while using a procedure that still leads to good results if it is incorrect.

Barron and Cover (1991) show that, for a wide variety of probabilistic models  $\mathcal{M}$ , there exist countable discretizations  $\mathcal{P} \subset \mathcal{M}$  and corresponding priors  $w$  on  $\mathcal{P}$  such that  $\frac{1}{n} E_{Z^n \sim P^*} [\text{RED}(2, \ddot{p}_2)]$  is equal to the minimax convergence rate in KL risk if  $\mathcal{M}$  is “non-parametric”, or equal to the minimax rate up to a  $\log n$ -factor if  $\mathcal{M}$  is “parametric”. Using Markov's inequality\*, WHP  $\text{RED}(2, \ddot{p}_2)$  on the data (as in (19)) is not larger than a constant factor times its expectation  $E_{Z^n \sim P^*} [\text{RED}(2, \ddot{p}_2)]$ . This implies that the standard MDL estimator also achieves the minimax rate in probability (up to a logarithmic factor in the parametric case). Hence, by (19), so do we. A similar story\* can be told if  $\langle \mathcal{P} \rangle$  is convex or if the weaker condition (4) holds; again, up to constant factors, the safe estimator performs as well as the two-part estimator, which converges at near-optimal rates.

**When the Model is Wrong** Define  $D_{\text{SQ}}^*(q, p) = E^*[(\log p(Z_i)/q(Z_i))^2]$ . Such a variation of generalized KL divergence was earlier considered by, e.g., Kleijn and van der Vaart (2006). The lemma below shows that if the model is wrong, then the value of  $\eta_{\text{CRIT}}(u)$  depends on the relation between  $D_{\text{SQ}}^*$  and  $D^*$ . The lemma is not really new, being a direct translation of existing results of e.g. Tsybakov (2004) from ‘ $\mathcal{F}$ -space with loss function LOSS’ to ‘ $\mathcal{P}$ -space with log-loss’.

**Lemma 5** *Assume the notations and conditions of Theorem 1. Suppose further (E) that for some  $A > 0$  and some  $0 \leq \nu \leq 1$ , for all  $p \in \mathcal{P}$ ,  $D_{\text{SQ}}^*(q, p) \leq A(D^*(q \parallel p))^\nu$ . Then, for all  $u > 0$ , we have*

$$\eta_{\text{CRIT}}(u) \geq \min \left\{ \frac{1}{2 \log V}, B \left( \frac{u}{n} \right)^{\frac{1-\nu}{2-\nu}} \right\} \quad \text{where } B = (2/eA)^{\frac{1}{2-\nu}}. \quad (20)$$

When  $\mathcal{P}$  represents a classification model containing the Bayes classifier, condition (E) above specializes to the celebrated condition of (Mammen and Tsybakov, 1999; Tsybakov, 2004); the  $\kappa$  in (Tsybakov, 2004) is equal to  $\nu^{-1}$  in our notation. In particular, we automatically have  $D_{\text{SQ}}^*(q, p) \leq (\log V)^2$  so (E) always holds for  $\nu = 0$  and  $A = (\log V)^2$ . Using (20) with these values, and using  $1/2 < \sqrt{2/e}$ , it follows that for all  $u \geq 1$ ,

$$\eta_{\text{CRIT}}(u) \geq \frac{1}{2\log V} \sqrt{\frac{u}{n}} \geq 1/\kappa_{\max}, \quad (21)$$

which explains why we could restrict  $\eta$  to  $\eta > \kappa_{\max}$ ; see Example 6. If (E) holds for some  $\nu > 0$  though, then  $\eta_{\text{CRIT}}(u)$  is of larger order than  $\sqrt{u/n}$  and things get better; see below Example 6.

**Example 6 [Worst-Case]** Let  $u \geq 1$  and let  $\eta = \eta_{\text{CRIT}}(u)$ . Using (18), and (21), we see that WHP,

$$\begin{aligned} D^*(q \parallel \check{p}_{\text{SAFE}}) &\leq \frac{C''}{n} \left( \text{RED}(2/\eta, \ddot{p}_{2/\eta}) + \frac{u}{\eta} + R'' \right) \\ &= \frac{C''}{n} \left( -2c\sqrt{\frac{n}{u}} \log w(\ddot{p}_{2/\eta}) - \log \frac{\ddot{p}_{2/\eta}(Z_i^n)}{q(Z_i^n)} + c\sqrt{u \cdot n} + R'' \right) \end{aligned}$$

for some constant  $c = 2 \log V$ . Differentiating with respect to  $u$  shows that a minimum is achieved\* for  $u \approx -2 \log w(\ddot{p}_{2/\eta})$ . The resulting expression becomes

$$C'' \cdot 2c\sqrt{\frac{-\log w(\ddot{p}_{2/\eta})}{n}} + C'' \left( -\frac{1}{n} \log \frac{\ddot{p}_{2/\eta}(Z_i^n)}{q(Z_i^n)} + R'' \right). \quad (22)$$

For classification models, this bound is familiar from the computational learning literature. Now suppose that (E) holds for some  $\nu > 0$ . Then we can achieve better bounds: by the reasoning below Theorem 3,  $\check{p}_{\text{SAFE}}$  converges at the same rate as the  $\kappa$ -MDL estimator with  $\kappa = \eta_{\text{CRIT}}(1)^{-1} = O(n^{\frac{1-\nu}{2-\nu}})$ . From (18) we see that in the special case that  $q \in \mathcal{P}$ ,  $w(q) > 0$ , (18) gives a rate in probability of  $(\log n)/n^{1/(2-\nu)}$ , which, for classification models, is equal to the minimax optimal rate in expectation (Tsybakov, 2004) up to a log factor. The next example illustrates this for  $\nu = 1$ ; in the full paper\* we will also provide examples involving regression and classification with  $0 < \nu < 1$ .

### Example 7 [Bayesian inconsistency and Tsybakov's Condition]

Grünwald and Langford (2007) showed that standard MDL and Bayesian inference can be inconsistent in various ways if  $P^* \notin \langle \mathcal{P} \rangle$ , for countable models  $\mathcal{P} = \{p_0, p_1, \dots\}$  that are really classification models, i.e.  $\mathcal{P} = \mathcal{P}_F$  with  $F = \{f_0, f_1, \dots\}$  with  $p_j = p_{f_j}$  as given by (2), where  $\mathcal{Y} = \{0, 1\}$  and LOSS is the 0/1-LOSS. In these examples,  $p_0$  has positive prior  $w(p_0) > 0$  independent of the sample size, and for some  $\delta > 0$ , for all  $j > 0$ , it holds  $D^*(p_0 \parallel p_j) > \delta$ , i.e.  $\text{RISK}(p_j) > \text{RISK}(p_0) + \beta^{-1}\delta$ . In the examples Tsybakov's condition (E) holds with  $\nu = 1$  but only for very large  $A$ . Since thus  $q = p_0$  and, by Lemma 5,  $\eta_{\text{CRIT}}(1) > 1/C$  for some very large constant independent of  $n$ , it follows from (18) that the safe estimator converges WHP at rate  $O((-\log w(p_0) + R'')/n) = O(\log n/n)$ , much faster than the worst-case  $O(1/\sqrt{n})$ . However, explicit calculation of  $\eta_{\text{CRIT}}(1)$  shows that it is indeed very small, and since standard MDL and Bayesian MAP use an  $\eta$  equal to 1 or 2, it comes as no surprise that in this scenario they do not converge at all, i.e. with probability 1, for all large  $n$ , they select a distribution/classifier  $p \neq p_0$ , as was shown formally by Grünwald and Langford (2007).

## 5. The Proofs

**Preliminary Results** Our main tool is Proposition 8 below, a bound for ratios of probability densities, similar to earlier inequalities by Barron and Cover (1991); Li (1999); Zhang (2006). Below  $\text{TR}$  is a function mapping distributions to other distributions (we use notation as in (6)).

**Proposition 8** *Let  $Z^n$  be i.i.d.  $\sim P^*$ . Let  $\mathcal{P}$  be a countable set of (possibly defective) conditional densities for  $Z^n$  and let  $\check{p}$  be an arbitrary estimator. Let  $\mathcal{Q}$  be another set of (possibly defective) conditional densities for  $Z^n$ . Let  $\text{TR} : \mathcal{P} \rightarrow \mathcal{Q}$  be a function mapping distributions in  $\mathcal{P}$  to distributions in  $\mathcal{Q}$ . Let  $w$  be a (potentially defective) probability mass function on  $\mathcal{P}$ . Let  $\eta > 0$ . Then WHP,*

$$d_\eta^*(\text{TR}(\check{p}) \parallel \check{p}) \leq \frac{1}{n} \left( -\log \frac{w(\check{p})\check{p}(Z_i^n)}{\text{TR}(\check{p})(Z_i^n)} + \frac{K}{\eta} + \frac{1}{\eta} \log \sum_{p \in \check{p}} w(p)^\eta \right). \quad (23)$$

**Proof** We bound the probability that (23) does *not* hold:

$$\begin{aligned} P^* \left( \eta d_\eta^*(\text{TR}(\check{p}) \parallel \check{p}) > \frac{1}{n} \left( -\eta \log \frac{w(\check{p})\check{p}(Z_i^n)}{\text{TR}(\check{p})(Z_i^n)} + K + \log \sum_{p \in \check{p}} w(p)^\eta \right) \right) &= \\ P^* \left( \log \left( \frac{\check{p}(Z_i^n)}{\text{TR}(\check{p})(Z_i^n)} \right)^\eta > -\log \frac{w^\eta(\check{p})}{\sum w^\eta(p)} + K + n \log E^* \left( \frac{\check{p}(Z_i)}{\text{TR}(\check{p})(Z_i)} \right)^\eta \right) &\leq \\ P^* \left( \text{There exists } p \in \mathcal{P} \text{ with } \left( \frac{p(Z_i^n)}{\text{TR}(p)(Z_i^n)} \right)^\eta > e^K \left( \frac{\sum w(p)^\eta}{w^\eta(p)} \right) E^* \left( \frac{p(Z^n)}{\text{TR}(p)(Z^n)} \right)^\eta \right) &\leq \\ \sum_{p \in \mathcal{P}} P^* \left( \left( \frac{p(Z_i^n)}{\text{TR}(p)(Z_i^n)} \right)^\eta > e^K \left( \frac{\sum w(p)^\eta}{w^\eta(p)} \right) E^* \left( \frac{p(Z^n)}{\text{TR}(p)(Z^n)} \right)^\eta \right) &\leq e^{-K}, \end{aligned}$$

where the equality is basic rewriting, the first inequality follows from exponentiating both sides, absorbing  $n$  into the expectation  $E^*$  (which can be done since the  $Z_i$  are i.i.d.) and weakening, the second is the union bound, and the third is an instance of Markov's inequality. ■

In some applications we set, for all  $p \in \mathcal{P}$ ,  $\text{TR}(p)$  equal to the best approximating density  $q$ , and then the first term on the right in (23) is equal to  $\text{RED}(1/\eta, \check{p})$ ; the inequality is then a weakening of Zhang's, who provides an expectation rather than an in-probability form. In other applications (e.g. below Eq. (37)),  $\text{TR}(p)$  actually varies with  $p$ , and in this form, the inequality is new. We will apply this proposition in two different ways. In the first type of application, the goal is to get a (high-probability) upper bound on the left-hand side of (23). In the second type of application, the goal is to upper bound  $-\log \text{TR}(\check{p})(Z_i^n)$ . Thus, we rewrite (23) equivalently as:

$$-\frac{1}{n} \log \text{TR}(\check{p})(Z_i^n) \leq -\frac{1}{n} \log w(\check{p})\check{p}(Z_i^n) + R, \text{ with } R = -d_\eta^*(\text{TR}(\check{p}) \parallel \check{p}) + \frac{K + \log \sum_{p \in \check{p}} w(p)^\eta}{n\eta}. \quad (24)$$

In such applications, we take a value of  $\eta$  guaranteeing  $d_\eta^*(\text{TR}(\check{p}) \parallel \check{p}) \geq 0$  or (as e.g. in Lemma 14), we allow  $d_\eta^*(\text{TR}(\check{p}) \parallel \check{p})$  to be negative but not too negative, so that the bound remains useful.

Let  $p, p' \in \mathcal{P}$  such that  $D^*(p \parallel p') \geq 0$ , and let\*  $\lambda^\circ := \arg \min_{\lambda \in [0,1]} D^*(p^* \parallel \text{MIX}(p, p', \lambda)) = \arg \max_{\lambda \in [0,1]} D^*(\text{MIX}(p, p', \lambda) \parallel p)$  (if more than one  $\lambda$  achieves the extremum, we take the smallest). Our second key result states that if  $D^*(\text{MIX}(p, p', \lambda^\circ) \parallel p)$  is small, then  $E^*[\text{MIX}(p, p', \lambda^\circ)(Z_i)/p(Z_i)]$  cannot be much larger than 1; equivalently  $d_1^*(p \parallel \text{MIX}(p, p', \lambda^\circ))$  cannot be much smaller than 0:

**Lemma 9** Let  $(\mathcal{P}, P^*)$  be as on page 401,  $V = V(\mathcal{P}, P^*)$  be as in (5), and let  $p, p' \in \mathcal{P}$  be such that  $D^*(p\|p') \geq 0$ , and  $\lambda^\circ$  be as above. (a) If  $\lambda^\circ = 0$  ( $p$  is closer to  $p^*$  than any mixture of  $p$  and  $p'$ ) then for all  $\lambda \in [0, 1]$ ,  $d_1^*(p\|\text{MIX}(p, p', \lambda)) \geq 0$ ; otherwise, (b),  $-d_1^*(p\|\text{MIX}(p, p', \lambda^\circ)) \leq 2V^2 D^*(\text{MIX}(p, p', \lambda^\circ)\|p)$ .

**Proof** Let  $g(\lambda) = D^*(\text{MIX}(p, p', \lambda)\|p)$ . Then  $g(0) = 0, g(1) \leq 0$ . We first need the following (proof straightforward by differentiation, see appendix):

**Proposition 10** 1.  $g'(0) = E^*\left(\frac{p'(Z_i)}{p(Z_i)}\right) - 1$ ; if  $\lambda^\circ = 0$  then (1a)  $g'(0) \leq 0$ ; if  $\lambda^\circ > 0$  then (1b)  $g'(0) > 0, g'(\lambda^\circ) = 0$  and  $g'(1) \leq 0$ ; 2. if  $p(Z_i) = p'(Z_i)$   $P^*$ -almost surely, then (2a)  $g'(\lambda) = g''(\lambda) = 0$  on  $\lambda \in [0, 1]$ . Otherwise (2b)  $g''(\lambda) < 0$  on  $[0, 1]$  and  $\max_{\lambda \in [0, 1]} |g''(\lambda)| \leq \min_{\lambda \in [0, 1]} V^2 |g''(\lambda)|$ .

Abbreviate  $d_1^*(p\|\text{MIX}(p, p', \lambda))$  to  $d^*(\lambda)$  and  $g(\lambda^\circ) = D^*(\text{MIX}(p, p', \lambda^\circ)\|p)$  to  $D^*$ , and note that

$$-d^*(\lambda) = \log E^*\left(\frac{(1-\lambda)p(Z_i)+\lambda p'(Z_i)}{p(Z_i)}\right) \leq E^*\left(\frac{(1-\lambda)p(Z_i)+\lambda p'(Z_i)}{p(Z_i)}\right) - 1 = \lambda g'(0). \quad (25)$$

In case (1a) and (2a) the result is now immediate, so assume (1b) and (2b). Then by a first-order Taylor approximation of  $g'$ , for some  $0 \leq \lambda_1 \leq \lambda^\circ$ ,  $g'(0) = g'(\lambda^\circ) - \lambda^\circ g''(\lambda_1) = \lambda^\circ |g''(\lambda_1)|$ , so that (25) gives  $-d^*(\lambda^\circ) \leq (\lambda^\circ)^2 |g''(\lambda_1)|$ . Also, by a 2nd-order Taylor expansion of  $g$  around  $\lambda^\circ$  we find, for some  $0 \leq \lambda_2 \leq \lambda^\circ$ , that  $0 = g(0) = (1/2)(\lambda^\circ)^2 g''(\lambda_2) + g(\lambda^\circ)$ , so  $D^* = (1/2)(\lambda^\circ)^2 |g''(\lambda_2)|$ . Combining with our expression for  $d^*(\lambda^\circ)$ , we get  $\frac{-d^*(\lambda^\circ)}{D^*} \leq 2 \frac{|g''(\lambda_1)|}{|g''(\lambda_2)|}$ . The result now follows by part (2b) of Proposition 10. ■

The next proposition is about varying exponents rather than mixture coefficients:

**Proposition 11** Let  $\mathcal{P} = \{p, p'\}$  be such that  $V(\mathcal{P}, P^*) < \infty$  and  $D^*(p\|p') > 0$ . Then (a): letting  $g(\eta) = \log E^*(p'(Z_i)/p(Z_i))^\eta = -\eta d_\eta^*(p\|p')$ , we have  $g(0) = 0$ ,  $g(\eta)$  is decreasing at  $\eta = 0$  and  $\exp(g(\eta))$  is strictly convex, so that there exists at most one  $\eta' > 0$  with  $g(\eta') = 0$ , and  $g(\eta)$  is increasing for  $\eta \geq \eta'$ . (b) Define  $\lambda^\circ$  as in Lemma 9. If  $\lambda^\circ = 0$  ( $p$  is closer to  $p^*$  than any mixture of  $p$  and  $p'$ ) then  $\forall \eta \in (0, 1)$ ,  $d_\eta^*(p\|p') > 0$ .

**Proof** (a) is just differentiation of  $E^*(p'(Z_i)/p(Z_i))^\eta$  (see proof of Lemma 5); details omitted. (b) is immediate from (a) because by Lemma 9, part (a),  $d_1^*(p\|p') = d_1^*(p\|\text{MIX}(p, p', 1)) = -g(1) \geq 0$ . ■

The following proposition provides the glue that ties all our inequalities together:

**Proposition 12 (log-Bonferroni)** Let  $\mathcal{J}$  be a finite or countably infinite set. Let  $\{Y_j\}_{j \in \mathcal{J}}$  be a collection of random variables, let  $\{a_j\}_{j \in \mathcal{J}}$  be a collection of constants in  $\mathbb{R}$  and let  $\{f_j\}_{j \in \mathcal{J}}$  be a collection of increasing functions  $\mathbb{R} \rightarrow \mathbb{R}$ . Suppose that for all  $j \in \mathcal{J}$ , WHP,  $Y_j \leq a_j + f_j(K)$ . Then, for any collection of positive numbers  $\{w_j\}_{j \in \mathcal{J}}$  such that  $\sum_{j \in \mathcal{J}} w_j = 1$ , we have, WHP,

$$\text{For all } j \in \mathcal{J}, Y_j \leq a_j + f_j(K - \log w_j).$$

This result is a straightforward consequence of the union bound that appears in one form or other in many COLT papers; for convenience there is a proof in the appendix.

**Notation common to the Proofs** In all proofs below we make use of the following concepts: let  $w$  be a prior for a countable set of densities  $\mathcal{P}$ . Let  $p \in \mathcal{P}$ . Relative to  $w$  and  $P^*$ , we define\* the *optimal density at p's description length* as

$$\text{OPT}(p) := \arg \min_{p' \in \mathcal{P}: -\log w(p') \leq \lceil -\log w(p) \rceil} D^*(p^* \| p) \quad (26)$$

For an estimator  $\check{p}$ ,  $\text{OPT}(\check{p})$  is itself a random variable, representing the best distribution (closest in KL divergence to  $p^*$ ) with prior no smaller (up to rounding) (or “complexity” no larger, up to rounding) than the  $\check{p}$  selected for the given data  $Z_1^n$ . We further define

$$\text{OPT}(\mathcal{P}) = \{p \in \mathcal{P} : p = \text{OPT}(p') \text{ for some } p' \in \mathcal{P}\}. \quad (27)$$

Now define for  $p \in \mathcal{P}$ ,  $\text{OPTMIX}(p) := \text{MIX}(\text{OPT}(p), p, \lambda^\circ)$ , where  $\lambda^\circ \in [0, 1]$  minimizes\*

$$E^*[-\log \text{MIX}(\text{OPT}(p), p, \lambda)(Z_1)] = E^*[-\log ((1 - \lambda)\text{OPT}(p)(Z_1) + \lambda p(Z_1))]. \quad (28)$$

Note that  $D^*(\text{OPTMIX}(\check{p}^\eta) \| \check{p}^\eta) = \max_{\lambda \in [0, 1]} D^*(\text{MIX}(\text{OPT}(\check{p}^\eta), \check{p}^\eta, \lambda) \| \check{p}^\eta) \geq D^*(\text{OPT}(\check{p}^\eta) \| \check{p}^\eta)$ . (29)

### 5.1. Proofs of Main Results

**Proof of Theorem 1** We first consider an arbitrary fixed  $\eta$  and a fixed  $u$ . We have:

$$D^*(q \| \check{p}) = D^*(q \| \text{OPT}(\check{p})) + \frac{1}{\eta} D^*(\text{OPT}(\check{p}^\eta) \| \check{p}^\eta) \leq D^*(q \| \text{OPT}(\check{p})) + \frac{1}{\eta} D^*(\text{OPTMIX}(\check{p}^\eta) \| \check{p}^\eta), \quad (30)$$

where the equality is straightforward from the definition of  $D^*$  and the inequality follows from (29). By Lemma 14, we can bound the term  $D^*(q \| \text{OPT}(\check{p}))$  from above and rewrite (30) to get, WHP,

$$D^*(q \| \check{p}) \leq \frac{C_\eta}{n} \left( -\frac{1}{\eta} \log \frac{w(\check{p}) \text{OPTMIX}(\check{p}^\eta)(Z_1^n)}{q^\eta(Z_1^n)} + \frac{u}{\eta'} + R_1 \right) + T \quad (31)$$

with  $C_\eta = 2 + 2\eta \log V$ ,  $\eta' = \min\{\eta, \eta_{\text{CRIT}}(u)\}$ ,  $R_1$  as in Lemma 14 and

$$T = -\frac{C_\eta}{\eta} d_1^*(\text{OPT}(\check{p}^\eta) \| \text{OPTMIX}(\check{p}^\eta)) + \frac{1}{\eta} D^*(\text{OPTMIX}(\check{p}^\eta) \| \check{p}^\eta), \quad (32)$$

We proceed to rewrite  $T$ .  $-d_1^*(\text{OPT}(\check{p}^\eta) \| \text{OPTMIX}(\check{p}^\eta))$  may very well be *positive*, but by Lemma 9, applied with  $(\mathcal{P}, p, p') \leftarrow (\mathcal{P}^{(\eta)}, \text{OPT}(\check{p})^\eta, \check{p}^\eta)$  (the notation indicates that e.g.  $\mathcal{P}$  in Lemma 9 is instantiated to  $\mathcal{P}^{(\eta)}$  as used above), we can bound (32) to get:  $T \leq \eta^{-1}(C_\eta C'_\eta + 1) D^*(\text{OPTMIX}(\check{p}^\eta) \| \check{p}^\eta)$ , where we used  $D^*(\text{OPTMIX}(\check{p}^\eta) \| \text{OPT}(\check{p})^\eta) \leq D^*(\text{OPTMIX}(\check{p}^\eta) \| \check{p}^\eta)$ , and  $C'_\eta = 2V^{2\eta}$ . Letting  $w'(p^\eta) := w(p)$ , this can be further bounded using Lemma 13 below, with  $(\check{p}, \mathcal{P}, w) \leftarrow (\check{p}^\eta, \mathcal{P}^{(\eta)}, w')$  (notation as explained above). This gives, WHP:

$$T \leq \frac{C_\eta}{n} \left( -\frac{c_\eta}{\eta} \log \frac{\check{p}^\eta(Z_1^n) w(\check{p})^2}{\text{SUPMIX}(\check{p}^\eta)(Z_1^n)} + R'_1 \right) = \frac{C_\eta}{n} (\text{CONV-LACK}(\eta, \check{p}) + R'_1), \quad (33)$$

where  $R'_1 = \eta^{-1}c_\eta 2K$  and  $c_\eta = C_\eta C'_\eta + 1$  and we used the definition of CONV-LACK. Now apply Proposition 8 as in (24), with  $(\mathcal{P}, \check{p}, \mathcal{Q}, \text{TR}(\cdot), w, \eta) \leftarrow (\mathcal{P}^{(\eta)}, \check{p}^\eta, \mathcal{Q}, \text{OPTMIX}(\cdot), w', 1)$ , where  $\mathcal{Q} = \{\text{OPTMIX}(p^\eta) \mid p \in \mathcal{P}\}$  and  $w'$  as above. Since, by Lemma 9, part (b),  $d_1^*(\text{OPTMIX}(p^\eta) \parallel \text{OPT}(p)^\eta) \geq 0$ , we find that WHP,  $-\log \text{OPTMIX}(\check{p}^\eta)(Z_i^n) \leq -\log w(\check{p})\check{p}^\eta(Z_i^n) + K$ . Substituting this and (33) into (31), we get with Proposition 12 (with  $|\mathcal{J}| = 3, w_j = 1/3$ ) that WHP,

$$D^*(q \parallel \check{p}) \leq \frac{C_\eta}{n} \left( -\frac{1}{\eta} \log \frac{w(\check{p})^2 \check{p}^\eta(Z_i^n)}{q^\eta(Z_i^n)} + \frac{u}{\eta'} + \text{CONV-LACK}(\eta, \check{p}) + R_2 + R'_2 \right), \quad (34)$$

where  $R_2 = (4(K + \log 3) + 4 \log(2 + \lceil -\log w(\check{p}) \rceil)) / \eta'$  and  $R'_2 = c_\eta(2(K + \log 3)) / \eta$ . The result now follows for a fixed value of  $u$  and  $\eta$ . To prove that it holds uniformly for  $u \in \{0, 1, 2, \dots, nL_{\max}\}$ ,  $\eta \in \{1, 1/2, 1/3, \dots, 1/\kappa_{\max}\}$ , we use Proposition 12; details omitted\*.

**Lemma 13** *Let  $(\mathcal{P}, P^*)$  and  $w$  be as on page 401. We have WHP,*

$$D^*(\text{OPTMIX}(\check{p}) \parallel \check{p}) \leq \frac{C_1}{n} \left( -\log \frac{\check{p}(Z_i^n) w^2(\check{p})}{\text{SUPMIX}(\check{p})(Z_i^n)} + 2K \right)$$

where  $C_1 = 2 + 2 \log V$  is a constant and SUPMIX is defined as in (7) above.

**Proof** By Proposition 15, applied with  $\eta = 1$ , we get

$$D^*(\text{OPTMIX}(\check{p}) \parallel \check{p}) \leq C_1 d_{1/2}^*(\text{OPTMIX}(\check{p}) \parallel \check{p}) - (C_1 - 1)d_1^*(\text{OPTMIX}(\check{p}) \parallel \check{p}) \leq C_1 d_{1/2}^*(\text{OPTMIX}(\check{p}) \parallel \check{p}),$$

where  $C_1 = 2 + 2 \log V$  and the final inequality follows because by Proposition 11,  $d_1^*(\text{OPTMIX}(\check{p}) \parallel \check{p}) \geq 0$ . We now let  $\mathcal{Q} = \{\text{MIX}(p_0, p_1, \lambda) : p_0, p_1 \in \mathcal{P}, \lambda \in [0, 1]\}$  be the set of two-component mixtures of elements of  $\mathcal{P}$ , and apply Proposition 8, with  $(\mathcal{P}, \check{p}, \mathcal{Q}, \text{TR}(\cdot), w, \eta) \leftarrow (\mathcal{P}, \check{p}, \mathcal{Q}, \text{OPTMIX}(\cdot), w^2, 1/2)$  (notation as below (32)). We get that, WHP,  

$$d_{1/2}^*(\text{OPTMIX}(\check{p}) \parallel \check{p}) \leq -\frac{1}{n} \log \frac{w^2(\check{p})\check{p}(Z_i^n)}{\text{OPTMIX}(\check{p})(Z_i^n)} + \frac{2K}{n} \leq -\frac{1}{n} \log \frac{w^2(\check{p})\check{p}(Z_i^n)}{\text{SUPMIX}(\check{p})(Z_i^n)} + \frac{2K}{n}. \quad \blacksquare$$

**Lemma 14** *Assume the conditions and notation of Theorem 1. For all  $0 < \eta \leq 1$ , WHP,*

$$D^*(q \parallel \text{OPT}(\check{p})) \leq \frac{C_\eta}{n} \left( -\frac{1}{\eta} \log \frac{w(\check{p})\text{OPTMIX}(\check{p}^\eta)(Z_i^n)}{q^\eta(Z_i^n)} + \frac{u}{\eta'} + R_1 \right) - \frac{C_\eta}{\eta} d_1^*(\text{OPT}(\check{p}^\eta) \parallel \text{OPTMIX}(\check{p}^\eta)),$$

where  $C_\eta = 2 + 2\eta \log V$ ,  $\eta' = \min\{\eta, \eta_{\text{CRIT}}(u)\}$  and remainder  $R_1 = (3K + 4 \log(2 + \lceil -\log w(\check{p}) \rceil)) / \eta'$ . (Note that  $d_1^*(\text{OPT}(\check{p}^\eta) \parallel \check{p}^\eta)$  may be negative).

**Proof** We apply Proposition 15 with  $\eta$  set to  $\eta'$ . This gives  $D^*(q \parallel \text{OPT}(\check{p})) \leq C_{\eta'} d_{\eta'/2}^*(q \parallel \text{OPT}(\check{p})) + (C_{\eta'} - 1)R$ , where  $R = -d_{\eta'}^*(q \parallel \text{OPT}(\check{p}))$ . By Proposition 11, part(a), if  $\eta'R \geq 0$  then, since  $\eta' \leq \eta_{\text{CRIT}}(u)$ , we have  $\eta'R \leq -\eta_{\text{CRIT}}(u)d_{\eta_{\text{CRIT}}(u)}^*(q \parallel \text{OPT}(\check{p})) \leq u/n$ . It follows that  $R \leq u/(\eta'n)$ , so we have

$$D^*(q \parallel \text{OPT}(\check{p})) \leq C_{\eta'} d_{\eta'/2}^*(q \parallel \text{OPT}(\check{p})) + \frac{1}{\eta'}(C_{\eta'} - 1)\frac{u}{n}. \quad (35)$$

Now fix some  $p_0 \in \mathcal{P}$ . Set  $w'(p_0) = 1$ ,  $\text{TR}(p_0) = q$  and apply Proposition 8 with  $(\mathcal{P}, \check{p}, \mathcal{Q}, \text{TR}(\cdot), w, \eta) \leftarrow (\mathcal{P}, p_0, \{q\}, \text{TR}(\cdot), \eta'/2)$ . The  $\check{p}$  in the proposition is a degenerate

estimator that is always equal to the fixed  $p_0$  and does not depend on the data, and  $\text{TR}(\check{p})$  is always equal to  $q$ . We get that WHP,

$$d_{\eta'/2}^*(q\|p_0) \leq \frac{1}{n} \left( -\log \frac{p_0(Z_i^n)}{q(Z_i^n)} + \frac{2K}{\eta'} \right). \quad (36)$$

Note that we can enumerate the elements of  $\text{OPT}(\mathcal{P})$  as given by (27) as  $\{p_1, p_2, \dots\}$  where, for all  $j$ , it holds  $j-1 \leq \lceil -\log w(p_j) \rceil \leq j$ . Using the log-Bonferroni Proposition 12 with  $|\mathcal{J}| = \mathbb{N}$ ,  $w_j = 1/j(j+1)$  and  $f_j(K) = 2K/\eta'(u)$ , we get that, WHP, uniformly for all  $p_j \in \text{OPT}(P)$ ,

$$\begin{aligned} d_{\eta'/2}^*(q\|p_j) &\leq \frac{1}{n} \left( -\log \frac{p_j(Z_i^n)}{q(Z_i^n)} + \frac{2}{\eta'} (K + 2 \log(j+1)) \right) \\ &\leq \frac{1}{n} \left( -\log \frac{p_j(Z_i^n)}{q(Z_i^n)} + \frac{2}{\eta'} (K + 2 \log(2 - \lceil -\log w(p_j) \rceil)) \right). \end{aligned} \quad (37)$$

We now let  $\mathcal{R} = \{\text{OPTMIX}(p^\eta) \mid p \in \mathcal{P}\}$ , and define a prior  $w'$  on  $\mathcal{R}$  with, for  $r \in \mathcal{R}$ ,  $w'(r) := w(p)$  for the  $p$  such that  $r = \text{OPTMIX}(p^\eta)$ . We set  $\text{TR}(p') := \text{OPT}(\check{p}^\eta)$ . We now use Proposition 8 again (in the form (24)) with  $(\mathcal{P}, \check{p}, \mathcal{Q}, \text{TR}(\cdot), w, \eta) \leftarrow (\mathcal{R}, \text{OPTMIX}(\check{p}^\eta), \text{OPT}(\mathcal{P})^\eta, \text{TR}(\cdot), w', 1)$ . (notation as below (32); effectively we use  $\text{OPTMIX}(\check{p}^\eta)$  as an estimator here.). We get, WHP,

$$-\frac{1}{n} \log \text{OPT}(\check{p}^\eta)(Z_i^n) \leq -\frac{1}{n} \log w(\check{p}) \text{OPTMIX}(\check{p}^\eta)(Z_i^n) - d_1^*(\text{OPT}(\check{p}^\eta)\|\text{OPTMIX}(\check{p}^\eta)) + \frac{K}{n}. \quad (38)$$

Dividing (38) by  $\eta$ , using  $\eta^{-1} \log \text{OPT}(\check{p}^\eta) = \log \text{OPT}(\check{p})$ , and then combining with (35) and (37), with  $p_j$  set to  $\text{OPT}(\check{p})$ , we get, WHP,

$$\begin{aligned} D^*(q\|\text{OPT}(\check{p})) &\leq C_{\eta'} \cdot \\ &\left( \frac{1}{n} \left( -\frac{1}{\eta} \log \frac{w(\check{p}) \text{OPTMIX}(\check{p}^\eta)(Z_i^n)}{q^\eta(Z_i^n)} + R \right) - \frac{1}{\eta} d_1^*(\text{OPT}(\check{p}^\eta)\|\text{OPTMIX}(\check{p}^\eta)) + \frac{1}{\eta' n} u \right), \end{aligned} \quad (39)$$

where  $R = (2/\eta')(K + 2 \log(2 - \lceil -\log w(p_j) \rceil)) + (1/\eta)K$ , which is no greater than  $R_1$  in the statement of the lemma. This proves the result for  $\eta \leq \eta_{\text{CRIT}}(u)$  (for then  $\eta' = \eta$ ). For the case that  $\eta > \eta_{\text{CRIT}}(u)$ , note that then  $C_\eta > C_{\eta'}$ . Because by definition of  $q$  the left-hand side of (39) must be nonnegative, we have WHP that both (39) holds and its right-hand side is nonnegative, so that with the same probability, (39) holds with  $C_{\eta'}$  replaced by  $C_\eta$ . The result follows.  $\blacksquare$

**Proposition 15** *Let  $P^*$  be a distribution on  $\mathcal{Z}$ , let  $p$  and  $q$  be conditional distributions for  $Y$  given  $X$ , and let  $V = V(\{p, q\}, P^*)$  defined as in (5). For all  $\eta \leq 1$  and all  $C_\eta \geq 2 + 2\eta \log V$ , we have*

$$D^*(q\|p) \leq C_\eta \cdot d_{\eta/2}^*(q\|p) - (C_\eta - 1)d_\eta^*(q\|p).$$

*In particular, if  $d_\eta^*(q\|p) \geq 0$ , then  $D^*(q\|p) \leq C_\eta d_{\eta/2}^*(q\|p)$ , i.e. the generalized KL divergence is upper bounded by a constant times the generalized Rényi divergence of order  $1/2\eta$ .*

**Proof** This result is a straightforward extension of a result due to Andrew Barron and Jonathan Li, published in Li's (1999) thesis. See Lemma 5.11, page 67 and Lemma 5.12, page 73 of (Li, 1999).  $d_\eta^*$  corresponds to  $\log c = \log \int fg/f^*$  in Li's notation;  $p^*$  corresponds to  $f$  in Li's notation,  $p(\cdot | x)^\eta$  corresponds to  $g$  in Li, and  $q(\cdot | x)^\eta$  corresponds to  $f^*$ . Our argument is slightly more involved than Li's since we allow conditioning on  $x$ ; this also accounts for the extra  $\eta \log V$  term in the constant  $C_\eta$ . For convenience, we provide a full proof in the appendix. ■

**Proof of Theorem 3** We fix a  $\hat{p}$  and  $\hat{\lambda}$  be such that  $\text{SUPMIX}(\check{p}^\eta) = \text{MIX}(\hat{p}^\eta, \check{p}^\eta, \hat{\lambda})$ , i.e.  $\hat{p}$  and  $\hat{\lambda}$  achieve\* the supremum in (7) applied with estimator  $\check{p}^\eta$ .

Let  $\dot{\lambda} = \arg \min_{\lambda \in \{0, 1/n, 2/n, \dots, 1\}} -\log \text{MIX}(\hat{p}^\eta, \check{p}^\eta, \lambda)(Z_i^n)$  be a discretized version of  $\hat{\lambda}$ . We have:

$$\begin{aligned} -\log \frac{\check{p}^\eta(Z_i^n)w^2(\check{p})}{\text{SUPMIX}(\check{p}^\eta)(Z_i^n)} &\leq -\log \frac{q^\eta(Z_i^n)}{\text{SUPMIX}(\check{p}^\eta)(Z_i^n)} - \log \frac{\check{p}^\eta(Z_i^n)w^2(\check{p})}{q^\eta(Z_i^n)} \\ &\leq -\log \frac{q^\eta(Z_i^n)}{\text{MIX}(\hat{p}^\eta, \check{p}^\eta, \dot{\lambda})(Z_i^n)} + V^\eta + \eta \text{RED}(2/\eta, \check{p}), \end{aligned} \quad (40)$$

where in the second inequality we used a simple first-order Taylor approximation, showing that  $-\log \text{MIX}(\hat{p}^\eta, \check{p}^\eta, \dot{\lambda})(Z_i^n) \leq -\log \text{MIX}(\hat{p}^\eta, \check{p}^\eta, \hat{\lambda})(Z_i^n) + V^\eta$  (details omitted). We now come to the crucial step: we will prove that WHP, we have

$$-\log \frac{q^\eta(Z_i^n)}{\text{MIX}(\hat{p}^\eta, \check{p}^\eta, \dot{\lambda})(Z_i^n)} \leq -\log w^2(\check{p}) + \log(n+1) + K + u. \quad (41)$$

This result follows by applying Proposition 8 to an extended model  $\mathcal{P}'$  with a prior  $w'$  defined, at sample size  $n$ , as follows:  $\mathcal{P}' = \{\text{MIX}(p_0^\eta, p_1^\eta, \lambda) : p_0^\eta, p_1^\eta \in \mathcal{P}, \lambda \in [0, 1]\}$ ; and, for  $\lambda \in \Lambda := \{0, 1/n, \dots, 1\}$ ,  $w'(\text{MIX}(p_0^\eta, p_1^\eta, \lambda)) := w(p_0^\eta) \cdot w(p_1^\eta)(n+1)^{-1}$ . Thus,  $\mathcal{P}'$  is the set of all two-component mixtures of  $\mathcal{P}^{(\eta)}$ ; and  $w'$  has its support on all two-component mixtures with  $\lambda \in \Lambda$ , and puts mass 0 on all other mixtures. Note that  $w'$  is indeed a prior, i.e.  $\sum_{p_0, p_1 \in \mathcal{P}, \lambda \in \Lambda} w'(\text{MIX}(p_0^\eta, p_1^\eta, \lambda)) \leq 1$ . We now apply Proposition 8 in the form (24) with, for all  $p' \in \mathcal{P}'$ ,  $\text{TR}(p') := q^\eta$ , and  $\eta$  in the proposition set to 1, and with the estimator that, for data  $z^n$ , chooses  $\text{MIX}(\hat{p}^\eta, \check{p}^\eta, \dot{\lambda})$ . That is, we set

$(\mathcal{P}, \check{p}, \mathcal{Q}, \text{TR}(\cdot), w, \eta) \leftarrow (\mathcal{P}', \text{MIX}(\hat{p}^\eta, \check{p}^\eta, \dot{\lambda}), \{q^\eta\}, \text{TR}(\cdot), w', 1)$ . This gives (41), where we also used (a), by definition,  $-\log w'(\text{MIX}(\hat{p}_0^\eta, \hat{p}_1^\eta, \dot{\lambda})) \leq -\log w(\check{p})^2 + \log(n+1)$ ; and (b): since  $\eta \leq \eta_{\text{CRIT}}(u)$ , we have that  $d_1^*(q^\eta \| p^\eta) \geq -u/n$  for both  $p = \check{p}$  and  $p = \hat{p}$ , which implies, from the definition of  $d_1^*$ , that  $d_1^*(q^\eta \| (1-\lambda)\hat{p}^\eta + \lambda\check{p}^\eta) \geq -u/n$  for all  $\lambda \in [0, 1]$ , in particular for  $\dot{\lambda}$ .

Combining (40) and (41), using the definition of CONV-LACK, it follows that, WHP,

$$\begin{aligned} \text{CONV-LACK} &\leq c_\eta \text{RED}(2/\eta, \check{p}) + \frac{c_\eta}{\eta} (-\log w^2(\check{p}) + \log(n+1) + K + u + V^\eta) \\ &= c_\eta \text{RED}(4/\eta, \check{p}) + \frac{c_\eta}{\eta} (\log(n+1) + K + u + V^\eta). \end{aligned}$$

Now with some relatively straightforward manipulations\* we get that we get, WHP,  $\text{RED}(4/\eta, \check{p}) \leq 3\text{RED}(2/\eta, \check{p}) + \frac{2K+2u+2\log 2}{\eta}$ , so that, using the log-Bonferroni Proposition 12 with  $\mathcal{J} = 2$ , the above becomes  $\text{CONV-LACK} \leq 3c_\eta \text{RED}(2/\eta, \check{p}) + R$ , and the result follows.

**Proof of Lemma 5** A second-order Taylor expansion of  $E^*(p(Z_i)/q(Z_i))^\eta = E^*(e^{\eta \log p(Z_i)/q(Z_i)})$  around  $\eta = 0$  shows that, for all  $\eta > 0$ , for some  $0 \leq \eta' \leq \eta$ , we have:

$$\begin{aligned} E^*\left(\frac{p(Z_i)}{q(Z_i)}\right)^\eta &= 1 - \eta E^*[-\log p(Z_i)/q(Z_i)] + \frac{1}{2}\eta^2 E^*\left(\frac{p(Z_i)}{q(Z_i)}\right)^{2\eta'} \left(\log \frac{p(Z_i)}{q(Z_i)}\right)^2 \leq \\ &\quad 1 - \eta D^* + \frac{1}{2}\eta^2 V^{2\eta} D_{\text{SQ}}^*, \end{aligned}$$

where we abbreviate  $D^*(q||p)$  to  $D^*$  and  $D_{\text{SQ}}^*(q,p)$  to  $D_{\text{SQ}}^*$ , and we replaced all factors in the expectation in the second order by their maximum. From now on we repeatedly use  $D^*(q||p) \geq 0$  which holds because  $q$  is best-approximating. It is sufficient to show that the right-hand side of this expression is bounded by  $1 + u/n$  if we plug in  $\eta \leq \eta_{\text{CRIT}}(u)$  as defined above. Dividing the inequality by  $\eta$  and using assumption (E), it is thus sufficient if we can show that

$$-D^* + \eta(D^*)^\nu \cdot b \leq \eta^{-1}(u/n) \quad (42)$$

where we set  $b = \frac{A}{2}V^{2\eta}$ . We may further assume  $(D^*)^{1-\nu} \leq \eta b$ ,  $(43)$

for if this does not hold, then  $-D^* = -(D^*)^\nu(D^*)^{1-\nu} \leq -(D^*)^\nu \eta b$  and then (42) holds trivially. Now first consider the case  $0 < \nu < 1$ . From (43) it follows that  $D^* \leq (\eta b)^{1/(1-\nu)}$ . By (42), it is thus sufficient if we can show that  $\eta \cdot (\eta b)^{\nu/(1-\nu)} b \leq \eta^{-1}u/n$ . Solving for  $\eta$  gives  $\eta^{2+\frac{\nu}{1-\nu}} \leq \frac{u}{n}b^{-1/(1-\nu)}$ , which can be rewritten to  $\eta \leq C$ , where  $C = \left(\frac{u}{n}\right)^{\frac{1-\nu}{2-\nu}} b^{-1/(2-\nu)}$ . Thus, weakening the requirement, it is sufficient if  $\eta \leq \min\{1/(2\log V), C\}$ . But if  $\eta \leq 1/(2\log V)$ , then  $b^{-1} \geq 2/(eA)$ , so it is also sufficient if  $\eta \leq \min\left\{\frac{1}{2\log V}, B\left(\frac{u}{n}\right)^{\frac{1-\nu}{2-\nu}}\right\}$ . (20) now follows for the case  $0 < \nu < 1$ . The limiting cases  $\nu = 0$  and  $\nu = 1$  can be handled similarly; we omit details.

## 6. Discussion and Future Work

The great advances we made were already summarized on page 3; but currently, our work also has at least two major restrictions : (a)  $V = V(\mathcal{P}, P^*)$  as in (5) must be bounded; and (b)  $V$  occurs in the definition of CONV-LACK, so that it must be known in order to apply the safe estimator. Neither restriction is problematic for classification models, as long we used a fixed  $\beta$  in (2); both are problematic for e.g. standard regression models though. As to (a), currently our results only hold for such models if  $P^*$  has bounded support. In future work, we hope to replace the strong  $V < \infty$  condition with a weaker condition on moments of  $P^*$ . As to (b), we do have a version of all our results in which  $V$  is replaced by its empirical counterpart  $\bar{V} = \sup_{i \in \{1, \dots, n\}} \sup_{p, p' \in \mathcal{P}} p(Z_{i|})/p'(Z_{i|})$ , but with worse constants. We hope to refine this in future work.

Another issue is that, even if known,  $V$  or  $\bar{V}$ , appearing in CONV-LACK, may be so large as to make the approach useless in practice (even aside from computational issues, which, in this preliminary study, we decided not to deal with at all). We should note though that our current results hold for arbitrary priors  $w$ , in particular, priors with very heavy tails. Most priors used in practice have lighter tails, i.e.  $\sum_{p \in \mathcal{P}} w^\rho(p) < \infty$  for some  $\rho < 1$ . For such priors, the theorems still hold for the prior  $w'$  defined as  $w'(p) \propto w^\rho(p)$  rather

than the original  $w$ . As a result, the safe estimator  $\check{p}_{\text{SAFE}}$  defined relative to  $w'$  rather than  $w$  will effectively choose simpler distributions (with higher  $w(p)$ ) for the same data, but all occurrences of  $V$  in our theorems can be replaced by  $V^p$ , which can lead to a serious improvement in the size of CONV-LACK. A related idea is to consider the  $\beta$  in predictor models  $\mathcal{F}$  as in (2) as an additional parameter, to be equipped with a prior and fitted to the data. Since  $q$  and  $\check{p}$  in Theorem 1 may then refer to different predictors with different  $\beta$ 's,  $\beta$  will act as a ‘local’ learning rate whereas  $\eta$ , shared by all distributions, is a ‘global’ learning rate. Preliminary investigations suggest that this leads to better bounds in some cases.

### Acknowledgments

Supported in part by the IST Programme of the EU, under the PASCAL NoE, IST-2002-506778. I would like to thank Andrew Barron, Tim van Erven and Rui Castro for some very useful discussions.

### References

- J.Y. Audibert. *PAC-Bayesian statistical learning theory*. PhD thesis, Université Paris VI, 2004.
- A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.
- A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Occam’s razor. *Information Processing Letters*, 24:377–380, 1987.
- L. Breiman. Statistical modeling: the two cultures (with discussion). *Statistical Science*, 16(3):199 –215, 2001.
- O. Catoni. *PAC-Bayesian Supervised Classification*. Lecture Notes-Monograph Series. IMS, 2007.
- A.P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77:605–611, 1982. Discussion: pages 611–613.
- J.L. Doob. Application of the theory of martingales. In *Le Calcul de Probabilités et ses Applications. Colloques Internationaux du Centre National de la Recherche Scientifique*, pages 23–27, Paris, 1949.
- P. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.
- P. Grünwald and J. Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66(2-3):119–149, 2007. DOI 10.1007/s10994-007-0716-7.
- P. D. Grünwald. Viewing all models as “probabilistic”. In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT’ 99)*, pages 171–182, 1999.

- B. Kleijn and A. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, 34(2), 2006.
- J.Q. Li. *Estimation of Mixture Models*. PhD thesis, Yale University, New Haven, CT, 1999.
- E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27: 1808–1829, 1999.
- D. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.
- A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.
- Tong Zhang. From  $\epsilon$ -entropy to KL entropy: analysis of minimum information complexity density estimation. *Annals of Statistics*, 34(5):2180–2210, 2006.

## Appendix A. Additional Proofs

**1. Condition (B) can be made to hold by adding one or a few distributions to  $\mathcal{P}$**  For example, in the classification case, it suffices to include the trivial distribution  $p_0$  into  $\mathcal{P}$ , where, for all  $x \in \mathcal{X}$ ,  $p_0(Y = 1 | X = x) = p_0(Y = 0 | X = x) = 1/2$ , and assign it some prior mass  $w_0(p_0) > 0$ . Then for all sequences  $z^n$ , for all  $\kappa \geq 1$ ,

$$\begin{aligned} -\log w(\ddot{p}_\kappa) &\leq -\log w(\ddot{p}_1) \leq -\log w(\ddot{p}_1) - \log \ddot{p}_1(z_\dagger^n) \\ &\leq -\log w_0 - \log p_0(z_\dagger^n) = -\log w_0 + n \log 2 \leq nL_{\max}, \end{aligned} \quad (44)$$

for suitably chosen  $L_{\max}$ . Clearly, this approach extends to all  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$  with finite or compact  $\mathcal{Y}$ . If  $\mathcal{Y}$  is not compact, then, by our assumption **(D)** that  $V(\mathcal{P}, P^*) < \infty$ , the interval  $[a, b]$  with  $a = \text{ess inf}_{Z \in \mathcal{Z}, p \in \mathcal{P}} p(Z_\dagger)$  and  $b = \text{ess sup}_{p \in \mathcal{P}, Z \in \mathcal{Z}} p(Z_\dagger)$ , is bounded. It then suffices to include a density  $p_{a,b}$  with prior  $w_0$  such that, for all  $x \in \mathcal{X}$ ,  $p_{a,b}(Y = \cdot | X = x)$  is uniform on  $[a, b]$ . If the end points on the interval are not known, we can discretize candidate end points to integers and put a prior  $v$  on these end points satisfying, for both end points  $c \in \{a, b\}$ ,  $-\log v(c) \approx 2 \log \min\{|c|, 1\}$  (Grünwald, 2007). We can define the defective distribution  $p_0(y | x) := \max_{a,b} p_{a,b}(y | x)v(a)v(b)$  and repeat the reasoning in (44).

**2. Equivalence of (4) and (12)** We will only show that equivalence holds in the idealized case in which the best-approximating  $q$  is actually a member of  $\mathcal{P}$ . This should be sufficient, since the goal of establishing equivalence is merely to give some intuition about the meaning of (12) (that's why we only put it in the appendix); the equivalence is not needed in any of our results, whose proofs invariably rely on (12) rather than (4). Assume then that  $D(p^* \| q) = \inf_{p \in \mathcal{P}} D(p^* \| p)$  and that  $q \in \mathcal{P}$ . We only show equivalence for  $\eta = 1$ ; extension to other  $\eta$  is immediate. We first prove (4)  $\Rightarrow$  (12). If (4) holds, then for all  $p \in \mathcal{P}$ ,  $D^*(q \| \text{MIX}(q, p, \lambda))$  has its minimum  $\lambda^\circ$  (as defined above Lemma 9) at  $\lambda^\circ = 0$ . It then follows by Lemma 9 that  $d_1^*(q \| p) = d_1^*(q \| \text{MIX}(q, p, 1)) \geq 0$ , and we see that (12) holds.

We next prove (12)  $\Rightarrow$  (4). Suppose that (12) holds for all  $p \in \mathcal{P}$ . Without loss of generality let  $\mathcal{P} = \{p_1, p_2, \dots\}$ . Then for any  $p'$  in the convex hull of  $\mathcal{P}$ , say  $p' = \sum_{j=1}^{\infty} \alpha_j p_j$

with all  $\alpha_j \geq 0$  and  $\sum \alpha_j = 1$ , we have  $E^*(p'(Z_i)/q(Z_i)) = \sum_{j=1}^{\infty} \alpha_j E^*(p_j(Z_i)/q(Z_i)) \leq 1$ . Thus  $E^*(p'(Z_i)/q(Z_i)) - 1 \leq 0$  and hence, by Proposition 10, part (1), the derivative of the concave function  $D^*(\text{MIX}(q, p', \lambda) \| q)$  is  $\leq 0$  at  $\lambda = 0$ . This implies that  $D(p^* \| q) \leq D(p^* \| \text{MIX}(q, p', \lambda))$  for all  $\lambda > 0$ , in particular  $D(p^* \| q) \leq D(p^* \| p')$ ; this shows that (4) holds.

### 3. Lemma 9 – proof of Proposition 10

Differentiation gives:

$$g'(\gamma) = \frac{d}{d\gamma} g(\gamma) = -E^* \left( \frac{p(Z) - p'(Z)}{(1-\gamma)p(Z) + \gamma p'(Z)} \right), \text{ in particular } g'(0) = E^* \left( \frac{p'(Z)}{p(Z)} \right) - 1, \quad (45)$$

which shows the first part of 1. We now first show part 2(a) and (b). Note that

$$g''(\gamma) = -E^* \left( \frac{p(Z) - p'(Z)}{(1-\gamma)p(Z) + \gamma p'(Z)} \right)^2.$$

For all  $\gamma \in [0, 1]$  and all  $Z$ , the denominator inside the expectation must be bounded from below by  $\underline{p} := \text{ess inf}_{Z,p \in \mathcal{P}} p(Z)$  and from above by  $\bar{p} := \text{ess sup}_{Z,p \in \mathcal{P}} p(Z)$ . We thus have, for all  $\gamma \in [0, 1]$ ,  $1/\bar{p}^2 E^*(p' - p)^2 \leq |g''(\gamma)| \leq 1/\underline{p}^2 E^*(p' - p)^2$ . Now suppose first that  $E^*(p'(Z_i) - p(Z_i))^2 = 0$ . Then  $p'(Z_i) = p(Z_i)$  almost surely, and  $g'(\lambda) = g''(\lambda) = 0$  on  $[0, 1]$ , almost surely, and part (2a) follows. If  $E^*(p'(Z_i) - p(Z_i))^2 > 0$ , then  $p'(Z_i) \neq p(Z_i)$  with positive probability, and part (2b) immediately follows. Having now established that  $g''(\lambda) \leq 0$  on  $[0, 1]$ , it follows by definition of  $\lambda^\circ$  that  $g'(0) > 0$  iff  $\lambda^\circ > 0$ . And since we assume  $D^*(p \| p') \geq 0$ , we have  $g(1) \leq g(0)$ , which implies that if  $\lambda^\circ > 0$ , then  $g'(\lambda^\circ) = 0$  and  $g'(1) \leq 0$ .

### 4. Proof of Proposition 12 (log-Bonferroni)

Let  $X_j := e^{-Y_j}$  and  $b_j = e^{-a_j}$ . The assumption implies that, for any collection  $\{K_j\}_{j \in \mathcal{J}}$  of positive real numbers, for all  $j \in \mathcal{J}$ ,

$$P^* \left( X_j \geq b_j e^{-f_j(K_j)} \right) \geq 1 - e^{-K_j},$$

or equivalently,

$$P^* \left( X_j < b_j e^{-f_j(K_j)} \right) < e^{-K_j}.$$

Now, for fixed  $K \geq 0$ , define  $K_j = K - \log w_j$ . By the union bound, we have

$$P^*(\mathcal{A}) < \sum_{j \in \mathcal{J}} e^{-K + \log w_j} = \sum_{j \in \mathcal{J}} w_j e^{-K}.$$

where  $\mathcal{A}$  is the event that for some  $j \in \mathcal{J}$ ,  $X_j < b_j e^{-f_j(K - \log w_j)}$ . This implies that for  $\bar{\mathcal{A}}$ , the complement of  $\mathcal{A}$ , we have

$$P^*(\bar{\mathcal{A}}) \geq 1 - \sum_{j \in \mathcal{J}} w_j e^{-K}.$$

The result now follows by noting that the event whose probability is bounded in the statement of the proposition is just  $\bar{\mathcal{A}}$ .

**5. Proof of Proposition 15 (Barron and Li's (1999) result)** Define, for given  $\eta, p^*, p$  and  $q$ , the *affinity relative to  $x$*  as  $A_x = \int_{y \in \mathcal{Y}} p^*(y | x) \cdot \left( \frac{p(y|x)}{q(y|x)} \right)^\eta$  and let

$$p^{\text{new}}(y | x) = \frac{1}{A(x)} p^*(y | x) \cdot \left( \frac{p(y|x)}{q(y|x)} \right)^\eta.$$

Next, recall that the *squared Hellinger distance*, between densities  $p$  and  $q$  on  $\mathcal{Y}$ , denoted by us as  $H^2(q, p)$ , is defined as

$$H^2(q, p) := \int_y (\sqrt{q(y)} - \sqrt{p(y)})^2 = 2 \left( 1 - \int_y \sqrt{q(y)p(y)} \right).$$

Also recall that the ordinary (nongeneralized) Rényi divergence of order  $1/2$  is given by  $d_{1/2}(q, p) = -2 \log \int_y \sqrt{q(y)p(y)} dy$ . Now, for  $u \geq 0$ , we have  $1 - u \leq -\log u$  (this follows from  $\log(1 + z) \geq z$  and substituting  $z = u - 1$ ). This implies the following well-known general relation between squared Hellinger distance and Rényi divergence:

$$H^2(q, p) \leq d_{1/2}(q \| p). \quad (46)$$

Moreover (Barron and Cover, 1991), when the ratio between  $p$  and  $q$  is bounded, then the standard (nongeneralized) KL divergence is upper-bounded by a multiple of the squared Hellinger distance. Yang and Barron (1999) proved the following precise relation:

$$D(q \| p) \leq (2 + \log V) H^2(q, p). \quad (47)$$

We will now use (46) and (47) to prove our result. We first need to clarify notation: for given  $x$ , the *generalized Rényi divergence between  $p$  and  $q$ , given  $x$*  is denoted as  $d_\eta^{*|x}(q(\cdot | x) \| p(\cdot | x))$  and defined as

$$d_\eta^{*|x}(q(\cdot | x) \| p(\cdot | x)) = -\frac{1}{\eta} \log E^* \left[ \left( \frac{p(Y|x)}{q(Y|x)} \right)^\eta \mid X = x \right].$$

We have for all  $C_\eta \geq 2 + 2\eta \log V$ , for each  $x \in \mathcal{X}$ , each  $\eta \leq 1$ ,

$$\begin{aligned} E^* \left[ -\log \frac{p(Y|x)}{q(Y|x)} \mid X = x \right] &= \frac{1}{\eta} \cdot E^* \left[ \log p^*(Y|x) - \log \left( p^*(Y|x) \left( \frac{q(Y|x)}{p(Y|x)} \right)^\eta \right) \mid X = x \right] \\ &\quad + \frac{1}{\eta} (\log A_x - \log A_x) \\ &= \frac{1}{\eta} D(p^*(\cdot|x) \| p^{\text{new}}(\cdot|x)) - \frac{1}{\eta} \log A_x \\ &\leq \frac{1}{\eta} C_\eta H^2(p^*(\cdot|x), p^{\text{new}}(\cdot|x)) - \frac{1}{\eta} \log A_x \\ &\leq \frac{1}{\eta} C_\eta d_{1/2}(p^*(\cdot|x), p^{\text{new}}(\cdot|x)) - \frac{1}{\eta} \log A_x \\ &= C_\eta \left( d_{\eta/2}^{*|x}(q(\cdot|x) \| p(\cdot|x)) + \frac{1}{\eta} \log A_x \right) - \frac{1}{\eta} \log A_x \\ &= C_\eta d_{\eta/2}^{*|x}(q(\cdot|x) \| p(\cdot|x)) + \frac{1}{\eta} (C_\eta - 1) \log A_x. \end{aligned}$$

Here the first two equalities are just rewriting. In the first inequality we used (47), the fact that  $P^*$ -almost surely,  $\sup_{X,Y} p^{\text{new}}(Y|X)/p^*(Y|X) \leq V^{2\eta}$ , and the fact that  $D(\cdot \| \cdot) \geq 0$ , and the second inequality is just (46). In the fifth line we used some basic rewriting. Using

the notation  $E_X^*$  to denote expectation of  $X$  under  $P_X^*$ , the marginal distribution of  $X$ , we thus get:

$$\begin{aligned} D^*(q\|p) &\leq C_\eta E_X^*[d_{\eta/2}^{*|X}(q(\cdot|X)\|p(\cdot|X))] + (C_\eta - 1)\frac{1}{\eta}E_X^*[\log A_X] \\ &\leq C_\eta d_{2/\eta}^*(q\|p) + (C_\eta - 1)\frac{1}{\eta}\log E_X^*[A_X] \\ &= C_\eta d_{\eta/2}^*(q\|p) - (C_\eta - 1)d_\eta^*(q\|p). \end{aligned}$$

where the second inequality is Jensen's and the final equality is just the definition of Rényi divergence.

GRÜNWALD

# Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization

**Elad Hazan**

*Technion - Israel Institute of Technology*

EHAZAN@IE.TECHNION.AC.IL

**Satyen Kale**

*Yahoo! Research*

SKALE@YAHOO-INC.COM

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We give a novel algorithm for stochastic strongly-convex optimization in the gradient oracle model which returns an  $O(\frac{1}{T})$ -approximate solution after  $T$  gradient updates. This rate of convergence is optimal in the gradient oracle model. This improves upon the previously known best rate of  $O(\frac{\log(T)}{T})$ , which was obtained by applying an online strongly-convex optimization algorithm with regret  $O(\log(T))$  to the batch setting.

We complement this result by proving that any algorithm has expected regret of  $\Omega(\log(T))$  in the online stochastic strongly-convex optimization setting. This lower bound holds even in the full-information setting which reveals more information to the algorithm than just gradients. This shows that any online-to-batch conversion is inherently suboptimal for stochastic strongly-convex optimization. This is the first formal evidence that online convex optimization is strictly more difficult than batch stochastic convex optimization.

**Keywords:** Stochastic Optimization, Regret Minimization

## 1. Introduction

Stochastic convex optimization has an inherently different flavor than standard convex optimization. In the stochastic case, a crucial resource is the number of data samples from the function to be optimized. This resource limits the precision of the output: given few samples there is simply not enough information to compute the optimum up to a certain precision. The error arising from this lack of information is called the *estimation error*.

The estimation error is independent of the choice of optimization algorithm, and it is reasonable to choose an optimization method whose precision is of the same order of magnitude as the sampling error: lesser precision is suboptimal, whereas much better precision is pointless (this issue was extensively discussed in Bottou and Bousquet (2007) and Shalev-Shwartz and Srebro (2008)). This makes first-order methods ideal for stochastic convex optimization: their error decreases as a polynomial in the number of iterations, usually one iteration per data point, and each iteration is extremely efficient.

In this paper we consider first-order methods for stochastic convex optimization. Formally, the problem of stochastic convex optimization is the minimization of a convex function on a convex, compact domain  $\mathcal{K}$ :

$$\min_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x}).$$

The stochasticity is in the access model: the only access to  $F$  is via a stochastic gradient oracle, which given any point  $\mathbf{x} \in \mathcal{K}$ , produces a random vector  $\hat{\mathbf{g}}$  whose expectation is a subgradient of  $F$  at the point  $\mathbf{x}$ , i.e.  $\mathbb{E}[\hat{\mathbf{g}}] \in \partial F(\mathbf{x})$ , where  $\partial F(\mathbf{x})$  denotes the subdifferential set of  $F$  at  $\mathbf{x}$ .

An important special case is when  $F(\mathbf{x}) = \mathbb{E}_Z[f(\mathbf{x}, Z)]$  (the expectation being taken over a random variable  $Z$ ), where for every fixed  $z$ ,  $f(\mathbf{x}, z)$  is a convex function of  $\mathbf{x}$ . The goal is to minimize  $F$  while given a sample  $z_1, z_2, \dots$  drawn independently from the unknown distribution of  $Z$ . A prominent example of this formulation is the problem of support vector machine training (see Shalev-Shwartz et al. (2009)).

An algorithm for stochastic convex optimization is allowed a budget of  $T$  calls to the gradient oracle. It sequentially queries the gradient oracle at consecutive points  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$ , and produces an approximate solution  $\bar{\mathbf{x}}$ . The *rate of convergence* of the algorithm is the expected excess cost of the point  $\bar{\mathbf{x}}$  over the optimum, i.e.  $\mathbb{E}[F(\bar{\mathbf{x}})] - \min_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x})$ , where the expectation is taken over the randomness in the gradient oracle and the internal random seed of the algorithm. The paramount parameter for measuring this rate is in terms of  $T$ , the number of gradient oracle calls.

Our first and main contribution is the first algorithm to attain the optimal rate of convergence in the case where  $F$  is  $\lambda$ -strongly convex, and the gradient oracle is  $G$ -bounded (see precise definitions in Section 2.1). After  $T$  gradient updates, the algorithm returns a solution which is  $O(\frac{1}{T})$ -close in cost to the optimum. Formally, we prove

**Theorem 1** *Assume that  $F$  is  $\lambda$ -strongly convex and the gradient oracle is  $G$ -bounded. Then there exists an algorithm that after at most  $T$  gradient updates returns a vector  $\bar{\mathbf{x}}$  such that for any  $\mathbf{x}^* \in \mathcal{K}$  we have*

$$\mathbb{E}[F(\bar{\mathbf{x}})] - F(\mathbf{x}^*) \leq O\left(\frac{G^2}{\lambda T}\right).$$

This matches the lower bound of Agarwal et al. (2010) up to constant factors.

The previously best known rate was  $O(\frac{\log(T)}{T})$ , and follows by converting a more general online convex optimization algorithm of Hazan et al. (2007) to the batch setting. This standard online-to-batch reduction works as follows. In the online convex optimization setting, in each round  $t = 1, 2, \dots, T$ , a decision maker (represented by an algorithm  $\mathcal{A}$ ) chooses a point  $\mathbf{x}_t$  in convex domain  $\mathcal{K}$ , and incurs a cost  $f_t(\mathbf{x}_t)$  for an adversarially chosen convex cost function  $f_t$ . In this model performance is measured by the *regret*, defined as

$$\text{Regret}(\mathcal{A}) := \sum_{t=1}^T f_t(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^T f_t(\mathbf{x}). \quad (1)$$

A regret minimizing algorithm is one that guarantees that the regret grows like  $o(T)$ . Given such an algorithm, one can perform batch stochastic convex optimization by setting  $f_t$  to be the function<sup>1</sup>  $f(\cdot, z_t)$ . A simple analysis then shows that the cost of the average point,  $\bar{x} = \frac{1}{T} \sum_{t=1}^T x_t$ , converges to the optimum cost at the rate of the *average* regret, which converges to zero.

---

1. Note that we are assuming that we have full access to the function  $f(\cdot, z_t)$  here, rather than just gradient information.

The best previously known convergence rates for stochastic convex optimization were obtained using this online-to-batch reduction, and thus these rates were equal to the average regret of the corresponding online convex optimization algorithm. While it is known that for general convex optimization, this online-to-batch reduction gives the optimal rate of convergence, such a result was not known for stochastic strongly-convex functions. In this paper we show that for stochastic strongly-convex functions, minimizing regret is strictly more difficult than batch stochastic strongly-convex optimization.

More specifically, the best known regret bound for  $\lambda$ -strongly-convex cost functions with gradients bounded in norm by  $G$  is  $O\left(\frac{G^2 \log(T)}{\lambda}\right)$  Hazan et al. (2007). This regret bound holds even for adversarial, not just stochastic, strongly-convex cost functions. A matching lower bound was obtained in Takimoto and Warmuth (2000) for the adversarial setting.

Our second contribution in this paper is a matching lower bound for strongly-convex cost functions that holds *even in the stochastic setting*, i.e. if the cost functions are sampled i.i.d from an unknown distribution. Formally:

**Theorem 2** *For any online decision-making algorithm  $\mathcal{A}$ , there is a distribution over  $\lambda$ -strongly-convex cost functions with norms of gradients bounded by  $G$  such that*

$$\mathbb{E}[\text{Regret}(\mathcal{A})] = \Omega\left(\frac{G^2 \log(T)}{\lambda}\right).$$

Hence, our new rate of convergence of  $O\left(\frac{G^2}{\lambda T}\right)$  is the first to separate the complexity of stochastic and online strongly-convex optimization. The following table summarizes our contribution with respect to the previously known bounds. The setting is assumed to be stochastic  $\lambda$ -strongly-convex functions with gradient norms bounded by  $G$ .

	Previous bound	New bound here
Convergence rate	$O\left(\frac{G^2 \log(T)}{\lambda T}\right)$ [Hazan et al. (2007)]	$O\left(\frac{G^2}{\lambda T}\right)$
Regret	$\Omega(1)$ [trivial bound]	$\Omega\left(\frac{G^2 \log(T)}{\lambda}\right)$

We also sharpen our results: Theorem 1 bounds the expected excess cost of the solution over the optimum by  $O\left(\frac{1}{T}\right)$ . We can also show high probability bounds. In situations where it is possible to evaluate  $F$  at any given point efficiently, simply repeating the algorithm a number of times and taking the best point found bounds the excess cost by  $O\left(\frac{G^2 \log(\frac{1}{\delta})}{\lambda T}\right)$  with probability at least  $1 - \delta$ . In more realistic situations where it is not possible to evaluate  $F$  efficiently, we can still modify the algorithm so that with high probability, the actual excess cost of the solution is bounded by  $O\left(\frac{\log \log(T)}{T}\right)$ :

**Theorem 3** *Assume that  $F$  is  $\lambda$ -strongly convex, and the gradient oracle is  $G$ -bounded. Then for any  $\delta > 0$ , there exists an algorithm that after at most  $T$  gradient updates returns a vector  $\bar{\mathbf{x}}$  such that with probability at least  $1 - \delta$ , for any  $\mathbf{x}^* \in \mathcal{K}$  we have*

$$F(\bar{\mathbf{x}}) - F(\mathbf{x}^*) \leq O\left(\frac{G^2(\log(\frac{1}{\delta}) + \log \log(T))}{\lambda T}\right).$$

### 1.1. Related work

For an in depth discussion of first-order methods, the reader is referred to Bertsekas (1999).

The study of lower bounds for stochastic convex optimization was undertaken in Nemirovski and Yudin (1983), and recently extended and refined in Agarwal et al. (2010).

Online convex optimization was introduced in Zinkevich (2003). Optimal lower bounds for the convex case, even in the stochastic setting, of  $\Omega(\sqrt{T})$  are simple and given in Cesa-Bianchi and Lugosi (2006). For exp-concave cost functions, Ordentlich and Cover (1998) gave a  $\Omega(\log T)$  lower bound on the regret, even when the cost functions are sampled according to a known distribution. For strongly convex functions, no non-trivial stochastic lower bound was known. Takimoto and Warmuth (2000) gave a  $\Omega(\log T)$  lower bound in the regret for adaptive adversaries. Abernethy et al. (2009) put this lower bound in a general framework for min-max regret minimization.

It has been brought to our attention that Juditsky and Nesterov (2010) have recently published a technical report that has a very similar algorithm to ours, and also obtain an  $O(\frac{1}{T})$  convergence rate. This work however was done independently and a preliminary version was published on arXiv (Hazan and Kale (2010)) before the technical report of Juditsky and Nesterov was available.

## 2. Setup and Background

### 2.1. Stochastic convex optimization

Consider the setting of stochastic convex optimization of a convex function  $F$  over a convex, compact set  $\mathcal{K} \subseteq \mathbb{R}^n$ . Let  $\mathbf{x}^*$  be a point in  $\mathcal{K}$  where  $F$  is minimized. We make the following assumptions:

1. We assume that  $F$  is  **$\lambda$ -strongly convex**: i.e., for any two points  $\mathbf{x}, \mathbf{y} \in \mathcal{K}$  and any  $\alpha \in [0, 1]$ , we have

$$F(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) \leq \alpha F(\mathbf{x}) + (1 - \alpha)F(\mathbf{y}) - \frac{\lambda}{2}\alpha(1 - \alpha)\|\mathbf{x} - \mathbf{y}\|^2.$$

$F$  is  $\lambda$ -strongly-convex, for example, if  $F(\mathbf{x}) = \mathbb{E}_Z[f(\mathbf{x}, Z)]$  and  $f(\cdot, z)$  is  $\lambda$ -strongly-convex for every  $z$  in the support of  $Z$ .

This implies  $F$  satisfies the following inequality (to see this, set  $\mathbf{y} = \mathbf{x}^*$ , divide by  $\alpha$ , and take the limit as  $\alpha \rightarrow 0^+$ ):

$$F(\mathbf{x}) - F(\mathbf{x}^*) \geq \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}^*\|^2 \tag{2}$$

This inequality holds even if  $\mathbf{x}^*$  is on the boundary of  $\mathcal{K}$ . In fact, (2) is the *only* requirement on the convexity of  $F$  for the analysis to work, we will simply assume that (2) holds.

2. Assume we have oracle access to compute an unbiased estimator of a subgradient of  $F$  at any point  $\mathbf{x}$ , denoted  $\hat{\mathbf{g}} \in \partial F(\mathbf{x})$ , whose  $\ell_2$  norm bounded by some known value  $\|\hat{\mathbf{g}}\| \leq G$ . Such a gradient oracle is called  **$G$ -bounded**.

3. Assume that the domain  $\mathcal{K}$  is endowed with an efficiently computable **projection operator**  $\Pi_{\mathcal{K}}(\mathbf{y}) = \arg \min_{\mathbf{x} \in \mathcal{K}} \|\mathbf{x} - \mathbf{y}\|$ .

Assumptions 1 and 2 above imply the following lemma:

**Lemma 4** *For all  $\mathbf{x} \in \mathcal{K}$ , we have  $F(\mathbf{x}) - F(\mathbf{x}^*) \leq \frac{2G^2}{\lambda}$ .*

**Proof** For any  $\mathbf{x} \in \mathcal{K}$ , let  $\hat{\mathbf{g}} \in \partial F(\mathbf{x})$  be a subgradient of  $F$  at  $\mathbf{x}$  such that  $\|\hat{\mathbf{g}}\| \leq G$  (using assumption 2). Then by the convexity of  $F$ , we have  $F(\mathbf{x}) - F(\mathbf{x}^*) \leq \hat{\mathbf{g}} \cdot (\mathbf{x} - \mathbf{x}^*)$ , so that  $F(\mathbf{x}) - F(\mathbf{x}^*) \leq G\|\mathbf{x} - \mathbf{x}^*\|$ . But assumption 1 implies that  $F(\mathbf{x}) - F(\mathbf{x}^*) \geq \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$ . Putting these together, we get that  $\|\mathbf{x} - \mathbf{x}^*\| \leq \frac{2G}{\lambda}$ . Finally, we get  $F(\mathbf{x}) - F(\mathbf{x}^*) \leq G\|\mathbf{x} - \mathbf{x}^*\| \leq \frac{2G^2}{\lambda}$ . Since  $\mathbf{x}^*$  is the minimizer of  $F$  on  $\mathcal{K}$ , the lemma follows. ■

## 2.2. Online Convex Optimization and Regret

Recall the setting of online convex optimization given in the introduction. In each round  $t = 1, 2, \dots, T$ , a decision-maker needs to choose a point  $\mathbf{x}_t \in \mathcal{K}$ , a convex set. Then nature provides a convex cost function  $f_t : \mathcal{K} \rightarrow \mathbb{R}$ , and the decision-maker incurs the cost  $f_t(\mathbf{x}_t)$ . The (adversarial) regret of the decision-maker is defined to be

$$\text{AdversarialRegret} := \sum_{t=1}^T f_t(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^T f_t(\mathbf{x}). \quad (3)$$

When the cost functions  $f_t$  are drawn i.i.d. from some unknown distribution  $D$ , (stochastic) regret is traditionally defined measured with respect to the expected cost function,  $F(\mathbf{x}) = \mathbb{E}_D[f_1(\mathbf{x})]$ :

$$\text{StochasticRegret} := \mathbb{E}_D \left[ \sum_{t=1}^T f_t(\mathbf{x}_t) \right] - T \min_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x}). \quad (4)$$

In either case, if the decision-making algorithm is randomized, then we measure the performance by the expectation of the regret taken over the random seed of the algorithm.

When cost functions are drawn i.i.d. from an unknown distribution  $D$ , it is easy to check that

$$\mathbb{E}_D \left[ \min_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^T f_t(\mathbf{x}) \right] \leq \min_{\mathbf{x} \in \mathcal{K}} \mathbb{E}_D \left[ \sum_{t=1}^T f_t(\mathbf{x}) \right],$$

by considering the point  $\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathcal{K}} \mathbb{E}_D \left[ \sum_{t=1}^T f_t(\mathbf{x}) \right]$ . So

$$\mathbb{E}_D[\text{AdversarialRegret}] \geq \text{StochasticRegret}.$$

Thus, for the purpose of proving lower bounds on the regret (expected regret in the case of randomized algorithms), it suffices to prove such bounds for StochasticRegret. We prove such lower bounds in Section 5. For notational convenience, henceforth the term “regret” refers to StochasticRegret.

### 3. The optimal algorithm and its analysis

Our algorithm is an extension of stochastic gradient descent. The new feature is the introduction of “epochs” inside of which standard stochastic gradient descent is used, but in each consecutive epoch the learning rate decreases exponentially.

---

**Algorithm 1** EPOCH-GD

---

```

1: Input: parameters  $\eta_1, T_1$  and total time  $T$ .
2: Initialize  $\mathbf{x}_1^1 \in \mathcal{K}$  arbitrarily, and set  $k = 1$ .
3: while  $\sum_{i=1}^k T_i \leq T$  do
4:   // Start epoch  $k$ 
5:   for  $t = 1$  to  $T_k$  do
6:     Query the gradient oracle at  $\mathbf{x}_t^k$  to obtain  $\hat{\mathbf{g}}_t$ 
7:     Update

$$\mathbf{x}_{t+1}^k = \prod_{\mathcal{K}} (\mathbf{x}_t^k - \eta_k \hat{\mathbf{g}}_t)$$

8:   end for
9:   Set  $\mathbf{x}_1^{k+1} = \frac{1}{T_k} \sum_{t=1}^{T_k} \mathbf{x}_t^k$ 
10:  Set  $T_{k+1} \leftarrow 2T_k$  and  $\eta_{k+1} \leftarrow \eta_k/2$ .
11:  Set  $k \leftarrow k + 1$ 
12: end while
13: return  $\mathbf{x}_1^k$ 
```

---

Our main result is the following theorem, which immediately implies Theorem 1.

**Theorem 5** *Set the parameters  $T_1 = 2$  and  $\eta_1 = \frac{1}{\lambda}$  in the EPOCH-GD algorithm. The final point  $\mathbf{x}_1^k$  returned by the algorithm has the property that  $\mathbb{E}[F(\mathbf{x}_1^k)] - F(\mathbf{x}^*) \leq \frac{8G^2}{\lambda T}$ . The total number of gradient updates is at most  $T$ .*

The intra-epoch use of standard gradient decent is analyzed using the following Lemma from Zinkevich (2003), which we prove here for completeness:

**Lemma 6 (Zinkevich (2003))** *Let  $\|\hat{\mathbf{g}}_t\| \leq G$ . Apply  $T$  iterations of the update  $\mathbf{x}_{t+1} = \prod_{\mathcal{K}} (\mathbf{x}_t - \eta \hat{\mathbf{g}}_t)$ . Then*

$$\frac{1}{T} \sum_{t=1}^T \hat{\mathbf{g}}_t \cdot (\mathbf{x}_t - \mathbf{x}^*) \leq \frac{\eta G^2}{2} + \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{2\eta T}.$$

**Proof** We use  $\|\mathbf{x}_t - \mathbf{x}^*\|^2$  as a potential function. Let  $\mathbf{x}'_{t+1} = \mathbf{x}_t - \eta \hat{\mathbf{g}}_t$ . We have

$$\|\mathbf{x}'_{t+1} - \mathbf{x}^*\|^2 = \eta^2 \|\hat{\mathbf{g}}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta \hat{\mathbf{g}}_t \cdot (\mathbf{x}_t - \mathbf{x}^*) \leq \eta^2 G^2 + \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta \hat{\mathbf{g}}_t \cdot (\mathbf{x}_t - \mathbf{x}^*).$$

Since  $\mathbf{x}_{t+1} = \prod_{\mathcal{K}} (\mathbf{x}'_{t+1})$  and  $\mathbf{x}^* \in \mathcal{K}$ , the fact that projections of an external point on a convex set reduces the distance to any point inside it (see Zinkevich (2003)) implies that  $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}'_{t+1} - \mathbf{x}^*\|^2$ . Putting these together, and rearranging, we get

$$\hat{\mathbf{g}}_t \cdot (\mathbf{x}_t - \mathbf{x}^*) \leq \frac{\eta G^2}{2} + \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2}{2\eta}.$$

Summing up over all  $t = 1, 2, \dots, T$ , we get the stated bound.  $\blacksquare$

**Lemma 7** *Apply  $T$  iterations of the update  $\mathbf{x}_{t+1} = \prod_{\mathcal{K}}(\mathbf{x}_t - \eta \hat{\mathbf{g}}_t)$ , where  $\hat{\mathbf{g}}_t$  is an unbiased estimator for a subgradient  $\mathbf{g}_t$  of  $F$  at  $\mathbf{x}_t$  satisfying  $\|\hat{\mathbf{g}}_t\| \leq G$ . Then*

$$\frac{1}{T} \mathbb{E} \left[ \sum_{t=1}^T F(\mathbf{x}_t) \right] - F(\mathbf{x}^*) \leq \frac{\eta G^2}{2} + \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{2\eta T}.$$

By convexity of  $F$ , we have the same bound for  $\mathbb{E}[F(\bar{\mathbf{x}})] - F(\mathbf{x}^*)$ , where  $\bar{\mathbf{x}} = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t$ .

**Proof** For a random variable  $X$  measurable w.r.t. the randomness until round  $t$ , let  $\mathbb{E}_{t-1}[X]$  denote its expectation conditioned on the randomness until round  $t-1$ . By the convexity of  $F$ , we get

$$F(\mathbf{x}_t) - F(\mathbf{x}^*) \leq \mathbf{g}_t \cdot (\mathbf{x}_t - \mathbf{x}^*) = \mathbb{E}_{t-1}[\hat{\mathbf{g}}_t \cdot (\mathbf{x}_t - \mathbf{x}^*)],$$

since  $\mathbb{E}_{t-1}[\hat{\mathbf{g}}_t] = \mathbf{g}_t$  and  $\mathbb{E}_{t-1}[\mathbf{x}_t] = \mathbf{x}_t$ . Taking expectations of the inequality, we get that

$$\mathbb{E}[F(\mathbf{x}_t)] - F(\mathbf{x}^*) \leq \mathbb{E}[\hat{\mathbf{g}}_t \cdot (\mathbf{x}_t - \mathbf{x}^*)].$$

Summing up over all  $t = 1, 2, \dots, T$  and applying Lemma 6, we get the required bound.  $\blacksquare$

Define  $V_k = \frac{G^2}{2^{k-2}\lambda}$ , and notice that the algorithm sets  $T_k = \frac{4G^2}{\lambda V_k}$  and  $\eta_k = \frac{V_k}{2G^2}$ . Define  $\Delta_k = F(\mathbf{x}_1^k) - F(\mathbf{x}^*)$ . Using Lemma 7 we prove the following key lemma:

**Lemma 8** *For any  $k$ , we have  $\mathbb{E}[\Delta_k] \leq V_k$ .*

**Proof** We prove this by induction on  $k$ . The claim is true for  $k = 1$  since  $\Delta_k \leq \frac{2G^2}{\lambda}$  by Lemma 4. Assume that  $\mathbb{E}[\Delta_k] \leq V_k$  for some  $k \geq 1$  and now we prove it for  $k+1$ . For a random variable  $X$  measurable w.r.t. the randomness defined up to epoch  $k+1$ , let  $\mathbb{E}_k[X]$  denote its expectation conditioned on all the randomness up to epoch  $k$ . By Lemma 7 we have

$$\begin{aligned} \mathbb{E}_k[F(\mathbf{x}_1^{k+1})] - F(\mathbf{x}^*) &\leq \frac{\eta_k G^2}{2} + \frac{\|\mathbf{x}_1^k - \mathbf{x}^*\|^2}{2\eta_k T_k} \\ &\leq \frac{\eta_k G^2}{2} + \frac{\Delta_k}{\eta_k T_k \lambda}, \end{aligned}$$

since  $\Delta_k = F(\mathbf{x}_1^k) - F(\mathbf{x}^*) \geq \frac{\lambda}{2} \|\mathbf{x}_1^k - \mathbf{x}^*\|^2$  by  $\lambda$ -strong convexity of  $F$ . Hence, we get

$$\mathbb{E}[\Delta_{k+1}] \leq \frac{\eta_k G^2}{2} + \frac{\mathbb{E}[\Delta_k]}{\eta_k T_k \lambda} \leq \frac{\eta_k G^2}{2} + \frac{V_k}{\eta_k T_k \lambda} = \frac{V_k}{2} = V_{k+1},$$

as required. The second inequality uses the induction hypothesis, and the last two equalities use the definition of  $V_k$  and the values  $\eta_k = \frac{V_k}{2G^2}$  and  $T_k = \frac{4G^2}{\lambda V_k}$ .  $\blacksquare$

We can now prove our main theorem:

**Proof** [Theorem 5.] The number of epochs made are given by the largest value of  $k$  satisfying  $\sum_{i=1}^k T_i \leq T$ , i.e.

$$\sum_{i=1}^k 2^i = 2(2^k - 1) \leq T.$$

This value is  $k^\dagger = \lfloor \log_2(\frac{T}{2} + 1) \rfloor$ . The final point output by the algorithm is  $\mathbf{x}_1^{k^\dagger+1}$ . Applying Lemma 8 to  $k^\dagger + 1$  we get

$$\mathbb{E}[F(\mathbf{x}_1^{k^\dagger+1})] - F(\mathbf{x}^*) = \mathbb{E}[\Delta_{k^\dagger+1}] \leq V_{k^\dagger+1} = \frac{G^2}{2^{k^\dagger-1}\lambda} \leq \frac{8G^2}{\lambda T},$$

as claimed. The while loop in the algorithm ensures that the total number of gradient updates is naturally bounded by  $T$ .  $\blacksquare$

#### 4. High probability bounds

While EPOCH-GD algorithm has a  $O(\frac{1}{T})$  rate of convergence, this bound is only on the expected excess cost of the final solution. In applications we usually need the rate of convergence to hold with high probability. Markov's inequality immediately implies that with probability  $1 - \delta$ , the actual excess cost is at most a factor of  $\frac{1}{\delta}$  times the stated bound. While this guarantee might be acceptable for not too small values of  $\delta$ , it becomes useless when  $\delta$  gets really small.

There are two ways of remedying this. The easy way applies if it is possible to evaluate  $F$  efficiently at any given point. Then we can divide the budget of  $T$  gradient updates into  $\ell = \log_2(1/\delta)$  consecutive intervals of  $\frac{T}{\ell}$  rounds each, and run independent copies of EPOCH-GD in each. Finally, we take the  $\ell$  solutions obtained, and output the best one (i.e. the one with the minimum  $F$  value). Applying Markov's inequality to every run of EPOCH-GD, with probability at least  $1/2$ , we obtain a point with excess cost at most  $\frac{64G^2\ell}{\lambda T} = \frac{64G^2 \log_2(1/\delta)}{\lambda T}$ , and so with probability at least  $1 - 2^{-\ell} = 1 - \delta$ , the best point has excess cost at most  $\frac{64G^2 \log_2(1/\delta)}{\lambda T}$ . This finishes the description of the easy way to obtain high probability bounds.

The easy way fails if it is not possible to evaluate  $F$  efficiently at any given point. For this situation, we now describe how using essentially the same algorithm with slightly different parameters, we can get a high probability guarantee on the quality of the solution. The only difference in the new algorithm, dubbed EPOCH-GD-PROJ, is that the update in line 7 requires a projection onto a smaller set, and becomes

$$\text{Update } \mathbf{x}_{t+1}^k = \prod_{\mathcal{K} \cap B(\mathbf{x}_1^k, \sqrt{2V_k/\lambda})} (\mathbf{x}_t^k - \eta_k \hat{\mathbf{g}}_t) \quad (5)$$

Here  $B(\mathbf{x}, r)$  denotes the  $\ell_2$  ball of radius  $r$  around the point  $x$ , and  $V_k = \frac{G^2}{2^{k-2}\lambda}$  as defined earlier. Since the intersection of two convex sets is also a convex set, the above projection can be computed via a convex program.

We prove the following high probability result, which in turn directly implies Theorem 3.

**Theorem 9** *Given  $\delta > 0$  for success probability  $1 - \delta$ , set  $\tilde{\delta} = \frac{\delta}{k^\dagger}$  for  $k^\dagger = \lfloor \log_2(\frac{T}{300} + 1) \rfloor$ . Set the parameters  $T_1 = 300 \log(1/\tilde{\delta})$  and  $\eta_1 = \frac{1}{3\lambda}$  in the EPOCH-GD-PROJ algorithm. The final point  $\mathbf{x}_1^k$  returned by the algorithm has the property that with probability at least  $1 - \delta$ , we have*

$$F(\mathbf{x}_1^k) - F(\mathbf{x}^*) \leq \frac{1200G^2 \log(1/\tilde{\delta})}{\lambda T}.$$

The total number of gradient updates is at most  $T$ .

The following lemma is analogous to Lemma 7, but provides a high probability guarantee.

**Lemma 10** *Let  $D$  be an upper bound on  $\|\mathbf{x}_1 - \mathbf{x}^*\|$ . Apply  $T$  iterations of the update  $\mathbf{x}_{t+1} = \prod_{K \cap B(\mathbf{x}_1, D)} (\mathbf{x}_t - \eta \hat{\mathbf{g}}_t)$ , where  $\hat{\mathbf{g}}_t$  is an unbiased estimator for the subgradient of  $F$  at  $\mathbf{x}_t$  satisfying  $\|\hat{\mathbf{g}}_t\| \leq G$ . Then for any  $\delta \in (0, 1)$ , with probability at least  $1 - \delta$  we have*

$$\frac{1}{T} \sum_{t=1}^T F(\mathbf{x}_t) - F(\mathbf{x}^*) \leq \frac{\eta G^2}{2} + \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{2\eta T} + \frac{4GD\sqrt{2\log(1/\delta)}}{\sqrt{T}}.$$

By the convexity of  $F$ , the same bound also holds for  $F(\bar{\mathbf{x}}) - F(\mathbf{x}^*)$ , where  $\bar{\mathbf{x}} = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t$ .

**Proof** Using the same notation as in the proof of Lemma 7, let  $\mathbb{E}_{t-1}[\hat{\mathbf{g}}_t] = \mathbf{g}_t$ , a subgradient of  $F$  at  $\mathbf{x}_t$ . Since as before,  $\mathbb{E}_{t-1}[\hat{\mathbf{g}}_t \cdot (\mathbf{x}_t - \mathbf{x}^*)] = \mathbf{g}_t \cdot (\mathbf{x}_t - \mathbf{x}^*)$ , the following defines as a martingale difference sequence:

$$X_t = \mathbf{g}_t \cdot (\mathbf{x}_t - \mathbf{x}^*) - \hat{\mathbf{g}}_t \cdot (\mathbf{x}_t - \mathbf{x}^*).$$

Note that  $\|\mathbf{g}_t\| = \|\mathbb{E}_{t-1}[\hat{\mathbf{g}}_t]\| \leq \mathbb{E}_{t-1}[\|\hat{\mathbf{g}}_t\|] \leq G$ , and so we can bound  $|X_t|$  as follows:

$$|X_t| \leq \|\mathbf{g}_t\| \|\mathbf{x}_t - \mathbf{x}^*\| + \|\hat{\mathbf{g}}_t\| \|\mathbf{x}_t - \mathbf{x}^*\| \leq 4GD,$$

where the last inequality uses the fact that  $\mathbf{x}^*, \mathbf{x}_t \in B(\mathbf{x}_1, D)$ , and hence by the triangle inequality  $\|\mathbf{x}_t - \mathbf{x}^*\| \leq \|\mathbf{x}_t - \mathbf{x}_1\| + \|\mathbf{x}_1 - \mathbf{x}^*\| \leq 2D$ .

By Azuma's inequality (see Lemma 12), with probability at least  $1 - \delta$ , the following holds:

$$\frac{1}{T} \sum_{t=1}^T \mathbf{g}_t \cdot (\mathbf{x}_t - \mathbf{x}^*) - \frac{1}{T} \sum_{t=1}^T \hat{\mathbf{g}}_t \cdot (\mathbf{x}_t - \mathbf{x}^*) \leq \frac{4GD\sqrt{2\log(1/\delta)}}{\sqrt{T}}. \quad (6)$$

By the convexity of  $F$ , we have  $F(\mathbf{x}_t) - F(\mathbf{x}^*) \leq \mathbf{g}_t \cdot (\mathbf{x}_t - \mathbf{x}^*)$ . Then, by using Lemma 6 and inequality (6), we get the claimed bound.  $\blacksquare$

We now prove the analogue of Lemma 8. In this case, the result holds with high probability. As before, define  $V_k = \frac{G^2}{2^{k-2}\lambda}$ , and notice that the algorithm sets  $T_k = \frac{600G^2 \log(1/\tilde{\delta})}{\lambda V_k}$  and  $\eta_k = \frac{V_k}{6G^2}$ .

**Lemma 11** For any  $k$ , with probability  $(1 - \tilde{\delta})^{k-1}$  we have  $\Delta_k \leq V_k$ .

**Proof** We prove this by induction on  $k$ . The claim is true for  $k = 1$  since  $\Delta_1 \leq \frac{2G^2}{\lambda}$  by Lemma 4. Assume that  $\Delta_k \leq V_k$  for some  $k \geq 1$  with probability at least  $(1 - \tilde{\delta})^{k-1}$  and now we prove it for  $k+1$ . We condition on the event that  $\Delta_k \leq V_k$ . Since  $\Delta_k \geq \frac{\lambda}{2} \|\mathbf{x}_1^k - \mathbf{x}^*\|^2$  by  $\lambda$ -strong convexity, this conditioning implies that  $\|\mathbf{x}_1^k - \mathbf{x}^*\| \leq \sqrt{2V_k/\lambda}$ , which explains the specification (see (5) for the radius of the ball for the projection in line 7 of EPOCH-GD-PROJ). So Lemma 10 applies with  $D = \sqrt{2V_k/\lambda}$  and hence we have with probability at least  $1 - \tilde{\delta}$ ,

$$\begin{aligned}\Delta_{k+1} &= F(\mathbf{x}_1^{k+1}) - F(\mathbf{x}^*) \\ &\leq \frac{\eta_k G^2}{2} + \frac{\|\mathbf{x}_1^k - \mathbf{x}^*\|^2}{2\eta_k T_k} + \frac{8G\sqrt{V_k}\sqrt{\log(1/\tilde{\delta})}}{\sqrt{\lambda T_k}} \quad (\text{by Lemma 10}) \\ &\leq \frac{\eta_k G^2}{2} + \frac{V_k}{\eta_k T_k \lambda} + \frac{8G\sqrt{V_k}\sqrt{\log(1/\tilde{\delta})}}{\sqrt{\lambda T_k}},\end{aligned}$$

since  $V_k \geq \Delta_k \geq \frac{\lambda}{2} \|\mathbf{x}_1^k - \mathbf{x}^*\|^2$  as above. For  $T_k = \frac{600G^2 \log(1/\tilde{\delta})}{\lambda V_k}$  and  $\eta_k = \frac{V_k}{6G^2}$  we get

$$F(\mathbf{x}_1^{k+1}) - F(\mathbf{x}^*) \leq \frac{V_k}{12} + \frac{V_k}{100 \log(1/\tilde{\delta})} + \frac{V_k}{3} \leq \frac{V_k}{2} = V_{k+1}.$$

Factoring in the conditioned event, which happens with probability at least  $(1 - \tilde{\delta})^{k-1}$ , overall, we get that  $\Delta_{k+1} \leq V_{k+1}$  with probability at least  $(1 - \tilde{\delta})^k$ . ■

We can now prove our high probability theorem:

**Proof** [Theorem 9] As in the proof of Theorem 1, we get that final epoch is  $k^\dagger = \lfloor \log_2(\frac{T}{300} + 1) \rfloor$ . The final point output is  $\mathbf{x}_1^{k^\dagger+1}$ .

By Lemma 11, we have with probability at least  $(1 - \tilde{\delta})^{k^\dagger}$  that

$$F(\mathbf{x}_1^{k^\dagger+1}) - F(\mathbf{x}^*) = \Delta_{k^\dagger+1} \leq V_{k^\dagger+1} = \frac{G^2}{2^{k^\dagger-1}\lambda} \leq \frac{1200G^2 \log(1/\tilde{\delta})}{\lambda T},$$

as claimed. Since  $\tilde{\delta} = \frac{\delta}{k^\dagger}$ , and hence  $(1 - \tilde{\delta})^{k^\dagger} \geq 1 - \delta$  as needed. The while loop in the algorithm ensures that the total number of gradient updates is bounded by  $T$ . ■

For completeness we state Azuma's inequality for martingales used in the proof above:

**Lemma 12 (Azuma's inequality)** Let  $X_1, \dots, X_T$  be a martingale difference sequence. Suppose that  $|X_t| \leq b$ . Then, for  $\delta > 0$ , we have

$$\Pr \left[ \sum_{t=1}^T X_t \geq b\sqrt{2T \ln(1/\delta)} \right] \leq \delta.$$

## 5. Lower bounds on stochastic strongly convex optimization

In this section we prove Theorem 2 and show that any algorithm (deterministic or randomized) for online stochastic strongly-convex optimization must have  $\Omega(\log(T))$  regret on some distribution. We start by proving a  $\Omega(\log T)$  lower bound for the case when the cost functions are 1-strongly convex and the gradient oracle is 1-bounded, and fine tune these parameters in the next subsection via an easy reduction.

In our analysis, we need the following standard lemma, which we reprove here for completeness. Here, for two distributions  $P, P'$  defined on the same probability space,  $d_{TV}(P, P')$  is the total variation distance, i.e.

$$d_{TV}(P, P') = \sup_A |P(A) - P'(A)|$$

where the supremum ranges over all events  $A$  in the probability space.

Let  $B_p$  be the Bernoulli distribution on  $\{0, 1\}$  with probability of obtaining 1 equal to  $p$ . Let  $B_p^n$  denote the product measure on  $\{0, 1\}^n$  induced by taking  $n$  independent Bernoulli trials according to  $B_p$  (thus,  $B_p^1 = B_p$ ).

**Lemma 13** *Let  $p, p' \in [\frac{1}{4}, \frac{3}{4}]$  such that  $|p' - p| \leq 1/8$ . Then*

$$d_{TV}(B_p^n, B_{p'}^n) \leq \frac{1}{2} \sqrt{(p' - p)^2 n}.$$

**Proof** Pinsker's inequality says that  $d_{TV}(B_p^n, B_{p'}^n) \leq \sqrt{\frac{1}{2} \text{RE}(B_p^n \| B_{p'}^n)}$ , where  $\text{RE}(B_p^n \| B_{p'}^n) = \mathbb{E}_{X \sim B_p^n} [\ln \frac{B_p^n(X)}{B_{p'}^n(X)}]$  is the relative entropy between  $B_p^n$  and  $B_{p'}^n$ . To bound  $\text{RE}(B_p^n \| B_{p'}^n)$ , note that the additivity of the relative entropy for product measures implies that

$$\text{RE}(B_p^n \| B_{p'}^n) = n \text{RE}(B_p \| B_{p'}) = n \left[ p \log \left( \frac{p}{p'} \right) + (1-p) \log \left( \frac{1-p}{1-p'} \right) \right], \quad (7)$$

Without loss of generality, assume that  $p' \geq p$ , and let  $p' = p + \varepsilon$ , where  $0 \leq \varepsilon \leq 1/8$ . Using the Taylor series expansion of  $\log(1+x)$ , we get the following bound

$$p \log \left( \frac{p}{p'} \right) + (1-p) \log \left( \frac{1-p}{1-p'} \right) = \sum_{i=1}^{\infty} \left[ \frac{(-1)^i}{p^{i-1}} + \frac{1}{(1-p)^{i-1}} \right] \varepsilon^i \leq \sum_{i=2}^{\infty} 4^{i-1} \varepsilon^i \leq \frac{\varepsilon^2}{2},$$

for  $\varepsilon \leq 1/8$ . Plugging this (7) and using Pinsker's inequality, we get the stated bound. ■

We now turn to showing our lower bound on expected regret. We consider the following online stochastic strongly-convex optimization setting: the domain is  $\mathcal{K} = [0, 1]$ . For every  $p \in [\frac{1}{4}, \frac{3}{4}]$ , define a distribution over strongly-convex cost functions parameterized by  $p$  as follows: choose  $X \in \{0, 1\}$  from  $B_p$ , and return the cost function

$$f(x) = (x - X)^2$$

With some abuse of notation, we use  $B_p$  to denote this distribution over cost functions.

Under distribution  $B_p$ , the expected cost function  $F$  is

$$F(x) := \mathbf{E}[f(x)] = p(x-1)^2 + (1-p)x^2 = x^2 + 2px + p = (x-p)^2 + c_p,$$

where  $c_p = p - p^2$ . The optimal point is therefore  $x^* = p$ , with expected cost  $c_p$ . The regret for playing a point  $x$  (i.e. excess cost over the minimal expected cost) is

$$F(x) - F(x^*) = (x-p)^2 + c_p - c_p = (x-p)^2.$$

Now let  $\mathcal{A}$  be a deterministic<sup>2</sup> algorithm for online stochastic strongly-convex optimization. Since the cost functions until time  $t$  are specified by a bit string  $X \in \{0, 1\}^{t-1}$  (i.e. the cost function at time  $t$  is  $(x - X_t)^2$ ), we can interpret the algorithm as a function that takes a variable length bit string, and produces a point in  $[0, 1]$ , i.e. with some abuse of notation,

$$\mathcal{A} : \{0, 1\}^{\leq T} \longrightarrow [0, 1],$$

where  $\{0, 1\}^{\leq T}$  is the set of all bit strings of length up to  $T$ .

Now suppose the cost functions are drawn from  $B_p$ . Fix a round  $t$ . Let  $X$  be the  $t-1$  bit string specifying the cost functions so far. Note that  $X$  has distribution  $B_p^{t-1}$ . For notational convenience, denote by  $\Pr_p[\cdot]$  and  $\mathbb{E}_p[\cdot]$  the probability of an event and the expectation of a random variable when the cost functions are drawn from  $B_p$ , and since these are defined by the bit string  $X$ , they are computed over the product measure  $B_p^{t-1}$ .

Let the point played by  $\mathcal{A}$  at time  $t$  be  $x_t = \mathcal{A}(X)$ . The regret (conditioned on the choice of  $X$ ) in round  $t$  is then

$$\text{regret}_t := (\mathcal{A}(X) - p)^2,$$

and thus the expected (over the choice of  $X$ ) regret of  $\mathcal{A}$  in round  $t$  is  $\mathbb{E}_p[\text{regret}_t] = \mathbb{E}_p[(\mathcal{A}(X) - p)^2]$ .

We now show that for any round  $t$ , for two distributions over cost functions  $B_p$  and  $B_{p'}$  that are close (in terms of  $|p - p'|$ ), but not too close, the regret of  $\mathcal{A}$  on at least one of the two distributions must be large.

**Lemma 14** *Fix a round  $t$ . Let  $\varepsilon \leq \frac{1}{8\sqrt{t}}$  be a parameter. Let  $p, p' \in [\frac{1}{4}, \frac{3}{4}]$  such that  $2\varepsilon \leq |p - p'| \leq 4\varepsilon$ . Then we have*

$$\mathbb{E}_p[\text{regret}_t] + \mathbb{E}_{p'}[\text{regret}_t] \geq \frac{1}{4}\varepsilon^2.$$

**Proof** Assume without loss of generality that  $p' \geq p + 2\varepsilon$ . Let  $X$  and  $X'$  be  $(t-1)$ -bit vectors parameterizing the cost functions drawn from  $B_p^{t-1}$  and  $B_{p'}^{t-1}$  respectively. Then

$$\mathbb{E}_p[\text{regret}_t] + \mathbb{E}_{p'}[\text{regret}_t] = \mathbb{E}_p[(\mathcal{A}(X) - p)^2] + \mathbb{E}_{p'}[(\mathcal{A}(X') - p')^2].$$

Now suppose the stated bound does not hold. Then by Markov's inequality, we have

$$\Pr_p[(\mathcal{A}(X) - p)^2 < \varepsilon^2] \geq 3/4,$$

---

2. We will remove the deterministic requirement shortly and allow randomized algorithms.

or in other words,

$$\Pr_p[\mathcal{A}(X) < p + \varepsilon] \geq 3/4. \quad (8)$$

Similarly, we can show that

$$\Pr_{p'}[\mathcal{A}(X') > p + \varepsilon] \geq 3/4, \quad (9)$$

since  $p' \geq p + 2\varepsilon$ . Now define the event

$$A := \{Y \in \{0, 1\}^{t-1} : \mathcal{A}(Y) > p + \varepsilon\}.$$

Now (8) implies that  $\Pr_p(A) < 1/4$  and (9) implies that  $\Pr_{p'}(A) \geq 3/4$ . But then by Lemma 13 we have

$$\frac{1}{2} < |\Pr_p(A) - \Pr_{p'}(A)| \leq d_{TV}(B_p^{t-1}, B_{p'}^{t-1}) \leq \frac{1}{2}\sqrt{(p' - p)^2(t - 1)} \leq \frac{1}{2}\sqrt{16\varepsilon^2(t - 1)} \leq \frac{1}{4},$$

a contradiction.  $\blacksquare$

We now show how to remove the deterministic requirement on  $\mathcal{A}$ :

**Corollary 15** *The bound of Lemma 14 holds even if  $\mathcal{A}$  is randomized:*

$$\mathbb{E}_{p,R}[\text{regret}_t] + \mathbb{E}_{p',R}[\text{regret}_t] \geq \frac{1}{4}\varepsilon^2,$$

where  $\mathbb{E}_{p,R}[\cdot]$  denotes the expectation computed over the random seed  $R$  of the algorithm as well as the randomness in the cost functions.

**Proof** Fixing the random seed  $R$  of  $\mathcal{A}$ , we get a deterministic algorithm, and then Lemma 14 gives the following bound on the sum of the conditional expected regrets:

$$\mathbb{E}_p[\text{regret}_t|R] + \mathbb{E}_{p'}[\text{regret}_t|R] \geq \frac{1}{4}\varepsilon^2.$$

Now taking expectations over the random seed  $R$ , we get the desired bound.  $\blacksquare$

Thus, from now on we allow  $\mathcal{A}$  to be randomized. We now show the desired lower bound on the expected regret:

**Theorem 16** *The expected regret for algorithm  $\mathcal{A}$  is at least  $\Omega(\log(T))$ .*

**Proof** We prove this by showing that there is one value of  $p \in [\frac{1}{4}, \frac{3}{4}]$  such that regret of  $\mathcal{A}$  when cost functions are drawn from  $B_p$  is at least  $\Omega(\log(T))$ .

We assume that  $T$  is of the form  $16 + 16^2 + \dots + 16^k = \frac{1}{15}(16^{k+1} - 16)$  for some integer  $k$ : if it isn't, we ignore all rounds  $t > T'$ , where  $T' = \frac{1}{15}(16^{k^*+1} - 16)$  for  $k^* = \lfloor \log_{16}(15T + 16) - 1 \rfloor$ , and show that in the first  $T'$  rounds the algorithm can be made to have  $\Omega(\log(T))$  regret. We now divide the time periods  $t = 1, 2, \dots, T'$  into consecutive epochs of length  $16, 16^2, \dots, 16^{k^*}$ . Thus, epoch  $k$ , denoted  $E_k$ , has length  $16^k$ , and consists of the time periods  $t = \frac{1}{15}(16^k - 16) + 1, \dots, \frac{1}{15}(16^{k+1} - 16)$ . We show the following claim now:

**Claim 1** *There exists a collection of nested intervals,  $[\frac{1}{4}, \frac{3}{4}] \supseteq I_1 \supseteq I_2 \supseteq I_3 \supseteq \dots$ , such that interval  $I_k$  corresponds to epoch  $k$ , with the property that  $I_k$  has length  $4^{-(k+3)}$ , and for every  $p \in I_k$ , for at least half the rounds  $t$  in epoch  $k$ , algorithm  $\mathcal{A}$  has  $\mathbb{E}_{p,R}[\text{regret}_t] \geq \frac{1}{8} \cdot 16^{-(k+3)}$ .*

As a consequence of this claim, we get that there is a value of  $p \in \bigcap_k I_k$  such that in every epoch  $k$ , the total regret is

$$\sum_{t \in E_k} \frac{1}{8} \cdot 16^{-(k+3)} \geq \frac{1}{2} 16^k \cdot \frac{1}{8} \cdot 16^{-(k+3)} = \frac{1}{16^4}.$$

Thus, the regret in every epoch is  $\Omega(1)$ . Since there are  $k^* = \Theta(\log(T))$  epochs total, the regret of the algorithm is at least  $\Omega(\log(T))$ . So now we prove the claim.

**Proof** We build the nested collection of intervals iteratively as follows. For notational convenience, define  $I_0$  to be some arbitrary interval of length  $4^{-3}$  inside  $[\frac{1}{4}, \frac{3}{4}]$ . Suppose for some  $k \geq 0$  we have found the interval  $I_k = [a, a + 4^{-(k+3)}]$ . We want to find the interval  $I_{k+1}$  now. For this, divide up  $I_k$  into 4 equal quarters of length  $\varepsilon = 4^{-(k+4)}$ , and consider the first and fourth quarters, viz.  $L = [a, a + 4^{-(k+4)}]$  and  $R = [a + 3 \cdot 4^{-(k+4)}, a + 4^{-(k+3)}]$ . We now show that one of  $L$  or  $R$  is a valid choice for  $I_{k+1}$ , and so the construction can proceed.

Suppose  $L$  is not a valid choice for  $I_{k+1}$ , because there is some point  $p \in L$  such that for more than half the rounds  $t$  in  $E_{k+1}$ , we have  $\mathbb{E}_{p,R}[\text{regret}_t] < 16^{-(k+1)}$ . Then we show that  $R$  is a valid choice for  $I_{k+1}$  as follows. Let  $H = \{t \in E_{k+1} : \mathbb{E}_{p,R}[\text{regret}_t] < \frac{1}{8} \cdot 16^{-(k+4)}\}$ . Now, we claim that for all  $p' \in R$ , and all  $t \in H$ , we must have  $\mathbb{E}_{p',R}[\text{regret}_t] > \frac{1}{8} \cdot 16^{-(k+4)}$ , which would imply that  $R$  is a valid choice for  $I_{k+1}$ , since by assumption,  $|H| \geq \frac{1}{2}|E_{k+1}|$ .

To show this we apply Lemma 14. Fix any  $p' \in R$  and  $t \in F$ . First, note that  $\varepsilon = 4^{-(k+4)} \leq \frac{1}{8\sqrt{t}}$ , since  $t \leq 16^{k+2}$ . Next, we have  $p' - p \geq 2\varepsilon$  (since we excluded the middle two quarters of  $I_k$ ), and  $|p - p'| \leq 4\varepsilon$  (since  $I_k$  has length  $4^{-(k+3)}$ ). Then Lemma 14 implies that

$$\mathbb{E}_{p,R}[\text{regret}_t] + \mathbb{E}_{p',R}[\text{regret}_t] \geq \frac{1}{4} \cdot 16^{-(k+4)},$$

which implies that  $\mathbb{E}_{p',R}[\text{regret}_t] \geq \frac{1}{8} \cdot 16^{-(k+4)}$  since  $\mathbb{E}_{p,R}[\text{regret}_t] < \frac{1}{8} \cdot 16^{-(k+4)}$ , as required. ■

■

### 5.1. Dependence on the gradient bound and on strong convexity

A simple corollary of the previous proof gives us tight lower bounds in terms of the natural parameters of the problem: the strong-convexity parameter  $\lambda$  and the upper bound on the norm of the subgradients  $G$ . The following Corollary implies Theorem 2.

**Corollary 17** *For any algorithm  $\mathcal{A}$ , there is distribution over  $\lambda$ -strongly convex cost functions with gradients bounded in norm by  $G$  such that the expected regret of  $\mathcal{A}$  is  $\Omega\left(\frac{G^2 \log(T)}{\lambda}\right)$ .*

**Proof** The online convex optimization setting we design is very similar: let  $\lambda, G \geq 0$  be given parameters. The domain is  $\mathcal{K} = [0, \frac{G}{\lambda}]$ . In round  $t$ , we choose  $X_t \in \{0, 1\}$  from  $B_p$ , and return

$$f_t(x) = \frac{\lambda}{2} \left( x - \frac{G}{\lambda} X_t \right)^2$$

as the cost function. Notice that the cost functions are always  $\lambda$ -strongly convex, and in addition, for any  $x \in \mathcal{K}$ , the gradient of the cost function at  $x$  is bounded in norm by  $G$ .

Denote  $x' = \frac{\lambda x}{G}$  to be the scaled decision  $x$ , mapping it from  $\mathcal{K}$  to  $[0, 1]$ . The expectation of the cost when playing  $x \in \mathcal{K}$  is given by

$$\mathbb{E}[f_t(x)] = \mathbb{E}_{X \sim B_p} \left[ \frac{\lambda}{2} \left( x - \frac{G}{\lambda} X_t \right)^2 \right] = \frac{G^2}{2\lambda} \mathbb{E}[(x' - X_t)^2] \quad (10)$$

Given an algorithm  $\mathcal{A}$  for this online convex optimization instance, we derive another algorithm,  $\mathcal{A}'$ , which plays points  $x' \in \mathcal{K}' = [0, 1]$  and receives the cost function  $(x' - X_t)^2$  in round  $t$  (i.e. the setting considered in Section 5). When  $\mathcal{A}$  plays  $x_t$  in round  $t$  and obtains cost function  $\frac{\lambda}{2} (x - \frac{G}{\lambda} X_t)^2$ , the algorithm  $\mathcal{A}'$  plays the point  $x'_t = \frac{\lambda}{G} x_t$  and receives the cost function  $(x' - X_t)^2$ .

The optimum point for the setting of  $\mathcal{A}$  is  $\frac{G}{\lambda} p$ , with expected cost  $\frac{G^2}{2\lambda}$  times the expected cost for the optimum point  $p$  for the setting of  $\mathcal{A}'$ . By equation (10), the cost of  $\mathcal{A}$  is  $\frac{G^2}{2\lambda}$  times that of  $\mathcal{A}'$ . Hence, the regret of  $\mathcal{A}$  is  $\frac{G^2}{2\lambda}$  times that of  $\mathcal{A}'$ .

By Theorem 16, there is a value of  $p$  such that the expected regret of  $\mathcal{A}'$  is  $\Omega(\log T)$ , and hence the expected regret of  $\mathcal{A}$  is  $\Omega\left(\frac{G^2 \log(T)}{\lambda}\right)$ , as required. ■

## 6. Conclusions

We have given an algorithm for stochastic strongly-convex optimization with an optimal rate of convergence  $O(\frac{1}{T})$ . The algorithm itself has an appealing feature of returning the average of the most recent points (rather than all points visited by the algorithm as in previous approaches). This is an intuitive feature which hopefully works well in practice for important applications such as support vector machine training.

Our analysis deviates from the common template of designing a regret minimization algorithm and then using online-to-batch conversion. In fact, we show that the latter approach is inherently suboptimal by our new lower bound on the regret of online algorithms for stochastic cost functions. This combination of results formally shows that the batch stochastic setting is strictly easier than its online counterpart, giving us tighter bounds.

A few questions remain open. The high-probability bound algorithm EPOCH-GD-PROJ has an extra factor of  $O(\log \log(T))$  in its convergence rate. Is it possible to devise an algorithm that has  $O(\frac{1}{T})$  convergence rate with high probability? We believe the answer is yes; the  $O(\log \log(T))$  is just an artefact of the analysis. In fact, as we mention in Section 4, if it is possible to evaluate  $F$  efficiently at any given point, then this dependence can be removed. Also, our lower bound proof is somewhat involved. Are there easier information theoretic arguments to give similar lower bounds?

## Acknowledgments

We thank an anonymous referee for several useful suggestions.

## References

- Jacob Abernethy, Alekh Agarwal, Peter L. Bartlett, and Alexander Rakhlin. A stochastic view of optimal regret through minimax duality. In *COLT*, 2009.
- Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *arXiv:1009.0571v1*, 2010.
- Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, September 1999. ISBN 1886529000.
- Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *NIPS*, 2007.
- Nicolò Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- Elad Hazan and Satyen Kale. An optimal algorithm for stochastic strongly-convex optimization. June 2010. URL <http://arxiv.org/abs/1006.2425>.
- Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.
- Anatoli Juditsky and Yuri Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. August 2010. URL <http://hal.archives-ouvertes.fr/docs/00/50/89/33/PDF/Strong-hal.pdf>.
- Arkadi S. Nemirovski and David B. Yudin. Problem complexity and method efficiency in optimization. John Wiley UK/USA, 1983.
- Erik Ordentlich and Thomas M. Cover. The cost of achieving the best portfolio in hindsight. *Math. Oper. Res.*, 23:960–982, November 1998.
- Shai Shalev-Shwartz and Nathan Srebro. SVM optimization: inverse dependence on training set size. In *ICML*, pages 928–935, 2008.
- Shai Shalev-Shwartz, Ohad Shamir, Karthik Sridharan, and Nati Srebro. Stochastic convex optimization. In *COLT*, 2009.
- Eiji Takimoto and Manfred K. Warmuth. The minimax strategy for gaussian density estimation. In *COLT*, pages 100–106, 2000.
- Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.

# A Close Look to Margin Complexity and Related Parameters

**Michael Kallweit**

and **Hans Ulrich Simon**

*Fakultät für Mathematik*

*Ruhr-Universität Bochum*

*D-44780 Bochum, Germany*

MICHAEL.KALLWEIT@RUB.DE

HANS.SIMON@RUB.DE

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

Concept classes can canonically be represented by sign-matrices, i.e., by matrices with entries 1 and  $-1$ . The question whether a sign-matrix (concept class)  $A$  can be learned by a machine that performs large margin classification is closely related to the “margin complexity” associated with  $A$ . We consider several variants of margin complexity, reveal how they are related to each other, and we reveal how they are related to other notions of learning-theoretic relevance like SQ-dimension, CSQ-dimension, and the Forster bound.

## 1. Introduction

Large margin classifiers implicitly use a feature map that transforms linearly inseparable data into feature vectors that can be linearly separated in feature space so as to achieve a (hopefully) large margin, which then leads to a small generalization error. Concept classes  $\mathcal{C}$  over a domain  $\mathcal{X}$  that can potentially be learned by large margin classifiers must therefore admit a linear arrangement consisting of hyperplanes and points (with the hyperplanes representing the concepts from  $\mathcal{C}$  and the points representing the instances from  $\mathcal{X}$ ) such that positive (resp. negative) examples appear as points lying in a positive (resp. negative) halfspace and having a certain “safety distance” to the corresponding separating hyperplane. In practice, a large “hard” margin cannot often be achieved so that softer notions of a margin come into play. Soft margins can be achieved by arrangements which occasionally put points close to the separating hyperplane (small margin) or, may be, even in the wrong half-space (negative margin). But one would still insist on something like a large “average margin”. This will (roughly) be captured by our notion of average margin complexity.

In this paper, we deal with sign-matrices (which represent finite concept classes: every column is a Boolean function and the rows correspond to the instances), and we study various notions of (average) margin complexity, where “high (average) margin complexity” means that even the best arrangement achieves a small (average) margin only. Sign-matrices with high average margin complexity represent concept classes that cannot be successfully learned by large margin classifiers (thereby indicating the limitations of this approach). In a seminal paper, Forster (2002) presented a lower bound on the margin complexity (and on the dimension complexity which, however, is not considered in this paper) of a sign-matrix in terms of its spectral norm. Loosely speaking, the Forster bound measures the “amount of orthogonality” that is contained in  $A$ . It achieves its maximal value for Hadamard matrices.

In this paper, we generalize the Forster bound by imposing probability distributions on the rows and the columns of  $A$ . In case of uniform distributions, the generalized bound collapses to the original-one. It is easy to construct matrices for which the original bound evaluates to a small number but, when the probability distributions are chosen properly, the generalized bound becomes large.

The SQ model of learning was introduced by Kearns (1998). It is an elegant abstraction from the PAC learning model of Valiant (1984). In this model, instead of having direct access to random examples (as in the PAC learning model) the learner obtains information about random examples via an oracle that provides estimates of various statistics about the unknown concept. Kearns showed that any learning algorithm that is successful in the SQ model can be converted, without much loss of efficiency, into a learning algorithm that is successful in the PAC learning model despite noise uniformly applied to the class labels of the examples. Furthermore, almost all concept classes known to be efficiently learnable in the PAC learning model can efficiently be learned in the SQ model too. This is why the SQ model attracted a lot of attention in the Computational Learning Community. “Correlational Statistical Queries” are statistical queries of a special form and lead in the obvious way to the CSQ model of learning. As shown by Bshouty and Feldman (2002), the two models coincide in case of a fixed distribution, but, as shown by Feldman (2008), the SQ model is exponentially more powerful in the distribution-independent setting. Blum et al. (2003) have shown that the number of statistical queries needed for weak SQ-learning under a fixed distribution is polynomially related to the SQ-dimension (defined w.r.t. the same distribution). Feldman (2008) has defined the CSQ-dimension and has shown that it plays a similar role for distribution-independent weak learning in the CSQ model. In the same paper, he shows furthermore that CSQ-learnability is equivalent to evolvability (a framework introduced by Valiant (2009) and designed so as to catch the computational aspects of evolution).

In this paper, we will be concerned with the relations that hold between the various notions of margin complexity on one hand and parameters like SQ-dimension or CSQ-dimension on the other hand. The main results are as follows:<sup>1</sup>

- By means of semi-definite programming duality, we show in Section 3 that the optimal margin (the smallest distance between one of the points and one of the hyperplanes in a margin-optimal arrangement) coincides with the optimal average margin (the average distance between points and hyperplanes in an optimal arrangement) provided that the underlying distribution (according to which the average is taken) is chosen in a worst-case fashion. More formally:

$$\text{mc}(A) = \max_Y \overline{\text{mc}}_Y(A)$$

- In Section 3.1, we complement the well-known lower bound  $\sqrt{mn}/\|A\|_2$  on the average margin complexity (w.r.t. uniform distributions on the rows and the columns of  $A$ ) by the upper bound  $mn/\|A\|_{tr}$ . More formally:

$$\frac{\sqrt{mn}}{\|A\|_2} \leq \overline{\text{mc}}(A) \leq \frac{mn}{\|A\|_{tr}}$$

---

1. The formal definitions needed for a precise understanding of the following statements are given in Section 2.

- In Section 3.2, we identify two families of matrices whose average margin complexity (w.r.t. uniform distributions on the rows and the columns) is determined exactly: Hadamard matrices and matrices composed of all reflections of a single Boolean function.
- In Sections 4, 5, and 6, we determine relations between the various notions of margin complexity, the various versions of the Forster bound, the SQ-dimension and the CSQ-dimension. Here is a quick overview over our results:
  - Let  $p, q$  denote vectors assigning probabilities to the rows and the columns of a matrix  $A$ , respectively. We show that the SQ-dimension w.r.t.  $p$  of a sign-matrix  $A$  is polynomially related to the generalized Forster bound and also polynomially related to the average margin complexity of  $A$  according to

$$\text{SQdim}_p(A) < 2 \cdot \max_q \text{FB}_{p,q}(A)^2 \leq 2 \cdot \max_q \overline{\text{mc}}_{p,q}(A)^2 < 2 \cdot \text{SQdim}_p(A) \cdot (\text{SQdim}_p(A) + 1)^2 .$$

- We reveal the following polynomial relationship between the CSQ-dimension of a matrix  $A \in \mathbb{R}^{m \times n}$  and the margin complexity of  $A$ :

$$\text{mc}(A) \leq \text{CSQdim}(A)^{1.5} \quad \text{and} \quad \text{CSQdim}(A) \leq \lceil 32 \ln(4mn) \cdot \text{mc}(A)^2 \rceil$$

- We show that

$$\text{SQdim}(A^T) < 2 \cdot \text{SQdim}(A) \cdot (\text{SQdim}(A) + 1)^2 .$$

This improves on  $\text{SQdim}(A^\top) \leq 32 \cdot \text{SQdim}(A)^4$ , a result that had been shown before by Sherstov (2008).

- We show that the generalized Forster bound is, up to a polynomial, not more effective than simply applying the classical bound to a properly chosen submatrix  $A''$  of  $A$ . More formally:

$$\max_{A''} \text{FB}(A'') \leq \max_{p,q} \text{FB}_{p,q}(A) < 64 \cdot (1 + o(1)) \cdot \max_{A''} \text{FB}(A'')^9$$

Although we are mainly interested in the study of sign-matrices, most of our notions and results deal with real-valued matrices because we do not want to impose unnecessary restrictions. A notable exception are the results in Section 5 which hold for sign-matrices only.

## 2. Definitions, Notations, and Facts

In this section, we provide the reader with the definitions and facts which will play a central role in the course of this paper.

**Vectors, Matrices, and Norms:** The all-ones vector in a finite-dimensional Euclidean space is simply denoted  $\mathbf{1}$ . The vector with value 1 in component  $k$  and zeros elsewhere is denoted  $e_k$ . The  $(d \times d)$  identity matrix is denoted by  $I_d$  or simply by  $I$ . Whenever the notation hides the dimension, say  $d$ , of the underlying Euclidean space, then  $d$  will be clear from context. The *Hadamard product* of two matrices  $A, B$  yields the matrix  $A \circ B = (a_{i,j} b_{i,j})$ , i.e., the matrices are multiplied entrywise. With  $\text{diag}(p)$  we denote the diagonal matrix build up from a vector  $p$  (i.e. the components of  $p$  are on the main diagonal and the remaining is zero). The *trace norm* of  $A \in \mathbb{R}^{m \times n}$ , denoted  $\|A\|_{tr}$ , is defined as the sum of all singular values of  $A$ . Let  $\|\cdot\|$  denote a vector norm. The notation  $\|A\|$  is understood as the norm of the  $mn$ -dimensional vector that results by concatenating the  $n$   $m$ -dimensional columns of  $A$  so as to form a single  $mn$ -dimensional vector. For example, the Euclidean norm applied to a matrix yields

$$\|A\|_2 = \sqrt{\sum_{i,j} A_{i,j}^2} ,$$

and this is sometimes called the *Frobenius norm* of  $A$ . The *operator norm* associated with  $\|\cdot\|$  is given by

$$\|A\| = \max_{\|v\|=1} \|Av\| = \max_{\|v\|\leq 1} \|Av\| .$$

For example, the operator norm associated with the Euclidean norm is given by

$$\|A\|_2 = \max_{\|v\|_2=1} \|Av\|_2 = \max_{\|v\|_2\leq 1} \|Av\|_2 ,$$

and this is sometimes called the *spectral norm* of  $A$ . We remind the reader to the following facts:

$$\begin{aligned} \|A\|_2 &= \|A^\top\|_2 , \quad \|AA^\top\|_2 = \|A^\top A\|_2 = \|A\|_2^2 , \quad \|A\|_2 \leq \|A\|_2 \\ \|A\|_2 &= \max_{\|u\|_2=\|v\|_2=1} u^\top Av = \max_{\|u\|_2,\|v\|_2\leq 1} u^\top Av \end{aligned} \tag{1}$$

By viewing matrices as vectors, we may consider the inner product of two matrices. An inner product without further specification refers to the standard scalar product. For example,  $\langle A, B \rangle = \sum_{i,j} A_{i,j} B_{i,j}$ . The dual of a norm  $\|\cdot\|$  is given by

$$\|u\|^* = \max_{\|v\|\leq 1} \langle u, v \rangle .$$

For example,  $L_\infty$  is the dual of  $L_1$ , and the trace norm is the dual of the spectral norm. It is well known that  $\|\cdot\|^{**} = \|\cdot\|$ , i.e., twofold dualization gives the original norm. Furthermore, for two norms  $\|\cdot\|_1, \|\cdot\|_2$  and every  $c > 0$ , we have

$$\|\cdot\|_1 \leq c \cdot \|\cdot\|_2 \Leftrightarrow \|\cdot\|_1^* \geq \frac{\|\cdot\|_2^*}{c} .$$

**SQ- and CSQ-Dimension:** Let  $p$  be a probability vector, i.e.,  $p$  has non-negative components that sum up to 1. Consider the inner product

$$\langle x, y \rangle_p := \sum_i p_i x_i y_i .$$

A collection of vectors  $u_1, \dots, u_d$  is said to be *almost p-orthogonal* if

$$\forall k \neq l \in \{1, \dots, d\} : |\langle u_k, u_l \rangle_p| \leq \frac{1}{d} .$$

The *SQ-dimension of a matrix*  $A \in \mathbb{R}^{m \times n}$  w.r.t.  $p$  is given by

$$\text{SQdim}_p(A) = \max\{d \in \{1, \dots, n\} : \text{there exist } d \text{ almost } p\text{-orthogonal column vectors in } A\} .$$

The *SQ-dimension of A* is given by

$$\text{SQdim}(A) = \max_p \text{SQdim}_p(A) .$$

A collection of (not necessarily different) vectors  $h_1, \dots, h_d \in [-1, 1]^m$  is said to be *universally correlated with A*  $\in \mathbb{R}^{m \times n}$  if, for every  $m$ -dimensional probability vector  $p$  and every  $j \in \{1, \dots, n\}$ , there exists  $k \in \{1, \dots, d\}$  such that  $|\langle h_k, A_j \rangle_p| \geq 1/d$ . The *CSQ-dimension* of  $A$  is given by

$$\text{CSQdim}(A) = \min\{d : \text{there exists a collection of } d \text{ vectors that is universally correlated with } A\} .$$

**Margin Complexity:** A *d-dimensional (homogeneous linear) arrangement for a matrix*  $A \in \mathbb{R}^{m \times n}$  is given by vectors

$$u_1, \dots, u_n; v_1, \dots, v_m \in \mathbb{R}^d$$

whose Euclidean norm is bounded by 1. With an arrangement  $\mathcal{A} = (u_1, \dots, u_n; v_1, \dots, v_m)$  for matrix  $A$ , we associate the *margin parameters*

$$\gamma_{i,j}(A|\mathcal{A}) = \langle u_i, v_j \rangle \cdot A_{i,j} .$$

The *margin complexity of A* is given by

$$\text{mc}(A) = \left( \max_{\mathcal{A}} \min_{i,j} \gamma_{i,j}(A|\mathcal{A}) \right)^{-1} .$$

It is easy to see that, for every matrix  $A$  without zero-entries (the matrices we are mainly interested in), there is always an arrangement that makes all margin parameters strictly positive, which implies that the margin complexity of  $A$  is strictly positive and finite. As outlined already by Linial et al. (2007) and Lee and Shraibman (2009), the so-called  $\gamma_2$ -norm and its dual,  $\gamma_2^*$ , are related to margin complexity as follows. Let  $r(M)$  denote the largest Euclidean norm of a row of the matrix  $M$ . With this notation  $\gamma_2$  and  $\gamma_2^*$ , satisfy the following equations (which, for the purpose of this paper, may also serve as a definition of these norms):

$$\gamma_2(A) = \min_{A=XY^\top} r(X)r(Y) \quad \text{and} \quad \gamma_2^*(A) = \max_{\mathcal{A}} \sum_{i,j} \gamma_{i,j}(A|\mathcal{A}) \tag{2}$$

Thus,  $\gamma_2^*(A)$  is basically the largest “total margin” that can be achieved by an arrangement for  $A$ .<sup>2</sup> Let  $Y = (y_{i,j})$  be an  $(m \times n)$ -dimensional matrix with non-negative entries that sum up to 1. The  $Y$ -average margin complexity of  $A$  is given by

$$\overline{\text{mc}}_Y(A) = (\gamma_2^*(Y \circ A))^{-1} .$$

In the special case where  $y_{i,j} = p_i q_j$  for two probability vectors  $p, q$  (so that  $Y = p \cdot q^\top$ ), we introduce the notations

$$\overline{\text{mc}}_{p,q}(A) := \overline{\text{mc}}_{pq^\top}(A) , \quad \overline{\text{mc}}_p(A) := \overline{\text{mc}}_{p,\mathbf{1}/n}(A) , \quad \overline{\text{mc}}(A) := \overline{\text{mc}}_{\mathbf{1}/m,\mathbf{1}/n}(A)$$

so that

$$\overline{\text{mc}}(A) = \left( \frac{1}{mn} \cdot \gamma_2^*(A) \right)^{-1} = \frac{mn}{\gamma_2^*(A)} . \quad (3)$$

Note that vector  $\mathbf{1}/n$  (resp.  $\mathbf{1}/m$ ) makes the columns (resp. rows) of  $A$  uniformly distributed. Considering the “smallest  $p$ -average margin” leads to the following definition:

$$\overline{\text{mc}}_{p,MIN}(A) = \left( \max_{\mathcal{A}} \min_j \sum_i p_i \gamma_{i,j}(A|\mathcal{A}) \right)^{-1} .$$

The various margin complexities are obviously related as follows:

$$\begin{aligned} \overline{\text{mc}}(A) &\leq \max_p \overline{\text{mc}}_p(A) \leq \max_{p,q} \overline{\text{mc}}_{p,q}(A) \leq \max_Y \overline{\text{mc}}_Y(A) \leq \text{mc}(A) \\ \max_q \overline{\text{mc}}_{p,q}(A) &\leq \overline{\text{mc}}_{p,MIN}(A) \end{aligned}$$

We will argue later that  $\max_Y \overline{\text{mc}}_Y(A) = \text{mc}(A)$  and  $\max_q \overline{\text{mc}}_{p,q}(A) = \overline{\text{mc}}_{p,MIN}(A)$ , but for the remaining inequalities the gap between the smaller and larger parameter can be exponentially large.

**Some Variants of the Forster bound:** It was shown by Forster and Simon (2006) that, for every  $A \in \mathbb{R}^{m \times n}$ ,

$$\overline{\text{mc}}(A) \geq \frac{\sqrt{mn}}{\|A\|_2} . \quad (4)$$

As for probability vectors  $p, q$ , we introduce the following notational convention:  $P$  and  $Q$  are defined as the diagonal matrices containing the components of  $p$  and  $q$ , respectively. That is  $P := \text{diag}(p)$  and  $Q := \text{diag}(q)$ . But keep in mind that this convention is *not* applied to letters different from  $P$  and  $Q$ . Let  $A$  be a real-valued matrix with  $m$  rows and  $n$  columns. Consider the following variant of the Forster bound:

$$\text{FB}_{p,q}(A) = \frac{1}{\|P^{1/2}AQ^{1/2}\|_2}$$

For  $q = \mathbf{1}/n$ , we simply write  $\text{FB}_p(A)$  instead of  $\text{FB}_{p,\mathbf{1}/n}$ . For this choice of  $q$ ,  $Q = \frac{1}{n}I_n$  and we obtain

$$\text{FB}_p(A) = \frac{\sqrt{n}}{\|P^{1/2}A\|_2} .$$

---

2. Note that the arrangement that maximizes the total margin may have some negative individual margin parameters.

Similarly, if  $p = 1/m$ , we simply write  $\text{FB}(A)$  instead of  $\text{FB}_p(A)$ . For this choice of  $p$ ,  $P = \frac{1}{m}I_m$  and we obtain

$$\text{FB}(A) = \frac{\sqrt{mn}}{\|A\|_2},$$

which is the “classical” Forster bound in (4). Here, and in what follows, we use the notation  $A'$  to indicate a sub-matrix of  $A$  that is formed by a subset of the columns of  $A$ . Let  $n(A') \leq n$  denote the number of columns in  $A'$ . Note that

$$\max_{A'} \text{FB}_p(A') \leq \max_q \text{FB}_{p,q}(A)$$

because  $\text{FB}_{p,q}$  collapses to  $\text{FB}_p(A')$  when the components of  $q$  are either 0 or  $1/n(A')$ , and the non-zero components are in one-to-one correspondence to the columns of  $A$  that are used to build  $A'$ .

**Semidefinite Programming (SDP):** We write  $A \succeq B$  iff  $A - B$  is a symmetric positive semi-definite matrix. The following definitions and facts about semi-definite programming are taken from Alizadeh (1995). A *standard primal SDP* is an optimization problem of the following form:

$$\min_X \langle C, X \rangle \quad \text{s.t. } \forall \rho = 1, \dots, r : \langle A_\rho, X \rangle = b_\rho, X \succeq 0 \quad (5)$$

Here, the matrices  $C, A_i$  are assumed to be symmetric. As in Linear Programming there is a duality theory for SDPs. The variables for the dual are denoted  $y_1, \dots, y_r$  (one dual variable per equality-constraint in the primal). We say that the equality-constraints of the primal *induce* the matrix  $\sum_{\rho=1}^r y_\rho A_\rho$ . The dual of (5) looks as follows:

$$\max_y \langle b, y \rangle \quad \text{s.t. } C - \sum_{\rho=1}^r y_\rho A_\rho \succeq 0$$

If the optimal values of the primal and dual are equal, we achieve “strong duality”. Among the well-known sufficient conditions for strong duality is the following-one (where “SCQ” means “Slater’s Constraint Qualification”).

**SCQ:** There exists  $y$  such that  $\sum_{\rho=1}^r y_\rho A_\rho$  is (strictly) positive definite.

Note that non-negativity constraints for individual variables  $w_i$  can be expressed within a constraint of the form  $X \succeq 0$  because the matrix  $X$  could be of the form  $\begin{bmatrix} X' & 0 \\ 0 & \text{diag}(w_1, \dots, w_s) \end{bmatrix}$ . We may therefore liberalize our definition of a standard primal SDP and allow constraints of the form  $w_i \geq 0$ .

### 3. Margin Maximization and its Dual

The fact that the optimal margin can be computed in polynomial time using semi-definite programming (SDP) had been observed first by Linial et al. (2007). In this section, we make use of this observation and express several variants of margin optimization as instances of

SDP. Throughout this section,  $A, M$  are  $(m \times n)$ -matrices,  $X$  is an  $(m+n) \times (m+n)$ -matrix containing variables of the primal SDP, index  $i$  ranges from 1 to  $m$ , index  $j$  ranges from 1 to  $n$ , and index  $k$  ranges from 1 to  $m+n$ .

We call into mind the fact that a semi-definite matrix  $X$  can be written in the form  $X = W^\top \cdot W$  (e.g., Cholesky-decomposition). If  $X$  has  $m+n$  rows and columns, respectively, then  $W$  has, say,  $d$  rows and  $m+n$  columns. Let  $W = [U \ V]$  be the decomposition of  $W$  with  $U$  containing the first  $m$  columns. Then,

$$W^\top \cdot W = [U \ V]^\top \cdot [U, V] = \begin{bmatrix} U^\top U & U^\top V \\ (U^\top V)^\top & V^\top V \end{bmatrix}.$$

Imposing constraints like  $X_{i,i} = \langle U_i, U_i \rangle = 1$ ,  $X_{m+j,m+j} = \langle V_j, V_j \rangle = 1$ , we can view  $X$  as a representation of an arrangement  $\mathcal{A}$  given by the columns of  $U$  and the columns of  $V$ . Note that  $\gamma_{i,j}(A|\mathcal{A}) = A_{i,j} \langle U_i, V_j \rangle = A_{i,j} X_{i,m+j}$ . This is why many variants of margin-maximization problems can be expressed as instances of SDP. The following results are applications of strong SDP-duality.

**Theorem 1** *For every  $A \in \mathbb{R}^{m \times n}$ :  $\text{mc}(A) = \max_Y \overline{\text{mc}}_Y(A)$ .*

**Proof** We will prove the equivalent statement

$$\max_{\mathcal{A}} \min_{i,j} \gamma_{i,j}(A|\mathcal{A}) = \min_Y \gamma_2^*(Y \circ A) \stackrel{(2)}{=} \min_Y \max_{\mathcal{A}} \sum_{i,j} \gamma_{i,j}(Y \circ A|\mathcal{A}). \quad (6)$$

Finding an arrangement  $\mathcal{A}$  for  $M = Y \circ A$  that maximizes  $\sum_{i,j} \gamma_{i,j}(M|\mathcal{A})$  can be expressed as a standard SDP-problem (with optimal value  $\gamma_2^*(Y \circ A)$ ) as follows:

$$\min_X -\frac{1}{2} \cdot \sum_{i,j} M_{i,j} (X_{i,m+j} + X_{m+j,i}) \quad \text{s.t. } \forall k : X_{k,k} = 1 \text{ and } X \succeq 0 \quad (7)$$

There are  $m+n$  equality constraints, which leads to dual variables  $y_1, \dots, y_{m+n}$ . The matrix induced by the equality-constraints equals  $\text{diag}(y_1, \dots, y_{m+n})$ . Obviously, condition SCQ is satisfied so that we have strong duality. The cost matrix of the primal is given by

$$C = \frac{1}{2} \cdot \begin{bmatrix} 0 & -M \\ -M^\top & 0 \end{bmatrix}$$

Thus, the dual problem (with variables  $-y_k/2$  substituted for  $y_k$  and  $Y \circ A$  substituted for  $M$ ) looks as follows:

$$\min_y \frac{1}{2} \cdot \sum_k y_k \quad \text{s.t. } \underbrace{\begin{bmatrix} \text{diag}(y_1, \dots, y_m) & -(Y \circ A) \\ -(Y \circ A)^\top & \text{diag}(y_{m+1}, \dots, y_{m+n}) \end{bmatrix}}_{=:S} \succeq 0 \quad (8)$$

Finding an arrangement  $\mathcal{A}$  for  $A$  that maximizes  $\min_{i,j} \gamma_{i,j}(A|\mathcal{A})$  can be expressed as a standard SDP-problem (with slack variables  $s_{i,j}$ ) as follows:

$$\min_{X,\mu,s} -\mu \quad \text{s.t. } \forall k : X_{k,k} = 1, \forall i,j : A_{i,j} (X_{i,m+j} + X_{m+j,i}) - s_{i,j} = 2\mu, X \succeq 0, s_{i,j} \geq 0, \mu \geq 0$$

The  $(m + n + mn + 1) \times (m + n + mn + 1)$ -matrix of primal variables is then given by

$$\begin{bmatrix} X & 0 & 0 \\ 0 & \text{diag}(s_{1,1}, \dots, s_{m,n}) & 0 \\ 0 & 0 & \mu \end{bmatrix}.$$

The dual variables are denoted  $y_k$  and  $y_{i,j}$ . Setting  $Y = (y_{i,j})$ , the matrix induced by the equality-constraints equals

$$\begin{bmatrix} \text{diag}(y_1, \dots, y_m) & Y \circ A & 0 & 0 \\ (Y \circ A)^\top & \text{diag}(y_{m+1}, \dots, y_{m+n}) & 0 & 0 \\ 0 & 0 & -\text{diag}(y_{1,1}, \dots, y_{m,n}) & 0 \\ 0 & 0 & 0 & -2 \sum_{i,j} y_{i,j} \end{bmatrix}.$$

It is easy to see that condition SCQ is satisfied: one may assign value  $-1$  to every variable  $y_{i,j}$  and a sufficiently large value to every variable  $y_k$  so that all eigenvalues must be strictly positive according to the Geršgorin Disc Theorem. Thus, we have strong duality. The dual problem (with variables  $-y_k/2$  substituted for  $y_k$  and  $y_{i,j}/2$  substituted for  $y_{i,j}$ ) looks as follows:<sup>3</sup>

$$\min_{Y,y} \frac{1}{2} \sum_k y_k \quad \text{s.t.} \quad \sum_{i,j} y_{i,j} = 1, \quad y_{i,j} \geq 0, \quad \begin{bmatrix} \text{diag}(y_1, \dots, y_m) & -(Y \circ A) \\ -(Y \circ A)^\top & \text{diag}(y_{m+1}, \dots, y_{m+n}) \end{bmatrix} \succeq 0 \quad (9)$$

By strong duality, (9) equals  $\max_{\mathcal{A}} \min_{i,j} \gamma_{i,j}(A|\mathcal{A})$ , and (8) equals  $\gamma_2^*(Y \circ A)$ . A comparison of (9) with (8) shows that (6) holds. ■

One can show that any arrangement of arbitrary dimension can be transformed (by virtue of Cholesky decomposition) into another arrangement of dimension at most  $m + n + mn + 1$  that achieves the same values for the respective margin parameters. Combined with a straightforward compactness and continuity argument this shows that there exists a maximizer  $\mathcal{A}^*$  for  $\min_{i,j} \gamma_{i,j}(A|\mathcal{A})$ , and there exists a minimizer  $Y^*$  for  $\gamma_2^*(Y \circ A)$ . According to (6), both problems have the same optimal value, say  $\gamma^*$ , i.e.,

$$\gamma^* := \min_{i,j} \gamma_{i,j}(A|\mathcal{A}^*) = \gamma_2^*(Y^* \circ A).$$

The following set  $K(A)$  represents the “hard part” of the matrix  $A \in \mathbb{R}^{m \times n}$  (thereby playing a similar role as support vectors in SVM optimization problems):

$$K(A|\mathcal{A}^*) := \{(i,j) : \gamma_{i,j}(A|\mathcal{A}) = \gamma^*\} \quad \text{and} \quad K(A) := \bigcap_{\mathcal{A}^*} K(A|\mathcal{A}^*)$$

In the definition of  $K(A)$ ,  $\mathcal{A}^*$  ranges over all maximizers for  $\min_{i,j} \gamma_{i,j}(A|\mathcal{A})$ . We say that  $Y$  is *centered* on  $K \subseteq \{1, \dots, m\} \times \{1, \dots, n\}$  if  $y_{i,j} = 0$  for all  $(i,j) \notin K$ . With these notations, the following holds:

---

3. The constraint  $C - \sum_\rho y_\rho A_\rho \succeq 0$  forces the  $y_{i,j}$  to be non-negative and to satisfy  $\sum_{i,j} y_{i,j} - 1 \geq 0$ , but it is obvious that, for an optimal assignment to the variables  $y_{i,j}$ , their values will sum up to 1 exactly.

**Corollary 2** 1. Every minimizer  $Y^*$  for  $\gamma_2^*(Y \circ A)$  is centered on  $K(A)$ .

$$2. \max_{\mathcal{A}} \min_{i,j} \gamma_{i,j}(A|\mathcal{A}) = \max_{\mathcal{A}} \min_{(i,j) \in K(A)} \gamma_{i,j}(A|\mathcal{A}).$$

### Proof

1. The claim is proved indirectly. Consider a matrix  $Y$  such that, for some  $(i', j') \notin K(A)$ ,  $y_{i',j'} > 0$ . According to the definition of  $K(A)$ , there must exist a maximizer  $\mathcal{A}^*$  for  $\min_{i,j} \gamma_{i,j}(A|\mathcal{A})$  such that  $(i', j') \notin K(A|\mathcal{A}^*)$ . This implies that  $\mathcal{A}^*$  achieves a  $Y$ -average margin strictly greater than  $\gamma^*$ . Thus,  $Y$  is not a minimizer for  $\gamma_2^*(Y \circ A)$ .
2. Let  $Y'$  range over all  $Y$  that are centered on  $K(A)$ . A straightforward modification of the proof of (6) shows that

$$\max_{\mathcal{A}} \min_{(i,j) \in K(A)} \gamma_{i,j}(A|\mathcal{A}) = \min_{Y'} \gamma_2^*(Y' \circ A) \quad (10)$$

Thus it suffices to show that

$$\min_Y \gamma_2^*(Y \circ A) = \min_{Y'} \gamma_2^*(Y' \circ A) .$$

But this is evident from the first part of the corollary.

■

The proof of the following result is similar to the proof of Theorem 1. It is found in Section A.

**Theorem 3** For every  $A \in \mathbb{R}^{m \times n}$  and every probability vector  $p$ :  $\overline{\text{mc}}_{p,\text{MIN}}(A) = \max_q \overline{\text{mc}}_{p,q}(A)$ .

### 3.1. Bounds on Average Margin Complexity

We make use of the inequalities

$$\|M\|_{tr} \leq \gamma_2^*(M) \leq \sqrt{mn} \cdot \|M\|_2 . \quad (11)$$

Because of (3), the second inequality is equivalent to (4), and it can also be found in (Linial et al., 2007). The first inequality is probably known as well but, since we are not aware of a proper reference, we will now provide the reader with a short proof for sake of completeness. Since the spectral norm is the dual of the trace norm, it suffices to show that

$$\gamma_2(M) \leq \|M\|_2 .$$

We denote the rank of  $M$  by  $r$ , and we make use of the singular value decomposition

$$M = U \cdot \text{diag}(\sigma_1, \dots, \sigma_r) \cdot V^\top = \underbrace{(U \cdot \text{diag}(\sqrt{\sigma_1}, \dots, \sqrt{\sigma_r}))}_{=:X} \cdot \underbrace{(V \cdot \text{diag}(\sqrt{\sigma_1}, \dots, \sqrt{\sigma_r}))^\top}_{=:Y} . \quad (12)$$

Here,  $U$  is an  $(m \times r)$ -matrix whose columns  $U_1, \dots, U_r$  have unit norm and are pairwise orthogonal. Likewise,  $V$  is an  $(n \times r)$ -matrix whose columns  $V_1, \dots, V_r$  have unit norm and

are pairwise orthogonal.  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$  are the singular values of  $M$ . As a matter of fact,  $\|M\|_2 = \sigma_1$ . Now obviously

$$\gamma_2(M) \stackrel{(2)}{\leq} r(X) \cdot r(Y) \leq \sqrt{\sigma_1} \cdot \sqrt{\sigma_1} = \|M\|_2$$

which concludes the verification of (11). Because of (3), (11) is equivalent to

$$\frac{\sqrt{mn}}{\|M\|_2} \leq \overline{\text{mc}}(M) \leq \frac{mn}{\|M\|_{tr}} . \quad (13)$$

### 3.2. Exact Determination of Average Margin Complexity

In general, the bounds (11) and (13) leave a gap. In this section, we consider families of matrices whose average margin complexity can be determined exactly: Hadamard matrices and matrices composed of all reflections of a single Boolean function.

By definition, a *Hadamard matrix*  $H$  of order  $n$  is a sign-matrix that satisfies  $H \cdot H^\top = n \cdot I$ .

**Corollary 4** *Let  $H$  be a Hadamard matrix of order  $n$ . Then,  $\overline{\text{mc}}(H) = \sqrt{n}$ .*

**Proof** A Hadamard matrix  $H$  of order  $n$  satisfies  $\sigma_1(H) = \dots = \sigma_n(H)$ . Thus,  $\|H\|_{tr} = n \cdot \|H\|_2$ , which makes the upper bound in (13) collapse to the lower bound in (13). ■

**Lemma 5** *If  $M \in \mathbb{R}^{n \times n}$  and the matrices  $U, V$  in its singular value decomposition (12) have only entries from  $\{\pm 1/\sqrt{n}\}$ , then  $\gamma_2^*(M) = n \cdot \|M\|_2$ .*

**Proof** According to (11),  $\gamma_2^*(M) \leq n \cdot \|M\|_2$ . By duality of norms, the converse direction is equivalent to  $\gamma_2(M) \leq \|M\|_{tr}/n$ . An inspection of (12) shows that, given our assumptions on the entries of  $U$  and  $V$ ,  $r(X) = r(Y) = \sqrt{\|M\|_{tr}/n}$ . Thus,  $\gamma_2(M) \leq r(X)r(Y) = \|M\|_{tr}/n$ , as required. ■

We briefly note that the assumptions of Lemma 5 can be weakened: it suffices to assume that the first columns of  $U$  and  $V$ , respectively, have entries from  $\{\pm 1/\sqrt{n}\}$ . The proof will then make use of strong SDP duality.

**Corollary 6** *Let  $f : \{-1, 1\}^d \rightarrow \{-1, 1\}$  be a Boolean function, and let  $L_\infty(f)$  denote the largest Fourier-coefficient in terms of absolute value. Consider the  $(2^d \times 2^d)$ -matrix  $F_{x,y} := f(x \circ y)$  where  $x \circ y := (x_1y_1, \dots, x_dy_d)$ . Then:  $\overline{\text{mc}}(F) = 1/L_\infty(f)$ .*

**Proof** Let  $\hat{F}$  be the matrix with the Fourier-coefficients of  $f$  on its main diagonal and zeros elsewhere. Let  $H$  denote the Silvester-type Hadamard matrix of order  $2^d$ . It is well-known (e.g., Doliwa et al., 2008) that the spectral decomposition of  $2^{-d}F$  has the form

$$2^{-d}F = 2^{-d/2}H \cdot \hat{F} \cdot 2^{-d/2}H .$$

This implies that  $\|F\|_2 = 2^d \cdot L_\infty(f)$  and that  $M = 2^{-d}F$  satisfies the assumptions of Lemma 5 so that  $\gamma_2^*(F) = 2^d \|F\|_2 = 2^{2d} L_\infty(f)$ . Thus, according to (3) with  $n = m = 2^d$ ,  $\overline{\text{mc}}(F) = 1/L_\infty(f)$ .  $\blacksquare$

#### 4. The Replication Trick

We have learned the replication trick from Sherstov (2008). He used it (together with the classical Forster bound on dimension complexity) to show that, for every sign-matrix  $A$ ,  $\text{SQdim}(A) = \max_p \text{SQdim}_p(A)$  is bounded from above by twice the square of the dimension complexity of  $A$ . Here, we will use the trick (together with (4)) for showing that, for every real-valued matrix  $A$ ,  $\overline{\text{mc}}_{p,q}(A) \geq \text{FB}_{p,q}(A)$ .

**Lemma 7** *Let  $A \in \mathbb{R}^{m \times n}$ . Let  $p$  be an  $m$ -dimensional probability vector with rational components  $r_i/R$  (so that  $\sum_{i=1}^m r_i = R$ ). Similarly, let  $q$  be an  $n$ -dimensional probability vector with rational components  $s_j/S$  (so that  $\sum_{j=1}^n s_j = S$ ). Let  $A_s$  be the matrix that results from  $A$  by duplicating the  $j$ -th column  $s_j$ -times. Let  $A_{r,s}$  denote the matrix that results from  $A_s$  by duplicating the  $i$ -th row  $r_i$ -times. With this notation, the following holds:*

$$\overline{\text{mc}}_{p,q}(A) = \overline{\text{mc}}(A_{r,s}) \text{ and } \sqrt{RS} \cdot \|P^{1/2}AQ^{1/2}\|_2 = \|A_{r,s}\|_2 \quad (14)$$

**Proof** We first show that  $\overline{\text{mc}}_{p,q}(A) \leq \overline{\text{mc}}(A_{r,s})$ . Any arrangement  $\mathcal{A} = (u_1, \dots, u_m; v_1, \dots, v_n)$  for  $A$  induces an arrangement  $\mathcal{A}'$  for  $A_{r,s}$  where the  $k$ -th duplicate of row  $i$  (resp. column  $j$ ) is represented by the (same) vector  $u_i$  (resp.  $v_j$ ). The average margin achieved by  $\mathcal{A}'$  equals

$$\frac{1}{RS} \sum_i \sum_j r_i s_j \langle u_i, v_j \rangle A_{i,j} = \sum_i \sum_j \frac{r_i s_j}{R S} \langle u_i, v_j \rangle A_{i,j} .$$

But the right hand-side equals the  $(p, q)$ -average margin achieved by  $\mathcal{A}$ .

Now, we show that  $\overline{\text{mc}}(A_{r,s}) \leq \overline{\text{mc}}_{p,q}(A)$ . To this end, we start with an arrangement  $\mathcal{A}'$  for  $A_{r,s}$  where the  $k$ -th duplicate of row  $i$  (resp. column  $j$ ) is represented by  $u_i(k)$  (resp.  $v_j(k)$ ). The average margin achieved by  $\mathcal{A}'$  equals

$$\begin{aligned} \frac{1}{RS} \sum_i \sum_j \sum_{k_i=1}^{r_i} \sum_{l_j=1}^{s_j} \langle u_i(k_i), v_j(l_j) \rangle A_{i,j} &= \frac{1}{RS} \sum_i \sum_j \left\langle \sum_{k_i=1}^{r_i} u_i(k_i), \sum_{l_j=1}^{s_j} v_j(l_j) \right\rangle A_{i,j} \\ &= \sum_i \sum_j \frac{r_i s_j}{R S} \left\langle \frac{1}{r_i} \sum_{k_i=1}^{r_i} u_i(k_i), \frac{1}{s_j} \sum_{l_j=1}^{s_j} v_j(l_j) \right\rangle A_{i,j} . \end{aligned}$$

But the final term coincides with the  $(p, q)$ -average margin that is achieved for  $A$  by the vectors

$$u_i = \frac{1}{r_i} \sum_{k_i=1}^{r_i} u_i(k_i) \text{ and } v_j = \frac{1}{s_j} \sum_{l_j=1}^{s_j} v_j(l_j) .$$

Note that, by the triangle inequality,  $\|u_i\|_2$  is bounded by 1 provided that  $\|u_i(1)\|_2, \dots, \|u_i(r_i)\|_2$  are bounded by 1. The analogous argument applies to  $v_j$ .

As for the second equation in (14), it suffices to show that  $\|\sqrt{S}AQ^{1/2}\|_2 = \|A_s\|_2$ . (We can then apply this equality with  $P^{1/2}A$  substituted for  $A$  and proceed with a symmetry argument.) Our proof for  $\|\sqrt{S}AQ^{1/2}\|_2 = \|A_s\|_2$  will make use of (1). We first show that  $\|\sqrt{S}AQ^{1/2}\|_2 \leq \|A_s\|_2$ . Note that the entry  $(i, j)$  of matrix  $\sqrt{S}AQ^{1/2}$  coincides with  $\sqrt{s_j}A_{i,j}$ . With any  $n$ -dimensional vector  $v$ , we associate the  $S$ -dimensional vector  $v'$  which is composed of sub-vectors  $v'(1), \dots, v'(n)$  of dimensions  $s_1, \dots, s_n$ , respectively, such that  $v'(j) = \frac{v_j}{\sqrt{s_j}} \cdot \mathbf{1}$ . Note that  $\|v'\|_2 = \|v\|_2$ . Furthermore note that

$$u^\top A_s v' = \sum_i \sum_j s_j u_i \frac{v_j}{\sqrt{s_j}} A_{i,j} = \sum_i \sum_j u_i v_j (\sqrt{s_j} A_{i,j}) = u^\top (\sqrt{S}AQ^{1/2})v .$$

Now, we show that  $\|A_s\|_2 \leq \|\sqrt{S}AQ^{1/2}\|_2$ . To this end, we consider an  $m$ -dimensional vector  $u$  and an  $S$ -dimensional vector  $v'$ . We can think of  $v'$  as being composed of  $s_j$ -dimensional sub-vectors  $v'(j)$  for  $j = 1, \dots, n$ . Then,

$$u^\top A_s v' = \sum_i \sum_j \sum_{k_j=1}^{s_j} u_i v'(j)_{k_j} A_{i,j} = \sum_j \left( \sum_i u_i A_{i,j} \right) \left( \sum_{k_j=1}^{s_j} v'(j)_{k_j} \right) .$$

Setting  $v_j := \frac{1}{\sqrt{s_j}} \sum_{k_j=1}^{s_j} v'(j)_{k_j}$ , the latter term equals

$$\sum_i \sum_j u_i v_j (\sqrt{s_j} A_{i,j}) = u^\top (\sqrt{S}AQ^{1/2})v .$$

Note that

$$v_j^2 = \frac{1}{s_j} \cdot \langle v'(j), \mathbf{1} \rangle^2 \leq \frac{1}{s_j} \cdot \langle v'(j), v'(j) \rangle \cdot \langle \mathbf{1}, \mathbf{1} \rangle = \langle v'(j), v'(j) \rangle ,$$

which implies that  $\|v\|_2 \leq \|v'\|_2$ . ■

**Corollary 8** For all probability vectors  $p, q$ :  $\overline{\text{mc}}_{p,q}(A) \geq \text{FB}_{p,q}(A)$ .

**Proof** With the notation from Lemma 7, the following holds for all rational probability vectors  $p, q$ :

$$\overline{\text{mc}}_{p,q}(A) = \overline{\text{mc}}(A_{r,s}) \stackrel{(4)}{\geq} \text{FB}(A_{r,s}) = \text{FB}_{p,q}(A) .$$

In order to generalize this equality to arbitrary probability vectors (with possibly non-rational components), we can use that fact that  $\mathbb{Q}$  is dense in  $\mathbb{R}$  and apply an obvious continuity argument. ■

## 5. SQ-Dimension and Margin-Complexity

In this section we focus on sign matrices. We establish two more relations (see Lemmas 9 and 10), and then we put all pieces together and arrive at the inequalities in (15) and (16). As a by-product, we obtain two results, see (17) and (18), which might be of independent interest.

**Lemma 9** *For every  $A \in \{-1, 1\}^{m \times n}$ :  $\overline{\text{mc}}_{p, \text{MIN}}(A) < \sqrt{\text{SQdim}_p(A)} \cdot (\text{SQdim}_p(A) + 1)$ .*

**Proof** Let  $d := \text{SQdim}_p(A)$ . Select a subset  $S = \{s(1), \dots, s(d)\} \subseteq \{1, \dots, n\}$  such that

$$\left( \forall k \neq l \in S : |\langle A_k, A_l \rangle_p| \leq \frac{1}{d} \right) \wedge \left( \forall j \in \{1, \dots, n\}, \exists k(j) \in \{1, \dots, d\} : |\langle A_j, A_{s(k(j))} \rangle_p| > \frac{1}{d+1} \right) .$$

Let  $\sigma_j := \text{sign}(\langle A_j, A_{s(k(j))} \rangle_p)$ . We define a  $d$ -dimensional arrangement for  $A$  as follows:

$$\left( \forall i = 1, \dots, m : u_i = \frac{1}{\sqrt{d}} \cdot (A_{i,s(1)}, \dots, A_{i,s(d)}) \right) \wedge \left( \forall j = 1, \dots, n : v_j = \sigma_j \cdot e_{k(j)} \right)$$

It follows that  $\langle u_i, v_j \rangle = \sigma_j \cdot A_{i,s(k(j))} / \sqrt{d}$ , and our embedding exhibits the margin parameters

$$\gamma_{i,j} = \langle u_i, v_j \rangle \cdot A_{i,j} = \frac{\sigma_j \cdot A_{i,s(k(j))} \cdot A_{i,j}}{\sqrt{d}} .$$

Averaging w.r.t. to  $p$  yields

$$\sum_i p_i \gamma_{i,j} = \frac{1}{\sqrt{d}} \cdot |\langle A_j, A_{s(k(j))} \rangle_p| > \frac{1}{\sqrt{d} \cdot (d+1)} .$$

Since this holds for every choice  $j$ , we get  $\overline{\text{mc}}_{p, \text{MIN}}(A) < \sqrt{d} \cdot (d+1)$ , as desired.  $\blacksquare$

The proof of the following result builds on a proof by Sherstov (2008) for a quite similar result:

**Lemma 10** *For every  $A \in \{-1, 1\}^{m \times n}$ :  $\text{SQdim}_p(A) < 2 \cdot \max_{A'} \text{FB}_p(A')^2$ .*

**Proof** Let  $d = \text{SQdim}_p(A)$ , and let  $S \subseteq \{1, \dots, n\}$  be chosen as in the proof of Lemma 9. Let  $A'$  be the submatrix that is formed by the columns  $s(1), \dots, s(d)$  of  $A$ . It follows that

$$C := A'^\top P A' = (P^{1/2} A')^\top (P^{1/2} A') \in \mathbb{R}^{d \times d}$$

has ones on the main diagonal and entries of absolute value at most  $1/d$  elsewhere. We apply an argument of Sherstov (2008) and conclude that

$$\|C\|_2 \leq \|C - I_d\|_2 + \|I_d\|_2 \leq \|C - I_d\|_2 + \|I_d\|_2 = \sqrt{\frac{d(d-1)}{d^2}} + 1 < 2 .$$

Note that

$$\|C\|_2 = \|P^{1/2} A'\|_2^2 .$$

The proof is now accomplished as follows:

$$\text{FB}_p(A')^2 \geq \frac{d}{\|P^{1/2}A'\|_2^2} = \frac{d}{\|C\|_2} > \frac{d}{2}$$

■

The combination of Lemma 9, Lemma 10, and Corollary 8 demonstrates that the parameters  $\text{SQdim}_p(A)$ ,  $\max_{A'} \text{FB}_p(A')$ ,  $\max_q \text{FB}_{p,q}(A)$ , and  $\max_q \overline{\text{mc}}_{p,q}(A) = \overline{\text{mc}}_{p,\text{MIN}}(A)$  are related as follows:

$$\text{SQdim}_p(A) < 2 \max_{A'} \text{FB}_p(A')^2 \leq 2 \max_q \text{FB}_{p,q}(A)^2 \leq 2 \max_q \overline{\text{mc}}_{p,q}(A)^2 < 2 \text{SQdim}_p(A)(\text{SQdim}_p(A)+1)^2 \quad (15)$$

Applying the operation “ $\max_p$ ” to (15), we get

$$\text{SQdim}(A) < 2 \cdot \max_{p,q} \text{FB}_{p,q}(A)^2 \leq 2 \cdot \max_{p,q} \overline{\text{mc}}_{p,q}(A)^2 < 2 \cdot \text{SQdim}(A) \cdot (\text{SQdim}(A)+1)^2. \quad (16)$$

Since  $\max_{p,q} \overline{\text{mc}}_{p,q}(A) = \max_{p,q} \overline{\text{mc}}_{p,q}(A^\top)$  — an analogous remark is valid for  $\max_{p,q} \text{FB}_{p,q}$  — it follows from (16) that

$$\text{SQdim}(A^\top) < 2 \cdot \text{SQdim}(A) \cdot (\text{SQdim}(A)+1)^2. \quad (17)$$

This improves on a result by Sherstov (2008): he used a polynomial relation between  $\text{SQdim}(A)$  and the discrepancy of  $A$  with respect to product distributions for showing that  $\text{SQdim}(A^\top) \leq 32 \cdot \text{SQdim}(A)^4$ .

Recall that, by convention,  $A'$  ranges over all sub-matrices of  $A$  which can be composed by (complete) columns of  $A$ . Let  $A''$  range over all sub-matrices of  $A$ . We claim that

$$\max_{A''} \text{FB}(A'') \leq \max_{p,q} \text{FB}_{p,q}(A) < 64 \cdot (1 + o(1)) \cdot \max_{A''} \text{FB}(A'')^9. \quad (18)$$

The first inequality is obvious. The last inequality is obtained by applying (15) twice, the first time on  $A$  and the second time on the transpose of  $A'$ .

## 6. CSQ-Dimension and Margin Complexity

Feldman (2008) has shown the following result. If a concept class  $\mathcal{C}$  over domain  $\mathcal{X}$  has CSQ-dimension  $d$ , then there exists a family  $W$  consisting of  $d$  Boolean base functions such every function in  $\mathcal{C}$  can be written as the majority of  $O(\log(|\mathcal{X}|)d^2)$  functions properly chosen from  $W$ . Viewing a sign-matrix  $A$  as a concept class, it is not hard to infer from that result an upper bound on  $\text{mc}(A)$  in terms of  $\text{CSQdim}(A)$ . However, a direct derivation of such an upper bound (as in the proof of the following lemma) leads to a tighter relationship:

**Lemma 11** *For every  $A \in \mathbb{R}^{m \times n}$ :  $\text{mc}(A) \leq \text{CSQdim}(A)^{1.5}$ .*

**Proof** Let  $d := \text{CSQdim}(A)$ , and let  $h_1, \dots, h_d \in \mathbb{R}^m$  be universally correlated with  $A$ . According to Lemma 1, there exists a matrix  $Y = (y_{i,j})$  such that  $\text{mc}(A) = \overline{\text{mc}}_Y(A)$ . We will present a  $d$ -dimensional arrangement  $u_1, \dots, u_m; v_1, \dots, v_n$  whose  $Y$ -average margin equals  $1/d^{1.5}$  (which proves the lemma). To this end, we set

$$u_i := \frac{1}{\sqrt{d}} \cdot (h_{i,1}, \dots, h_{i,d})$$

where  $h_{i,k}$  denotes the  $i$ -th component of vector  $h_k$ . Note that  $\|u_i\|_2 \leq 1$ . Furthermore, let  $Y_j := y_{1,j} + \dots + y_{m,j}$ , and let the  $m$ -dimensional probability vector  $p_j$  be given by  $(p_j)_i := y_{i,j}/Y_j$ . Because  $h_1, \dots, h_d$  is universally correlated with  $A$ , the following holds. For every  $j$ , there exists  $k(j) \in \{1, \dots, d\}$  such that  $|\langle h_{k(j)}, A_j \rangle_{p_j}| \geq 1/d$ . Now we set  $\sigma_j := \text{sign}(\langle h_{k(j)}, A_j \rangle_{p_j})$ ,  $v_j := \sigma_j \cdot e_{k(j)}$ , and we bound the  $Y$ -average margin from below as follows:

$$\sum_i \sum_j y_{i,j} \langle u_i, v_j \rangle A_{i,j} = \frac{1}{\sqrt{d}} \cdot \sum_j Y_j \sigma_j \sum_i \frac{y_{i,j}}{Y_j} h_{i,k(j)} A_{i,j} = \frac{1}{\sqrt{d}} \cdot \sum_j Y_j |\langle h_{k(j)}, A_j \rangle_{p_j}| \geq \frac{1}{d^{1.5}}$$

■

As for the converse direction, we get the following result:

**Lemma 12** *For every  $A \in \mathbb{R}^{m \times n}$  and  $\ell(m, n) := 32 \ln(4mn)$ :*

$$\text{CSQdim}(A) \leq \lceil \ell(m, n) \cdot \text{mc}(A)^2 \rceil$$

**Proof** Consider an arrangement  $\mathcal{A}$  that maximizes  $\gamma := \min_{i,j} \gamma_{i,j}(A|\mathcal{A})$  so that  $\text{mc}(A) = 1/\gamma$ . It is well-known<sup>4</sup> that  $\mathcal{A}$  can be transformed into another arrangement  $\mathcal{A}' = (u_1, \dots, u_m; v_1, \dots, v_n)$  that is  $d$ -dimensional for  $d := \lceil \ell(m, n)/\gamma^2 \rceil$  and still satisfies  $\min_{i,j} \gamma_{i,j}(A|\mathcal{A}') \geq \gamma/2$ . For every  $k \in \{1, \dots, d\}$ , let  $u_{i,k}$  denote the  $i$ -th component of  $u_k$ . We will show that  $h_1, \dots, h_d$  given by

$$h_k = (u_{1,k}, \dots, u_{m,k})$$

is universally correlated with  $A$ . To this end, let  $p$  be an arbitrary but fixed  $m$ -dimensional probability vector, and let  $v'_j = \|v_j\|_1^{-1} \cdot v_j$  so that

$$\|v'_j\|_1 = \sum_{k=1}^d |v'_{j,k}| = 1 . \quad (19)$$

Note that  $\|v_j\|_1 \leq \sqrt{d}$  since  $\mathcal{A}'$  is a  $d$ -dimensional arrangement. It follows that

$$\min_{i,j} \langle u_i, v'_j \rangle A_{i,j} \geq \frac{\gamma}{2\sqrt{d}} ,$$

---

4. This is a typical application of random projections (see Johnson and Lindenstrauss, 1984; Arriaga and Vempala, 1999). E.g., apply Corollary 19 in the paper by Ben-David et al. (2002).

and the following holds for every  $j \in \{1, \dots, n\}$ :

$$\begin{aligned} \frac{\gamma}{2\sqrt{d}} &\leq \sum_i p_i \langle u_i, v'_j \rangle A_{i,j} = \sum_i p_i A_{i,j} \sum_{k=1}^d |v'_{j,k}| \text{sign}(v_{j,k}) u_{i,k} \\ &= \sum_{k=1}^d |v'_{j,k}| \sum_i p_i \text{sign}(v_{j,k}) u_{i,k} A_{i,j} = \sum_{k=1}^d |v'_{j,k}| \langle \text{sign}(v_{j,k}) h_k, A_j \rangle_p \end{aligned}$$

The latter sum is a convex combination of inner products because of (19), and, as the above calculation shows, the inner products achieve a value of at least  $\gamma/(2\sqrt{d})$  on the average. By the pigeon-hole principle, there exists  $k(j) \in \{1, \dots, d\}$  such that

$$\text{sign}(v_{j,k(j)}) \cdot \langle h_{k(j)}, A_j \rangle_p \geq \frac{\gamma}{2\sqrt{d}} .$$

It is easily checked that  $\gamma/(2\sqrt{d}) \geq 1/d$  (by solving this inequality for  $d$  and comparing with the above definition of  $d$ ). It follows that, as announced above,  $h_1, \dots, h_d$  is universally correlated with  $A$ .  $\blacksquare$

**Conclusions:** Looking back, we have seen a hierarchy of margin optimization problems, and the dual versions of these problems nicely reflect why the optimal values become smaller when we go up in the hierarchy. In the dual setting, we are always faced with a problem of maximizing the total margin of a matrix of the form  $Y \circ A$  (which is the  $Y$ -average margin of  $A$ ). The crucial issues are the structure of the matrix  $Y$  and whether its choice is under control of “nature” or under control of an “intelligent adversary”:

- (a) The easiest problem from the perspective of the margin-maximizer results when  $Y$  is of the form  $p q^\top$  for fixed and “benign”  $p, q$ . Here “benign” means that the distribution  $p$  on the rows of  $A$  (= instances of the domain) and the distribution  $q$  on the columns of  $A$  (= possible target concepts) are resulting from a learning application (and not from settings within a worst-case analysis). In this situation the goal of the margin-maximizer roughly corresponds to achieving a reasonably large “soft margin” on the average.
- (b) The problem becomes harder when  $q$  (Case 1) or both of  $p, q$  (Case 2) are under control of an adversary so that  $\max_q \overline{\text{mc}}_{p,q}(A) = \overline{\text{mc}}_{p,\text{MIN}}(A)$  in Case 1 and  $\max_{p,q} \overline{\text{mc}}_{p,q}(A) = \max_p \overline{\text{mc}}_{p,\text{MIN}}(A)$  in Case 2 would be the appropriate complexity measures. Maximizing over all choices of  $q$  means choosing the target concept in a worst-case fashion. Maximizing over all choices of  $p$  (as in Case 2) means that the domain distribution is chosen in a worst-case fashion although it is still fixed (because the chosen arrangement may depend on  $p$ ). Because of the polynomial relation between average margin complexity and the SQ-dimension, Case 1 corresponds to weak learning in the SQ model under a fixed distribution. A similar remark is valid for Case 2 but here we have to cope with the hardest fixed distribution.
- (c) The hardest problem results when an adversary controls  $Y$ , and  $Y$  is an arbitrary matrix with non-negative entries summing up to 1 (as opposed to a matrix of the form

$pq^\top$ , which is the special case where  $Y$  has rank 1). Now  $\max_Y \overline{\text{mc}}_Y(A) = \text{mc}(A)$  is the appropriate complexity measure, and the learning goal is to achieve a reasonably large hard margin for every possible target concept. Because of the polynomial relation between  $\text{mc}(A)$  and the CSQ-dimension, the learning goal can be achieved iff the concept class is distribution-independently weakly learnable in the CSQ model.

Feldman (2008) has shown that there exist classes (e.g., Boolean decision lists) which are distribution independently (weakly or strongly)<sup>5</sup> learnable in the SQ model but not (not even weakly) in the CSQ model. This also shows that  $\max_{p,q} \overline{\text{mc}}_{p,q}(A)$  and  $\max_Y \overline{\text{mc}}_Y(A)$  are not polynomially related. (There is even an exponential gap.) Thus imposing the rank 1 constraint on  $Y$  makes much of a difference.

**Open Problems:** The level of distribution-independent SQ-learning is located somewhere between (b) and (c). It would be interesting to find a combinatorial parameter (or another variant of margin optimization?) that characterizes this level. A parameter of this kind must be lower-bounded by the SQ-dimension and upper-bounded by the CSQ-dimension. It would furthermore be interesting to find a concept class that separates distribution-independent SQ-learning from SQ-learning w.r.t. the hardest fixed distribution.

The correspondence between maximization of the average margin and typical soft-margin optimization problems would be more convincing if we replaced  $\gamma_{i,j}(A|\mathcal{A})$  by  $\min\{\gamma_{i,j}(A|\mathcal{A}), \gamma\}$  for some  $\gamma > 0$  so that few extremely large margin parameters cannot provide compensation for many small or negative margin parameters.<sup>6</sup> It would be interesting to know whether results similar to the ones in this paper can be shown for this “average clipped margin”.

## References

- Farid Alizadeh. Interior point methods to semidefinite programming with applications to combinatorial optimization. *SIAM Journal on Optimization*, 5(13), 1995.
- Rosa I. Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Proceedings of the 40'th Annual Symposium on the Foundations of Computer Science*, pages 616–623, 1999.
- Javed A. Aslam and Scott E. Decatur. General bounds on statistical query learning and PAC learning with noise via hypothesis boosting. *Information and Computation*, 141(2): 85–118, 1998.
- Shai Ben-David, Nadav Eiron, and Hans U. Simon. Limitations of learning via embeddings in euclidean half-spaces. *Journal of Machine Learning Research*, 3:441–461, 2002.
- Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the Association on Computing Machinery*, 50(4):506–519, 2003.

---

5. Because of the Boosting-result by Aslam and Decatur (1998) weak learners can be transformed into strong learners in this model without much loss of efficiency.

6. Furthermore note the connection to using hinge-loss in soft-margin optimization problems.

- Nader H. Bshouty and Vitaly Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2:359–395, 2002.
- Thorsten Doliwa, Michael Kallweit, and Hans U. Simon. Dimension and margin bounds for reflection-invariant kernels. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 157–167, 2008.
- Vitaly Feldman. Evolvability from learning algorithms. In *Proceedings of the 2008 ACM International Symposium on Theory of Computing*, pages 619–628, 2008.
- Jürgen Forster. A linear lower bound on the unbounded error communication complexity. *Journal of Computer and System Sciences*, 65(4):612–625, 2002.
- Jürgen Forster and Hans U. Simon. On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theoretical Computer Science*, 350(1):40–48, 2006.
- W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mapping into Hilbert spaces. *Contemp. Math.*, 26:189–206, 1984.
- Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the Association on Computing Machinery*, 45(6):983–1006, 1998.
- Troy Lee and Adi Shraibman. Lower bounds in communication complexity. *Foundations and Trends in Theoretical Computer Science*, 3(4):263–399, 2009.
- Nati Linial, Shahar Mendelson, Gideon Schechtman, and Adi Shraibman. Complexity measures of sign matrices. *Combinatorica*, 27(4):439–463, 2007.
- Alexander Sherstov. Halfspace matrices. *Computational Complexity*, 17(2):149–178, 2008.
- Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
- Leslie G. Valiant. Evolvability. *Journal of the Association on Computing Machinery*, 56(1):3:1–3:21, 2009.

## Appendix A. Proof of Theorem 3

We will prove the equivalent statement

$$\max_{\mathcal{A}} \min_j \sum_i p_i \gamma_{i,j}(A|\mathcal{A}) = \min_q \gamma_2^*(pq^\top \circ A) . \quad (20)$$

We know from the proof of Theorem 1 that  $\gamma_2^*(pq^\top \circ A)$  coincides with (8) provided that  $Y = pq^\top$ . Let us now discuss the left hand-side of (20). Setting  $M := P \cdot A$ , we obtain

$\min_j \sum_i p_i \gamma_{i,j}(A|\mathcal{A}) = \min_j \sum_i \gamma_{i,j}(M|\mathcal{A})$ .<sup>7</sup> Finding an arrangement  $\mathcal{A}$  for  $M$  that maximizes  $\min_j \sum_i \gamma_{i,j}(M|\mathcal{A})$  can be expressed as a standard SDP-problem (with slack variables  $s_j$ ) as follows:

$$\min_{X, \mu, s} -\mu \quad \text{s.t. } \forall k : X_{k,k} = 1, \forall j : \sum_i M_{i,j}(X_{i,m+j} + X_{m+j,i}) - s_j = 2\mu, X \succeq 0, \mu \geq 0, s_j \geq 0$$

The  $(m+2n+1) \times (m+2n+1)$ -matrix of primal variables is then given by

$$\begin{bmatrix} X & 0 & 0 \\ 0 & \text{diag}(s_1, \dots, s_n) & 0 \\ 0 & 0 & \mu \end{bmatrix}.$$

The dual variables are denoted  $y_k$  and  $q_j$ . The matrix induced by the equality-constraints equals

$$\begin{bmatrix} \text{diag}(y_1, \dots, y_m) & M \cdot Q & 0 & 0 \\ (M \cdot Q)^\top & \text{diag}(y_{m+1}, \dots, y_{m+n}) & 0 & 0 \\ 0 & 0 & -Q & 0 \\ 0 & 0 & 0 & -2(q_1 + \dots + q_n) \end{bmatrix}.$$

It is easy to see that condition SCQ is satisfied. Thus, we have strong duality. The dual problem (with variables  $-y_k/2$  substituted for  $y_k$ ,  $q_j/2$  substituted for  $q_j$ , and  $P \cdot A$  substituted for  $M$ ) looks as follows:

$$\min_{q,y} \frac{1}{2} \sum_k y_k \quad \text{s.t. } \sum_j q_j = 1, q_j \geq 0, \begin{bmatrix} \text{diag}(y_1, \dots, y_m) & -(P \cdot A \cdot Q) \\ -(P \cdot A \cdot Q)^\top & \text{diag}(y_{m+1}, \dots, y_{m+n}) \end{bmatrix} \succeq 0 \quad (21)$$

By strong duality,  $\max_{\mathcal{A}} \min_j \sum_i p_i \gamma_{i,j}(A|\mathcal{A})$  equals (21). As discussed above,  $\gamma_2^*(pq^\top \circ A)$  equals (8) provided that  $Y = pq^\top$  so that  $Y \circ A = pq^\top \circ A = P \cdot A \cdot Q$ . A comparison of (21) and (8) shows that (20) holds.

---

7. We remind the reader to the convention  $P = \text{diag}(p)$  and  $Q = \text{diag}(q)$ .

# Maximum Likelihood vs. Sequential Normalized Maximum Likelihood in On-line Density Estimation

**Wojciech Kotłowski**

*Centrum Wiskunde & Informatica*

KOTLOWSK@CWI.NL

**Peter Grünwald**

*Centrum Wiskunde & Informatica*

PDG@CWI.NL

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

The paper considers sequential prediction of individual sequences with log loss (online density estimation) using an exponential family of distributions. We first analyze the regret of the maximum likelihood (“follow the leader”) strategy. We find that this strategy is (1) suboptimal and (2) requires an additional assumption about boundedness of the data sequence. We then show that both problems can be addressed by adding the currently predicted outcome to the calculation of the maximum likelihood, followed by normalization of the distribution. The strategy obtained in this way is known in the literature as the *sequential normalized maximum likelihood* or *last-step minimax* strategy. We show for the first time that for general exponential families, the regret is bounded by the familiar  $(k/2) \log n$  and thus optimal up to  $O(1)$ . We also show the relationship to the Bayes strategy with Jeffreys’ prior.

**Keywords:** List of keywords

## 1. Introduction

The game of sequential prediction of individual sequences with log loss (online density estimation) is defined in the following way. Let  $x_1, x_2, \dots \in \mathcal{X}^*$ , be a sequence of outcomes revealed one at a time. After observing  $x^n = x_1, x_2, \dots, x_n$ , a forecaster assigns a probability distribution on  $\mathcal{X}$ , denoted  $P(\cdot | x^n)$ . Then, after  $x_{n+1}$  is revealed, the forecaster incurs the *log loss*  $-\log P(x_{n+1} | x^n)$ . The performance of the strategy is measured relative to the best in a reference set of strategies, which we call the *model*  $\mathcal{P}$ . The difference between the accumulated loss of the prediction strategy and the best strategy in the model is called the *regret*. The goal is to minimize the regret in the worst case over all possible data sequences.

We assume the model  $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$  is an exponential family of distributions, examples of which include normal, Bernoulli, multinomial, Gamma, Poisson, Pareto, geometric distributions and many others. If there is a known time horizon  $n$  of the game (maximal number of outcomes), the minimax strategy for the game is the *normalized maximum likelihood* (NML) strategy (Shtarkov, 1987; Rissanen, 1996). If the parameter space of a  $k$ -dimensional exponential family is constrained to a compact subset  $\Theta_0$ , NML achieves regret  $\frac{k}{2} \log n + O(1)$  for all data sequences. NML, however, requires knowledge of the time horizon and is impractical to calculate in many situations. A particularly simple and popular prediction strategy is the *maximum likelihood* (ML) strategy, (also known as “follow the

leader”), which predicts the next outcome  $x_n$  by using the distribution  $P_{\hat{\mu}_{n-1}}$ , with  $\hat{\mu}_{n-1}$  being the ML estimator based on the  $n - 1$  past outcomes. The ML strategy, contrary to NML, belongs to the family of *plug-in* strategies which in each iteration predict with one of the strategies from the model.

Despite the popularity of the “follow the leader” approach, guarantees on the regret of the ML strategy were only obtained for some special exponential families, such as normal, Bernoulli and Gamma distributions (Freund, 1996; Azoury and Warmuth, 2001). In this paper, we prove general bounds which hold for any exponential family. We show (Theorem 4, Section 3) that if the parameter space is constrained to a compact subset  $\Theta_0$ , and if the outcomes are bounded within a ball of radius  $B$ , the regret can be upper bounded by  $C \log n + O(1)$ , where  $C$  is a constant depending on  $B$  and  $\Theta_0$ . We also prove (Theorem 5) that the bound is essentially tight. In other words, (1) the ML strategy requires boundedness of the data sequence, and (2) the rate of the regret growth is still logarithmic, but the constant in front of  $\log n$  can be very large, especially when  $B$  is large. Moreover, Theorem 5 implies that those two drawbacks are shared among all plug-in strategies.

We also show, however, that both problems can be addressed by adding the currently predicted outcome to the calculation of the maximum likelihood, followed by normalization of the distribution. Typically, the new strategy predicts with a distribution  $P(x_n|x^{n-1})$  proportional to  $P_{\hat{\mu}_n}(x_n)$ . Usually, this distribution will not be equal to any of the distributions  $P_\mu$  within the model, so the new strategy is not plug-in. The strategy obtained in this way is known as *sequential normalized maximum likelihood* (Rissanen and Roos, 2007; Roos and Rissanen, 2008) (SNML). It was discovered with a different motivation in mind: Rissanen and Roos noticed that its predictions coincide with those of the NML distribution under the assumption that the current iteration is the last iteration. Therefore, it can be viewed as an approximation to NML for which the time horizon of the game does not need to be known. A similar idea, though restricted to strategies within the model (plug-in strategies), was introduced by Takimoto and Warmuth (2000) under the name *last-step minimax*.

In this paper, we develop bounds on the worst-case regret for SNML for general exponential families (such bounds had been unknown so far). As our main result, in Theorem 7, we prove that the regret of the SNML strategy is at most  $\frac{k}{2} \log n + O(1)$ , which matches, up to the  $O(1)$  term, the minimax regret bound. This issue is important from a practical point of view, as SNML constitutes an interesting and effective algorithm for online density estimation and model selection. However, our results are also interesting from a conceptual point of view, as the answer to the following question: how much do we lose if we base our decision in a given moment by looking only one step ahead instead of looking at the whole possible future up to a given time horizon? Our results suggest that we do not lose anything substantial, at least asymptotically; for some models, it turns out that we don’t even lose anything: in Section 5 we show that in some cases (but not always) the SNML strategy coincides with the *Bayes* strategy, when the prior distribution is chosen to be a *Jeffreys’ prior*. Moreover, we prove that when the two strategies are equal, they are also equal to the NML strategy and thus minimax optimal.

**Related Work** Sequential prediction with log loss has been extensively studied in learning theory, in the framework of *prediction with expert advice* (Cesa-Bianchi and Lugosi, 2006). It also plays an important role in information theory: a key result based on the Kraft

inequality (Cover and Thomas, 1991) states that, ignoring rounding issues, for every length function  $L$  of a uniquely decodable code, there is a probability distribution  $P$  such that  $L(x) = -\log P(x)$  and vice versa. Thus, at least when  $\mathcal{X}$  is countable, any prediction strategy can also be thought of as a *universal source coding algorithm*; the cumulative logarithmic loss corresponds exactly to the incurred codelength. As Rissanen's theory of Minimum Description Length (MDL) learning (Barron et al., 1998; Grünwald, 2007) is based on universal coding, a sequential prediction strategy with log loss defines an MDL model selection criterion. Similarly, in statistics, Dawid's theory of prequential model assessment (Dawid, 1984) is based on sequential prediction.

The ML strategy for exponential families was considered by Freund (1996) and Azoury and Warmuth (2001), with regret bounds proven for the particular cases of normal, Bernoulli and Gamma distributions. Grünwald and de Rooij (2005) showed the following: let the model  $\mathcal{P}$  be an arbitrary 1-dimensional exponential family ( $k = 1$ ). Suppose the outcomes are i.i.d. by some distribution  $P^*$ , possibly outside the model. Then the expected regret of the ML plug-in strategy is  $(1/2c) \log n + O(1)$  where  $c$  is the variance of an outcome under the true distribution  $P^*$  divided by the variance under the element of the model  $P_\theta$  that minimizes the Kullback-Leibler divergence  $D(P^* \| P_\theta)$ . In general,  $c$  can be much smaller than 1. Moreover, it was shown by Grünwald and Kotłowski (2010) that *no* plug-in estimator can achieve  $c = 1$ . Our Theorems 4 and 5 are essentially extensions of this result to individual-sequence settings. Dasgupta and Hsu (2007) considered Gaussian density estimation with unknown mean and covariance matrix and obtained a much worse linear bound on the regret, excluding the possibility of a logarithmic bound. Their results do not contradict ours, since the set of reference strategies (distributions in the exponential family) was not constrained to be in a compact subset of the parameter space, which is necessary to obtain logarithmic bounds (interestingly, if one considers the regret conditioned on the first outcome then for some models it is possible after all to get logarithmic bound even with full parameter space, as we show in Section 5; we pose as an open problem whether this result extends to arbitrary exponential families). Raginsky et al. (2009) considered a plug-in strategy based on Bregman projections and proved regret bounds for general exponential families; their strategy, however, is different from those considered here. Hazan et al. (2007) proved logarithmic regret bounds on the follow the leader strategy in online convex optimization, however the assumptions of their theorem do not match online density estimation with exponential families. Kotłowski et al. (2010) considered “following the ‘flattened’ leader”, an improvement over the ML strategy, “slightly” outside the model, achieving the optimal regret bound. However, this flattened-leader strategy still requires boundedness of the data sequence.

The idea of including the current observation to the calculation of maximum likelihood was considered by (Rissanen and Roos, 2007; Roos and Rissanen, 2008), though with a different motivation in mind. The regret bounds were not given apart from specific cases. A similar idea, though restricted to strategies within the model (plug-in strategies), was introduced by Takimoto and Warmuth (2000) under the name *last-step minimax* (the relation is made precise in Section 4).

The paper is organized as follows. We introduce the mathematical context for our results in Section 2. We then analyze the ML strategy in Section 3, proving the regret bounds

which reveal suboptimal behavior in the worst case. Then, we introduce the SNML strategy in Section 4 and prove optimal regret bounds. We give some examples for particular exponential families in Section 5 and discuss the relationship between SNML and Bayes with Jeffreys' prior in Section 6. We end with a conclusion in Section 7.

## 2. Notation and Definitions

### 2.1. Exponential Family

Let  $\mathcal{X}$  be a set of outcomes, taking values either in a finite or countable set, or in a subset of Euclidean space. Exponential family models (Barndorff-Nielsen, 1978) are families of distributions on  $\mathcal{X}$  with densities  $P_\theta(x) = e^{\theta^T \phi(x) - \psi(\theta)} h(x)$ , defined relative to a random variable  $\phi : \mathcal{X} \rightarrow \mathbb{R}^k$  (called *sufficient statistic*) and a function  $h : \mathcal{X} \rightarrow [0, \infty)$ . The function  $\psi(\theta) = \log \int_{x \in \mathcal{X}} e^{\theta^T \phi(x)} h(x) dx$  (the integral to be replaced by a sum for countable  $\mathcal{X}$ ) is called a *partition function*, and  $\Theta = \{\theta \in \mathbb{R}^k : \psi(\theta) < \infty\}$  is called the *natural parameter space*. We only consider *regular* exponential families, when  $\Theta$  is an open and convex subset of  $\mathbb{R}^k$ , and the representation is *minimal*, i.e. the functions  $\phi_i(x), i = 1, \dots, k$ , are linearly independent. Moreover, without loss of generality, we will make the simplifying assumption that  $\phi(x) \equiv x$ , i.e. the exponential family is in the canonical form. All results in this paper are valid for a more general  $\phi$ . The function  $\psi(\theta)$  is differentiable infinitely often, and strictly convex on  $\Theta$ . A standard result for exponential families states (Barndorff-Nielsen, 1978) that the gradient  $\mu = \nabla_\theta \psi(\theta)$  is the mean value vector of  $x$ ,  $\mu = \mathbb{E}_\theta[x]$ , while the Hessian  $\nabla_\theta^2 \psi(\theta) = E_\theta[-\nabla_\theta^2 \log P_\theta(x)] = I(\theta)$  coincides with the *Fisher information* matrix in the natural parameterization, is positive definite and equal to the covariance matrix  $\text{Cov}_\theta(x)$ . Strict convexity of  $\psi$  implies that the function  $\mu(\theta)$  is invertible, and thus suggests reparameterizing the distribution by  $\mu$ . The function  $\mu(\theta)$  maps parameters in the natural parameterization to the *mean value* parameterization  $\Xi = \mu(\Theta)$ . It is a diffeomorphism (Barndorff-Nielsen, 1978), and thus  $\Xi$  is also an open convex set of  $\mathbb{R}^k$ . The inverse  $\theta(\mu)$  maps back to the natural parametrization. Moreover,  $\nabla_\mu \theta(\mu) = E_\mu[-\nabla_\mu^2 \log P_\mu(x)] = I(\mu)$  is the Fisher information in the mean-value parametrization, which is equal to the inverse covariance matrix  $\text{Cov}_\mu^{-1}(x)$ .

The KL-divergence between distributions  $P_\theta$  and  $P_{\theta'}$ :

$$\mathbb{E}_\theta \left[ \log \frac{P_\theta(x)}{P_{\theta'}(x)} \right] = \mathbb{E}_\mu \left[ \log \frac{P_\mu(x)}{P_{\mu'}(x)} \right] = (\theta - \theta')^T \mu - \psi(\theta) + \psi(\theta'), \quad (1)$$

where  $\mu = \mu(\theta)$  and  $\mu' = \mu(\theta')$ , is denoted by  $D(\theta||\theta')$  or by  $D(\mu||\mu')$ , depending on the context.

Let  $\Theta_0 \subseteq \Theta$  be any nonempty convex subset of  $\Theta$ . Given the data sequence  $x^n$ , the *maximum likelihood* (ML) estimate  $\hat{\theta}_n$  relative to  $\Theta_0$  is defined as:

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta_0} P_\theta(x^n) = \arg \min_{\theta \in \Theta_0} -\log P_\theta(x^n), \quad (2)$$

or equivalently as:

$$\hat{\mu}_n := \mu(\hat{\theta}_n) = \arg \min_{\mu \in \Xi_0} -\log P_\mu(x^n),$$

where  $\Xi_0 = \mu(\Theta_0)$  is also convex. By rewriting  $-\log P_\theta(x^n) = -n(\theta^T \bar{x}_n - \psi(\theta)) - \log h(x^n)$ , where  $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ , we see that  $\nabla_\theta - \log P_\theta(x^n) = -n(\bar{x}_n - \mu(\theta))$ . This means than when  $\bar{x}_n \in \Xi_0$ , then  $\hat{\mu}_n = \bar{x}_n$ . More generally, we can exploit the fact that  $-\log P_\theta(x^n)$  is a convex function of  $\theta$ , and thus the necessary condition for a minimum of a convex function on a convex set states (Boyd and Vandenberghe, 2004) that  $\nabla_{\hat{\theta}_n} - \log P_{\hat{\theta}_n}(x^n)(\theta - \hat{\theta}_n) \geq 0$  for all  $\theta \in \Theta_0$ , which implies:

$$(\hat{\mu}_n - \bar{x}_n)^T (\theta - \hat{\theta}_n) \geq 0 \quad (3)$$

for all  $\theta \in \Theta_0$ . Condition (3) has a nice interpretation in terms of Bregman projections. Assuming  $\bar{x}_n \in \Xi$ , we can rewrite (3) as:

$$D(\bar{x}_n \| \mu) \geq D(\hat{\mu}_n \| \mu) + D(\bar{x}_n \| \hat{\mu}_n),$$

for all  $\mu \in \Xi_0$ , which is closely related to the generalized Pythagorean inequality for Bregman divergences (Cesa-Bianchi and Lugosi, 2006). Another expression which we are going to use is:

$$D(\mu_1 \| \mu_2) - D(\mu_1 \| \mu_3) = D(\mu_3 \| \mu_2) + (\theta_2 - \theta_3)^T (\mu_3 - \mu_1), \quad (4)$$

for all  $\mu_1, \mu_2, \mu_3 \in \Xi_0$ , where  $\theta_i = \theta(\mu_i)$ ,  $i = 1, 2, 3$ . This can be derived by writing KL-divergences on both sides according to (1).

## 2.2. Sequential Prediction

At every iteration  $n = 1, 2, \dots$ , the prediction  $P(\cdot | x^{n-1})$  depends on the past outcomes  $x^{n-1}$  and has the form of a probability distribution on  $\mathcal{X}$ , and therefore can be considered as a conditional of the joint distribution of outcomes in  $\mathcal{X}^n$ , which is  $P(x^n) = \prod_{i=1}^n P(x_i | x^{i-1})$ . Conversely, any probability distribution  $P$  on the set  $\mathcal{X}^n$  defines a prediction strategy induced by its conditional distributions  $P(\cdot | x^i)$  for  $0 \leq i < n$  (Cesa-Bianchi and Lugosi, 2006; Grünwald, 2007). The performance of the strategy  $P$  on the outcome sequence  $x^n$  is measured relative to the best strategy in the model (reference set of strategies)  $\mathcal{P}$  by the *regret*, defined as:

$$\mathcal{R}(P; x^n) = \sum_{i=1}^n -\log P(x_i | x^{i-1}) - \inf_{P_\theta \in \mathcal{P}} \sum_{i=1}^n -\log P_\theta(x_i) = -\log P(x^n) - \inf_{P_\theta \in \mathcal{P}} -\log P_\theta(x^n). \quad (5)$$

The regret is the difference in cumulative losses incurred so far by the prediction strategy and the best strategy (distribution) in the model. We are usually interested in the worst-case regret,  $\mathcal{R}(P; n) = \sup_{x^n} \mathcal{R}(P; x^n)$ . Unfortunately, for most common exponential families  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$  (extended to sequences by the i.i.d. assumption),  $\mathcal{R}(P; n)$  cannot be made finite even for  $n = 1$ , whatever  $P$  is. Indeed, let  $\mathcal{P}$  be a one-dimensional normal family  $N(\theta, 1)$  with fixed unit variance, so that  $\Theta = \mathbb{R}$ . For every strategy  $P$ , we must have  $P(x_1) \rightarrow 0$  as  $x_1 \rightarrow \infty$  (otherwise  $P$  would not be normalizable), and therefore  $-\log P(x_1) \rightarrow \infty$ . On the other hand,  $\inf_{P_\theta \in \mathcal{P}} -\log P_\theta(x_1) = -\log P_{x_1}(x_1) = \frac{1}{2} \log 2\pi$ , so that by increasing  $x_1$ , we can make the regret as large as we want.

Therefore, to obtain non-trivial regret bounds, we choose a compact convex subset  $\Theta_0 \subset \Theta$  and define  $\mathcal{P} = \{P_\theta : \theta \in \Theta_0\}$ ; equivalently, we can choose  $\Xi_0 = \mu(\Theta_0)$  and define

$\mathcal{P} = \{P_\mu : \mu \in \Xi_0\}$  (we will use both interchangeably). Then,

$$\mathcal{R}(P; x^n) = -\log P(x^n) + \log P_{\hat{\mu}_n}(x^n),$$

where  $\hat{\mu}_n$  is the ML estimator relative to  $\Xi_0$ .

Let  $P$  be a prediction strategy. If for every  $n$ ,  $P(x_n|x^{n-1}) \in \mathcal{P}$ , i.e.  $P(x_n|x^{n-1}) = P_{\bar{\mu}_{n-1}}(x_n)$  for some  $\bar{\mu}_{n-1} = \bar{\mu}_{n-1}(x^{n-1})$ , we call such  $P$  a *plug-in strategy*. In other words, a plug-in strategy always predicts with one of the distributions from the model. An example of a plug-in strategy is the *maximum likelihood* (or *follow the leader*) strategy defined as  $P_{\text{ML}}(x_n|x^{n-1}) = P_{\bar{\mu}_{n-1}}(x_n)$ .

There is, however, an advantage in using strategies which are not in the model. An important out-model strategy is the *normalized maximum likelihood (NML)* strategy, defined as:

$$P_{\text{NML}}(x^n) = \frac{\sup_{\theta \in \Theta_0} P_\theta(x^n)}{\int_{\mathcal{X}^n} \sup_{\theta \in \Theta_0} P_\theta(z^n) dz^n} = \frac{\sup_{\mu \in \Xi_0} P_\mu(x^n)}{\int_{\mathcal{X}^n} \sup_{\mu \in \Xi_0} P_\mu(z^n) dz^n}. \quad (6)$$

NML is known to be the *minimax* prediction strategy for the log-loss game: it can be shown (see, e.g. Cesa-Bianchi and Lugosi (2001)) that:

$$\inf_P \sup_{x^n} \mathcal{R}(P; x^n) = \mathcal{R}(P_{\text{NML}}; n),$$

where the infimum is over all, both in-model (plug-in) and out-model, prediction strategies. The value of  $\mathcal{R}(P_{\text{NML}}; n)$  is also known: if  $\mathcal{P}$  is a  $k$ -dimensional exponential family and  $\Xi_0$  is a closed convex subset of  $\Xi$  with non-empty interior, then

$$\mathcal{R}(P_{\text{NML}}, x^n) = \frac{k}{2} \log n + O(1). \quad (7)$$

For a proof, see e.g. (Grünwald, 2007). (7) is the famous ‘ $k$  over  $2 \log n$  formula’, refinements of which lie at the basis of practical approximations to MDL and Bayesian learning (Grünwald, 2007). Since the NML strategy is minimax, a worst-case regret of  $\frac{k}{2} \log n + O(1)$  is optimal.

### 3. Regret Bounds for ML Strategy

In this section, we analyze the performance of ML strategy and show, that under an additional boundedness assumption on the data sequence, one can get a non-trivial regret bound. The bound, however, reveals suboptimal behavior and dependence on the range of the data sequence. Later, we show the bound is essentially unimprovable for any plug-in strategy.

Before we prove the main theorem, we need several propositions, which will also be useful in the next section, while proving the regret bound for SNML.

**Proposition 1** Let  $\bar{y}_n = \frac{(n-1)\hat{\mu}_{n-1} + x_n}{n}$ , and let  $\tilde{\mu}_n = \arg \min_{\mu \in \Xi_0} D(\bar{y}_n \| \mu)$ . Then:

$$-\log P_{\hat{\mu}_{n-1}}(x^n) + \log P_{\tilde{\mu}_n}(x^n) \leq nD(\bar{y}_n \| \hat{\mu}_{n-1}) - nD(\bar{y}_n \| \tilde{\mu}_n). \quad (8)$$

**Proof** From the definition of  $\tilde{\mu}_n$ , we have  $D(\bar{y}_n\|\tilde{\mu}_n) \leq D(\bar{y}_n\|\hat{\mu}_n)$ , so that:

$$\begin{aligned}
 D(\bar{y}_n\|\hat{\mu}_{n-1}) - D(\bar{y}_n\|\tilde{\mu}_n) &\geq D(\bar{y}_n\|\hat{\mu}_{n-1}) - D(\bar{y}_n\|\hat{\mu}_n) && \text{(from definition of } \tilde{\mu}_n\text{)} \\
 &= D(\hat{\mu}_n\|\hat{\mu}_{n-1}) + (\hat{\theta}_{n-1} - \hat{\theta}_n)^T(\hat{\mu}_n - \bar{y}_n) && \text{(from (4))} \\
 &= D(\hat{\mu}_n\|\hat{\mu}_{n-1}) + (\hat{\theta}_{n-1} - \hat{\theta}_n)^T(\hat{\mu}_n - \bar{x}_n) + (\hat{\theta}_{n-1} - \hat{\theta}_n)^T(\bar{x}_n - \bar{y}_n) \\
 &= D(\hat{\mu}_n\|\hat{\mu}_{n-1}) + (\hat{\theta}_{n-1} - \hat{\theta}_n)^T(\hat{\mu}_n - \bar{x}_n) + \frac{n}{n-1}(\hat{\theta}_{n-1} - \hat{\theta}_n)^T(\bar{x}_{n-1} - \hat{\mu}_{n-1}) \\
 &\geq D(\hat{\mu}_n\|\hat{\mu}_{n-1}) + (\hat{\theta}_{n-1} - \hat{\theta}_n)^T(\hat{\mu}_n - \bar{x}_n) && \text{(from (3))} \\
 &= -(\hat{\theta}_{n-1} - \hat{\theta}_n)^T\bar{x}_n + \psi(\hat{\theta}_{n-1}) - \psi(\hat{\theta}_n) \\
 &= \frac{1}{n}(-\log P_{\hat{\mu}_{n-1}}(x^n) + \log P_{\hat{\mu}_n}(x^n)).
 \end{aligned}$$

■

Proposition 1 states that if we pretend  $\hat{\mu}_{n-1}$  (rather than  $\bar{x}_{n-1}$ ) is the sufficient statistic in the previous iteration, and do all the updates accordingly, then the drop of the KL divergence from the data to the ML estimator per iteration will decrease. Thus, if we imagine the data is generated by an adversary trying to maximize the regret, it does not pay for him/her to choose  $\bar{x}_{n-1}$  outside  $\Xi_0$  (since then  $\hat{\mu}_{n-1} \neq \bar{x}_{n-1}$ ).

We now show that within a compact set  $\Xi_0$ , the KL-divergence behaves approximately as a quadratic form:

**Proposition 2** Let  $\Xi_1$  be a compact subset of  $\Xi$ . Then, for all  $\mu, \mu' \in \Xi_1$ ,

$$D(\mu\|\mu') \leq \frac{1}{2}(\mu - \mu')^T I(\mu')(\mu - \mu') + C\|\mu - \mu'\|^3,$$

where  $C < \infty$  depends on  $\Xi_1$ .

**Proof** We need two standard results regarding the properties of KL divergence (see, e.g. Barndorff-Nielsen (1978); Grünwald (2007)): for any  $\mu, \mu' \in \Xi$ , it holds:

1.  $D(\mu\|\mu') \geq 0$  and the equality only holds for  $\mu = \mu'$ ,
2. For exponential families,  $\nabla_\mu^2 D(\mu\|\mu') = I(\mu)$ .

By Taylor expanding  $D(\mu\|\mu')$  around  $\mu'$  up to the second order, we get:

$$D(\mu\|\mu') = D(\mu'\|\mu') + \nabla_\mu D(\mu\|\mu')^T|_{\mu=\mu'}(\mu - \mu') + \frac{1}{2}(\mu - \mu')^T I(\bar{\mu})(\mu - \mu'),$$

for some  $\bar{\mu}$  between  $\mu$  and  $\mu'$ . Due to the first property the zeroth order term disappears; the second order term also disappears because the gradient vanishes at the minimum, so we have:

$$\begin{aligned}
 D(\mu\|\mu') &= \frac{1}{2}(\mu - \mu')^T I(\bar{\mu})(\mu - \mu') = \frac{1}{2}(\mu - \mu')^T I(\mu')(\mu - \mu') + \frac{1}{2}(\mu - \mu')^T(I(\bar{\mu}) - I(\mu'))(\mu - \mu') \\
 &\leq \frac{1}{2}(\mu - \mu')^T I(\mu')(\mu - \mu') + \frac{1}{2}\|I(\bar{\mu}) - I(\mu')\| \|\mu - \mu'\|^2,
 \end{aligned} \tag{9}$$

where  $\|\cdot\|$  denotes vector or matrix norm, depending on the context. Taylor expanding  $I(\bar{\mu})$  around  $\mu'$  up to the first order gives  $I(\bar{\mu}) = I(\mu') + \nabla I(\tilde{\mu})^T(\bar{\mu} - \mu')$ , for some  $\tilde{\mu}$  between  $\bar{\mu}$  and  $\mu'$ . From that we get:

$$\|I(\bar{\mu}) - I(\mu')\| \leq \|\nabla I(\tilde{\mu})\| \|\bar{\mu} - \mu'\| \leq C \|\bar{\mu} - \mu'\|, \quad (10)$$

where  $C = \sup_{\mu \in \Xi_1} \|\nabla I(\mu)\|$  is finite due to compactness of  $\Xi_1$  and continuity of all derivatives of the information matrix. It follows from the definition of  $\bar{\mu}$  that  $\|\bar{\mu} - \mu'\| \leq \|\mu - \mu'\|$ ; using this in (10) and plugging the result into (9) finishes the proof. ■

**Proposition 3** *Let the data sequence  $x_1, x_2, \dots$  be such that  $\|x_n\| \leq B$  for all  $n$ . Then, for all large  $n$ ,*

$$\log P_{\hat{\mu}_n}(x^n) - \log P_{\hat{\mu}_{n-1}}(x^n) \leq \frac{1}{2n} (\hat{\mu}_{n-1} - x_n)^T I(\hat{\mu}_{n-1})(\hat{\mu}_{n-1} - x_n) + \frac{C}{n^2}, \quad (11)$$

where  $C$  depends on both  $\Xi_0$  and  $B$ .

**Proof** Proposition 1 states that the left hand side of (11) is upper bounded by  $nD(\bar{y}_n \|\hat{\mu}_{n-1}) - nD(\bar{y}_n \|\tilde{\mu}_n)$ . Let  $\Xi_1 \subset \Xi$  be a compact set such that  $\Xi_0 \subset \Xi_1$  and:

$$\inf_{\mu \in \Xi \setminus \Xi_1, \mu' \in \Xi_0} \|\mu - \mu'\| \geq \delta$$

for some  $\delta > 0$ . In other words,  $\Xi_0$  and the outside of  $\Xi_1$  never come arbitrarily close to each other. Such a set always exists, because  $\Xi$  is open, while  $\Xi_0$  is compact. Compactness of  $\Xi_0$  also imply that  $\|\hat{\mu}_n\| \leq C_{\Xi_0}$  for some  $C_{\Xi_0} < \infty$ . Due to boundedness of  $x_n$  we have:

$$\|\bar{y}_n - \hat{\mu}_{n-1}\| = \left\| \frac{x_n - \hat{\mu}_{n-1}}{n} \right\| \leq \frac{B + C_{\Xi_0}}{n} \leq \delta,$$

for all sufficiently large  $n$ , which implies that  $\bar{y}_n \in \Xi_1$ . Using first Proposition 1, and then Proposition 2 with  $\mu = \bar{y}_n$  and  $\mu' = \hat{\mu}_{n-1}$  for compact set  $\Xi_1$ , we get:

$$\begin{aligned} \log P_{\hat{\mu}_n}(x^n) - \log P_{\hat{\mu}_{n-1}}(x^n) &\leq nD(\bar{y}_n \|\hat{\mu}_{n-1}) - nD(\bar{y}_n \|\tilde{\mu}_n) \leq nD(\bar{y}_n \|\hat{\mu}_{n-1}) \\ &\leq \frac{n}{2} (\bar{y}_n - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(\bar{y}_n - \hat{\mu}_{n-1}) + nC \|\bar{y}_n - \hat{\mu}_{n-1}\|^3 \\ &= \frac{1}{2n} (x_n - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(x_n - \hat{\mu}_{n-1}) + \frac{C}{n^2} \|x_n - \hat{\mu}_{n-1}\|^3 \\ &\leq \frac{1}{2n} (x_n - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(x_n - \hat{\mu}_{n-1}) + \frac{C(B + C_{\Xi_0})^3}{n^2}. \end{aligned}$$

■

With the propositions stated above, we are able to prove the main theorem of this section:

**Theorem 4** Let the data sequence  $x_1, x_2, \dots$  be such that  $\|x_n\| \leq B$ . Then,

$$\mathcal{R}(P_{\text{ML}}; x^n) \leq \frac{I_{\Xi_0}(B + C_{\Xi_0})^2}{2} \log n + O(1),$$

where  $C_{\Xi_0} = \max_{\mu \in \Xi_0} \|\mu\|$  and  $I_{\Xi_0} = \max_{\mu \in \Xi_0} \|I(\mu)\|$ .

**Proof** Proposition 3 states that there exists  $n_0$  such that for all  $n \geq n_0$ ,

$$\log P_{\hat{\mu}_n}(x^n) - \log P_{\hat{\mu}_{n-1}}(x^n) \leq \frac{1}{2n} (\hat{\mu}_{n-1} - x_n)^T I(\hat{\mu}_{n-1})(\hat{\mu}_{n-1} - x_n) + \frac{C}{n^2} \leq \frac{I_{\Xi_0}(B + C_{\Xi_0})^2}{2n} + \frac{C}{n^2}$$

For  $n < n_0$ ,

$$\log P_{\hat{\mu}_n}(x^n) - \log P_{\hat{\mu}_{n-1}}(x^n) \leq -\log P_{\hat{\mu}_{n-1}}(x^n) = \bar{x}_n \hat{\mu}_{n-1} - \psi(\hat{\theta}_{n-1}) \leq C < \infty,$$

due to compactness of  $\Xi_0$  and boundedness of  $\bar{x}_n$ . Using those bounds, we get:

$$\begin{aligned} \mathcal{R}(P_{\text{ML}}; x^n) &= \sum_{i=1}^n -\log P_{\hat{\theta}_{i-1}}(x_i) + \log P_{\hat{\theta}_n}(x^n) = \sum_{i=1}^n -\log P_{\hat{\theta}_{i-1}}(x_i) + \log P_{\hat{\theta}_i}(x^i) - \log P_{\hat{\theta}_{i-1}}(x^{i-1}) \\ &= \sum_{i=1}^n -\log P_{\hat{\theta}_{i-1}}(x^i) + \log P_{\hat{\theta}_i}(x^i) \leq O(1) + \sum_{i=n_0}^n -\log P_{\hat{\theta}_{i-1}}(x^i) + \log P_{\hat{\theta}_i}(x^i) \\ &\leq O(1) + \frac{I_{\Xi_0}(B + C_{\Xi_0})^2}{2} \sum_{i=n_0}^n \frac{1}{i} + C \sum_{i=n_0}^n \frac{1}{n^2} = \frac{I_{\Xi_0}(B + C_{\Xi_0})^2}{2} \log n + O(1). \end{aligned}$$

■

Theorem 4 states that when playing against distributions from a compact subset of the parameter space, the ML strategy achieves the logarithmic regret growth. However, the constant factor in front of the logarithm can be very large, especially when the bound on the data sequence  $B$  is large.

An important question is whether the bound in Theorem 4 is improvable or whether any of the assumptions of the theorem can be relaxed? The answer appears to be negative. First, without restricting the reference strategies to the compact subset  $\Xi_0$ , one cannot account for a logarithmic regret at all. Dasgupta and Hsu (2007) consider predicting with Gaussian distributions with unknown mean and covariance without restricting the parameter space, and show that the ML strategy will incur linear regret growth. Second, the following theorem shows that one cannot avoid the boundedness assumption and the bound in Theorem 4 is essentially unimprovable:

**Theorem 5** Let  $k = 1$ , i.e. the exponential family is one-dimensional. Let  $P$  be a plug-in prediction strategy, i.e.  $P(x_n|x^{n-1}) = P_{\bar{\mu}_{n-1}}(x_n)$ , for some  $\bar{\mu}_{n-1} = \bar{\mu}_{n-1}(x^{n-1}) \in \Xi$ ,  $n = 1, 2, \dots$ . Then, for Lebesgue almost all  $\mu \in \Xi$  (all except Lebesgue measure zero set), there exists a data sequence  $x_1, x_2, \dots$ , such that  $\|x_n - \mu\| \leq B$ , for which:

$$\mathcal{R}(P; x^n) \geq \frac{I(\mu)B^2}{2} \log n + O(1).$$

**Proof** We will use a theorem from Grünwald and Kotłowski (2010), which states for any plug-in strategy  $P$  and one dimensional exponential family  $\mathcal{M}$ , when the outcomes are generated i.i.d. by some distribution  $P^*$ , with  $\mathbb{E}_{P^*}[x] = \mu^* \in \Xi$ ,

$$\mathbb{E}_{P^*}[\mathcal{R}(P; x^n)] \geq \frac{1}{2} \frac{\text{var}_{P^*}(x)}{\text{var}_{P_{\mu^*}}(x)} \log n + O(1), \quad (12)$$

for Lebesgue almost all  $\mu^* \in \Xi$ , where  $\text{var}$  denotes the variance.

We will apply the theorem with  $P^*$  which distributes its mass equally on the two outcomes  $x = \mu + B$  and  $x = \mu - B$ . Then,  $\mathbb{E}_{P^*}[x] = \mu^* = \mu$  and  $\text{var}_{P^*}(x) = B^2$ . Moreover,  $\text{var}_{P_{\mu^*}}(x) = I^{-1}(\mu)$ . Since the bound (12) holds in expectation, it must also hold for some particular data sequence.  $\blacksquare$

#### 4. Regret Bounds for Sequential Normalized Maximum Likelihood Strategy

In the previous section, it was shown that the ML strategy behaves suboptimally and cannot achieve logarithmic regret without bounding the data sequence. In this section, we present a modification of the ML strategy, that achieves logarithmic regret without any boundedness assumptions on the data, and also achieves the optimal constant in front of the logarithm.

The modification is based on calculating the ML estimator of the data sequence *including the current observation*. In other words, the strategy predicts  $x_n$  with a distribution proportional to  $P_{\hat{\mu}_n}(x_n)$ :  $P(x_n|x^{n-1}) \propto P_{\hat{\mu}_n}(x_n)$ . The proportionality constant differs from unity since  $P_{\hat{\mu}_n}(x_n)$  does not normalize properly anymore ( $\hat{\mu}_n$  depends on  $x_n$ ). If  $\hat{\mu}_n^x$  denotes the ML estimator for the data sequence  $x_1, \dots, x_{n-1}, x$  (i.e.  $\hat{\mu}_n^{x_n} = \hat{\mu}_n$ ), we can write the strategy as follows:

$$P(x_n|x^{n-1}) = \frac{P_{\hat{\mu}_n}(x^n)}{\int_{\mathcal{X}} P_{\hat{\mu}_n^x}(x^{n-1}, x) dx}. \quad (13)$$

Typically, the strategy will *not* predict with one of the distributions from the model. The strategy (13) is known as *sequential normalized maximum likelihood* (SNML) (Rissanen and Roos, 2007; Roos and Rissanen, 2008) and will be denoted  $P_{\text{SNML}}$ . It was arrived at from a different starting point: by noticing that (13) is the prediction of the NML distribution in the  $n$ -th iteration, with time horizon  $n$ . In other words, SNML predicts as NML, assuming that the current iteration is the last iteration. Therefore, contrary to NML, the time horizon of the game does not need to be known. As noted by Rissanen and Roos (2007), the prediction of the SNML strategy can be defined as the solution to the following minimax problem:

$$P_{\text{SNML}}(\cdot|x_{n-1}) = \arg \min_{P(\cdot|x_{n-1})} \max_{x_n} \mathcal{R}(P; x^n) = \arg \min_{P(\cdot|x_{n-1})} \max_{x_n} \left\{ -\log P(x_n|x^{n-1}) + \log P_{\hat{\mu}_n}(x^n) \right\}. \quad (14)$$

A similar idea, though restricted to strategies within the model, was introduced by Takimoto and Warmuth (2000) under the name *last-step minimax*. It is defined as (14), except that the argmin is only over the distributions from the model  $\mathcal{P}$ . However, the strategy obtained in such a way is a plug-in strategy, and plug-in strategies were already ruled out from having optimal regret bound by Theorem 5.

Surprisingly, the behavior of the worst-case regret for (13) for general exponential families has not been studied before. At the same time, this issue seems to be important, not only from a practical point of view (as SNML constitutes an effective algorithm for online density estimation and model selection), but also from a conceptual point of view. It raises the following question: how much do we loose if we base our decision in a given moment by looking only one step ahead instead of looking at the whole possible future up to a given time horizon. In this section, we show that we do not loose much, at least asymptotically: the regret of the SNML strategy is at most  $\frac{k}{2} \log n + O(1)$  for all exponential families, which matches, up to the  $O(1)$  term, the minimax regret bound (in Section 5 we will see that for some exponential families even the difference in the  $O(1)$  terms is negligible).

We start by rewriting the regret of SNML strategy using (in the second line) telescoping:

$$\begin{aligned}\mathcal{R}(P; x^n) &= -\log P(x^n) + \log P_{\hat{\mu}_n}(x^n) \\ &= \sum_{i=1}^n -\log P(x_i | x^{i-1}) + \log P_{\hat{\mu}_i}(x^i) - \log P_{\hat{\mu}_{i-1}}(x^{i-1}) \\ &= \sum_{i=1}^n \log \int_{\mathcal{X}} P_{\hat{\mu}_i^x}(x^{i-1}, x) dx - \log P_{\hat{\mu}_{i-1}}(x^{i-1}) \\ &= \sum_{i=1}^n \log \int_{\mathcal{X}} \exp \left\{ \log P_{\hat{\mu}_i^x}(x^{i-1}, x) - \log P_{\hat{\mu}_{i-1}}(x^{i-1}) \right\} dx.\end{aligned}\quad (15)$$

We will bound each term of the sum separately. To this end, we need the following lemma, which states that the integral is negligibly small outside a ball around  $\hat{\mu}_{n-1}$ , slowly growing with  $n$ .

**Lemma 6** *Fix an arbitrary  $\alpha > 0$  and let  $\mathcal{B}_n(\alpha) = \{x \in \mathcal{X}: \|x - \hat{\mu}_{n-1}\| \leq n^\alpha\}$ . Then, for every  $\gamma > 0$ , there exists a constant  $C_\gamma > 0$  such that for all large  $n$ ,*

$$\int_{\mathcal{X} \setminus \mathcal{B}_n(\alpha)} \exp \left\{ \log P_{\hat{\mu}_n^x}(x^{n-1}, x) - \log P_{\hat{\mu}_{n-1}}(x^{n-1}) \right\} dx \leq C_\gamma n^{-\gamma}. \quad (16)$$

**Proof** Let us denote the left hand side of (16) as  $A_n$ . Rewriting the integral, we get:

$$A_n = \int_{\mathcal{X} \setminus \mathcal{B}_n(\alpha)} P_{\hat{\mu}_n^x}(x) \exp \left\{ \log P_{\hat{\mu}_n^x}(x^{n-1}) - \log P_{\hat{\mu}_{n-1}}(x^{n-1}) \right\} dx \leq \int_{\mathcal{X} \setminus \mathcal{B}_n(\alpha)} P_{\hat{\mu}_n^x}(x) dx,$$

because from the definition of  $\hat{\mu}_{n-1}$ ,  $P_{\hat{\mu}_{n-1}}(x^{n-1}) \geq P_{\hat{\mu}_n^x}(x^{n-1})$ .

We will use the fact that exponential families have all central moments finite (Barndorff-Nielsen, 1978). This means, that  $\int_{\mathcal{X}} \|x - \mu\|^\beta P_\mu(x) dx < \infty$ , for all  $\beta = 1, 2, \dots$ , so if  $x \rightarrow \infty$ ,  $P_\mu(x)$  must converge to 0 faster than any monomial  $\|x - \mu\|^{-\beta}$  for the integral to be finite. This implies that for any  $\beta > 0$ ,  $P_\mu(x) \leq C_{\mu,\beta} \|x - \mu\|^{-\beta}$  for some  $C_{\mu,\beta} < \infty$ . Moreover,  $C_\beta := \sup_{\mu \in \Xi_0} C_{\mu,\beta}$  is finite. Otherwise, we could form a monotonic sequence  $\mu_i$  with  $C_{\mu_i,\beta} > C_{\mu_{i-1},\beta} + 1$ ,  $i = 1, 2, \dots$ ; due to compactness of  $\Xi_0$ , this sequence has a subsequence converging to  $\mu^* \in \Xi_0$  and due to monotonicity,  $C_{\mu^*,\beta}$  cannot be finite, which is a contradiction. Therefore, we can write:

$$A_n \leq \int_{\mathcal{X} \setminus \mathcal{B}_n(\alpha)} P_{\hat{\mu}_n^x}(x) dx \leq C_\beta \int_{\mathcal{X} \setminus \mathcal{B}_n(\alpha)} \|x - \hat{\mu}_n^x\|^{-\beta}.$$

From the triangle inequality,  $\|x - \hat{\mu}_n^x\| \geq \|x - \hat{\mu}_{n-1}\| - \|\hat{\mu}_n^x - \hat{\mu}_{n-1}\|$ . Since the former is at least  $n^\alpha$  for  $x \notin \mathcal{B}(\alpha)$ , while the latter is bounded, for  $n$  large enough,  $\|x - \hat{\mu}_n^x\| \geq \frac{1}{2}\|x - \hat{\mu}_{n-1}\|$  and therefore:

$$A_n \leq C_\beta 2^\beta \int_{\|x\| \geq n^\alpha} \|x\|^{-\beta} dx \leq C' n^{\alpha(k-\beta)}.$$

Setting  $C_\gamma = C'$  and  $\gamma = \alpha(\beta - k)$  finishes the proof.  $\blacksquare$

We are now ready to prove the main theorem:

**Theorem 7** *Assume the setting of Section 2.1; in particular, let the ML distribution be defined as in (2) where  $\Theta_0$  (and hence  $\Xi_0$ ) is compact. Let  $P$  be the SNML strategy. Then,*

$$\mathcal{R}(P; x^n) \leq \frac{k}{2} \log n + O(1),$$

where the constant in  $O(1)$  depends on  $\Xi_0$ , but does not depend on the data sequence  $x^n$ .

**Proof** Using the simple fact that  $\log(a+b) \leq \max\{0, \log(a)\} + b$  for  $a, b \geq 0$ , and applying Lemma 6, we can bound:

$$\log \int_{\mathcal{X}} e^{\log P_{\hat{\mu}_n^x}(x^{n-1}, x) - \log P_{\hat{\mu}_{n-1}}(x^{n-1})} dx \leq \max \left\{ 0, \log \int_{\mathcal{B}_n(\alpha)} e^{\log P_{\hat{\mu}_n^x}(x^{n-1}, x) - \log P_{\hat{\mu}_{n-1}}(x^{n-1})} dx \right\} + C_\gamma n^{-\gamma}. \quad (17)$$

Let us denote the integral over  $\mathcal{B}_n(\alpha)$  on the right hand side of (17) as  $S_n$ . We have:

$$\begin{aligned} S_n &= \log \int_{\mathcal{B}_n(\alpha)} \exp \left\{ \log P_{\hat{\mu}_n^x}(x^{n-1}, x) - \log P_{\hat{\mu}_{n-1}}(x^{n-1}, x) + \log P_{\hat{\mu}_{n-1}}(x) \right\} dx \\ &= \log \int_{\mathcal{B}_n(\alpha)} P_{\hat{\mu}_{n-1}}(x) \exp \left\{ \log P_{\hat{\mu}_n^x}(x^{n-1}, x) - \log P_{\hat{\mu}_{n-1}}(x^{n-1}, x) \right\} dx \\ &\leq \log \int_{\mathcal{B}_n(\alpha)} P_{\hat{\mu}_{n-1}}(x) \exp \left\{ nD(\bar{y}_n^x \| \hat{\mu}_{n-1}) - nD(\bar{y}_n^x \| \tilde{\mu}_n^x) \right\} dx \quad (\text{Proposition 1}) \\ &\leq \log \int_{\mathcal{B}_n(\alpha)} P_{\hat{\mu}_{n-1}}(x) \exp \left\{ nD(\bar{y}_n^x \| \hat{\mu}_{n-1}) \right\} dx, \end{aligned}$$

where  $\bar{y}_n^x = \frac{(n-1)\hat{\mu}_{n-1} + x}{n}$  and  $\tilde{\mu}_n^x = \arg \min_{\mu \in \Xi_0} D(\bar{y}_n^x \| \mu)$ .

Let  $\Xi_1 \subset \Xi$  be a compact set such that  $\Xi_0 \subset \Xi_1$  and:  $\inf_{\mu \in \Xi \setminus \Xi_1, \mu' \in \Xi_0} \|\mu - \mu'\| \geq \delta$  for some  $\delta > 0$ , as in the proof of Proposition 3. Let us choose  $\alpha < 1/4$  in the definition of  $\mathcal{B}_n(\alpha)$ . Then, for  $n$  large enough,

$$\|\bar{y}_n^x - \hat{\mu}_{n-1}\| = \frac{\|x - \hat{\mu}_{n-1}\|}{n} \leq n^{\alpha-1} < \delta$$

for sufficiently large  $n$ , which means that  $\bar{y}_n^x \in \Xi_1$ . Using Proposition 2 with  $\mu = \bar{y}_n^x$  and  $\mu' = \hat{\mu}_{n-1}$  for compact set  $\Xi_1$  and  $x \in \mathcal{B}_n(\alpha)$ , we get:

$$\begin{aligned} nD(\bar{y}_n^x \| \hat{\mu}_{n-1}) &\leq \frac{n}{2} (\bar{y}_n^x - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(\bar{y}_n^x - \hat{\mu}_{n-1}) + nC \|\bar{y}_n^x - \hat{\mu}_{n-1}\|^3 \\ &= \frac{1}{2n} (x - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(x - \hat{\mu}_{n-1}) + \frac{C}{n^2} \|x - \hat{\mu}_{n-1}\|^3 \\ &\leq \frac{1}{2n} (x - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(x - \hat{\mu}_{n-1}) + C' n^{3\alpha-2} \end{aligned}$$

Let  $B(x)$  be a function, equal to the right hand side of the above inequality, if  $x \in \mathcal{B}_n(\alpha)$ , and  $B(x) = 0$  for  $x \notin \mathcal{B}_n(\alpha)$ . Since  $I(\mu)$  is continuous on  $\Xi$ , and  $\Xi_0$  is compact,  $\sup_{\mu \in \Xi_0} \|I(\mu)\| = I_{\Xi_0} < \infty$ , and thus  $B(x)$  is bounded by:

$$B(x) \leq \frac{1}{2n} I_{\Xi_0} n^{2\alpha} + C' n^{3\alpha-2} = C'' n^{2\alpha-1},$$

(because  $\alpha < 1/4$ ). Note that:

$$\log \int_{\mathcal{B}_n(\alpha)} P_{\hat{\mu}_{n-1}}(x) e^{B(x)} dx \leq \log \int_{\mathcal{X}} P_{\hat{\mu}_{n-1}}(x) e^{B(x)} dx = \log \mathbb{E} [e^{B(x)}].$$

We can therefore use Hoeffding's lemma,  $\log \mathbb{E} [e^{B(x)}] \leq \mathbb{E} [B(x)] + \frac{(C'')^2}{8} n^{2(2\alpha-1)}$  (Cesa-Bianchi and Lugosi, 2006), and bound  $\mathbb{E} [B(x)]$ :

$$\begin{aligned} \mathbb{E} [B(x)] &\leq \frac{1}{2n} \mathbb{E} [(x - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1})(x - \hat{\mu}_{n-1})] + C' n^{3\alpha-2} \\ &= \frac{1}{2n} \mathbb{E} [\text{Tr} \{ (x - \hat{\mu}_{n-1})(x - \hat{\mu}_{n-1})^T I(\hat{\mu}_{n-1}) \}] + C' n^{3\alpha-2} \\ &= \frac{1}{2n} \text{Tr} \{ \text{Cov}_{\hat{\mu}_{n-1}}(x) I(\hat{\mu}_{n-1}) \} + C' n^{3\alpha-2} = \frac{k}{2n} + C' n^{3\alpha-2}. \end{aligned}$$

Thus, we finally get:

$$\log \int_{\mathcal{X}} e^{\log P_{\hat{\mu}_n^x}(x^{n-1}, x) - \log P_{\hat{\mu}_{n-1}}(x^{n-1})} dx \leq \frac{k}{2n} + C' n^{3\alpha-2} + \frac{(C'')^2}{8} n^{2(2\alpha-1)} + C_\gamma n^{-\gamma} = \frac{k}{2n} + o(n^{-1}),$$

for sufficiently large  $\gamma$ ,  $\alpha < 1/4$ , for all large  $n$ . Since:

$$\log \int_{\mathcal{X}} P_{\hat{\mu}_i^x}(x) \exp \{ \log P_{\hat{\mu}_i^x}(x^{i-1}) - \log P_{\hat{\mu}_{i-1}}(x^{i-1}) \} dx \leq \log \int_{\mathcal{X}} P_{\hat{\mu}_i^x}(x) dx \leq \log \int_{\mathcal{X}} P_{\hat{\mu}_1^x}(x) dx,$$

which is the minimax (NML) regret for  $n = 1$  and thus finite (Section 2.2), the terms in (15) are finite and well-controlled for small  $n$ . Thus we conclude that  $\mathcal{R}(P; x^n) \leq \frac{k}{2} \log n + O(1)$ .

■

## 5. Examples

In this section, we calculate and analyze SNML for few examples of commonly used exponential families.

### 5.1. Bernoulli Distribution

The SNML for Bernoulli was first considered by Takimoto and Warmuth (2000) as the *last-step minimax* algorithm. The simplest derivation is in the mean value parametrization, in which the Bernoulli distribution looks like:

$$P_\mu(x) = \mu^x (1 - \mu)^{1-x},$$

where  $x \in \{0, 1\}$  and  $\Xi = (0, 1)$ ; note that we need to exclude from  $\Xi$  two extreme points 0 and 1 to be consistent with the assumptions made in this paper. We set  $\Xi_0 = [\epsilon, 1 - \epsilon]$ . The ML estimator  $\hat{\mu}_n$  to relative  $\Xi_0$  equals:

$$\hat{\mu}_n = \min\{1 - \epsilon, \max\{\epsilon, \bar{x}_n\}\},$$

i.e.  $\hat{\mu}_n$  is  $\bar{x}_n$  truncated to the range  $[\epsilon, 1 - \epsilon]$ . Then, the SNML strategy can easily be computed from:

$$P_{\text{SNML}}(x_n = 1|x^{n-1}) = \frac{(\hat{\mu}_n^1)^{k+1}(1 - \hat{\mu}_n^1)^{n-k}}{(\hat{\mu}_n^1)^{k+1}(1 - \hat{\mu}_n^1)^{n-k} + (\hat{\mu}_n^0)^k(1 - \hat{\mu}_n^0)^{n-k+1}},$$

where  $k = (n - 1)\bar{x}_{n-1}$  is the number of ones in the past, and  $\hat{\mu}_n^x$  for  $x \in \{0, 1\}$  is defined as usual. One can also show that even for the case  $\Xi = [0, 1]$  and  $\epsilon = 0$ , although not covered by our theorem, the regret of the strategy is bounded by  $\frac{1}{2}\log(n+1) + \frac{1}{2}$  (Takimoto and Warmuth, 2000), and is superior even to the celebrated Krichevsky-Trofimov estimator (Cesa-Bianchi and Lugosi, 2006).

## 5.2. Exponential Distribution

The distribution has the form:

$$P_\theta(x) = \frac{1}{\mu}e^{-x/\mu},$$

with  $\Xi = (0, \infty)$ . The strategy becomes particularly simple if we take  $\Xi_0 = \Xi$  (although this case is not covered by Theorem 7). Like in the Bernoulli case, the ML estimator is equal to the truncation of  $\bar{x}_n$  into the range  $\Xi_0$ , so that for  $\Xi_0 = (0, \infty)$ ,  $\hat{\mu}_n = \bar{x}_n$ . Thus:

$$P_{\text{SNML}}(x_n|x^{n-1}) \propto \sup_{\mu \in \Xi_0} P_\mu(x^n) = \bar{x}_n^{-n} e^{-\frac{\sum_{i=1}^n x_i}{\bar{x}_n}} = \frac{e^{-n} n^n}{((n-1)\bar{x}_{n-1} + x_n)^n},$$

which, after proper normalization, becomes:

$$P_{\text{SNML}}(x_n|x^{n-1}) = \frac{(n-1)^n \bar{x}_{n-1}^{n-1}}{((n-1)\bar{x}_{n-1} + x_n)^n}.$$

One can directly show that the per-round regret increase  $\mathcal{R}(P_{\text{SNML}}; x^n) - \mathcal{R}(P_{\text{SNML}}; x^{n-1})$  equals  $n \log \frac{n}{n-1} - 1$ , which is upper bounded by  $\frac{1}{2(n-1)}$  except the first iteration. Indeed, choosing  $\Xi_0 = \Xi$  implies infinite regret in the first trial. That being said, in the rest of the game such a choice of  $\Xi_0$  does not seem to be harmful anymore.

## 5.3. Gaussian Distribution, Fixed Variance

The  $k$ -dimensional Gaussian distribution  $N(x|\mu, \Sigma)$  reads:

$$P_\mu(x) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \right\},$$

and without loss of generality we can assume  $\Sigma = I$ , the identity matrix (we can always rotate the coordinate system so that it matches with the principal axes of  $\Sigma$ , and rescale it

by the inverses of the eigenvalues). We set  $\Xi = \mathbb{R}^k$  and we will use  $\Xi_0 = \{\mu: \|\mu\| \leq R\}$ . The ML estimator  $\hat{\mu}_n$  relative to  $\Xi_0$  equals:

$$\hat{\mu}_n = \begin{cases} \bar{x}_n & \text{if } \|\bar{x}_n\| \leq R, \\ \frac{R}{\|\bar{x}_n\|} \bar{x}_n & \text{if } \|\bar{x}_n\| > R. \end{cases}$$

The derivation of the SNML strategy simplifies a lot if we choose  $R \rightarrow \infty$ . Then,  $\hat{\mu}_n = \bar{x}_n$  and:

$$P_{\text{SNML}}(x_n|x^{n-1}) \propto e^{-\frac{1}{2} \sum_{i=1}^n \|x_i - \bar{x}_n\|^2}.$$

A bit of algebra reveals that:

$$\sum_{i=1}^n \|x_i - \bar{x}_n\|^2 = \sum_{i=1}^{n-1} \|x_i - \bar{x}_{n-1}\|^2 + \frac{n-1}{n} \|x_n - \bar{x}_{n-1}\|^2,$$

so that  $P_{\text{SNML}}(x_n|x^{n-1}) \propto e^{-\frac{1}{2} \frac{n-1}{n} \|x_n - \bar{x}_{n-1}\|^2}$ , which means that SNML predicts with  $N(x|\bar{x}_{n-1}, \frac{n}{n-1}I)$ . Note, that although SNML predict with a Gaussian distribution, it is *not* a plug-in type strategy, as the predictive distribution is *outside* the model (the variance is not equal to  $I$ ). Although Theorem 7 does not apply for  $R = \infty$ , one can directly show that the per-round regret increase equals  $\frac{k}{2} \log \frac{n}{n-1}$ , which gives the regret  $\frac{k}{2} \log n$  if we do *not* count the regret in the first iteration (which is, again, infinite).

#### 5.4. Gaussian Distribution, Unknown Mean and Variance

As a final example, consider the family of one-dimensional Gaussian distributions with unknown mean  $\mu$  and variance  $\sigma^2$ . Writing down the distribution in the natural parameterization:

$$P_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \frac{\mu^2}{2\sigma^2} + \log \sigma \right\},$$

shows that the exponential family is two-dimensional with sufficient statistic  $(x, x^2)$ . Similarly as in the previous cases, setting  $\Xi_0 = \Xi$  significantly simplifies all the derivations. The ML estimator becomes  $\hat{\mu}_n = \bar{x}_n$  and  $\hat{\sigma}_n^2 = n^{-1} \sum (x_i - \bar{x}_n)^2$ . The SNML strategy predicts with:

$$P_{\text{SNML}}(x_n|x^{n-1}) \propto (2\pi\hat{\sigma}_n^2)^{-n/2} \exp \left\{ -\frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{2\hat{\sigma}_n^2} \right\} = \frac{e^{-\frac{n}{2} n^{n/2}}}{(2\pi(n-1))^{n/2}} \frac{1}{(\hat{\sigma}_{n-1}^2 + \frac{1}{n} (x_n - \bar{x}_{n-1})^2)^{n/2}},$$

which after normalization gives:

$$P_{\text{SNML}}(x_n|x^{n-1}) = \frac{\Gamma(n/2)}{\Gamma(1/2)\Gamma((n-1)/2)} (n\hat{\sigma}_{n-1}^2)^{-1/2} \left( 1 + \frac{(x_n - \bar{x}_{n-1})^2}{n\hat{\sigma}_{n-1}^2} \right)^{-n/2}.$$

One can directly show that the per-round regret increase

$$\mathcal{R}(P_{\text{SNML}}; x^n) - \mathcal{R}(P_{\text{SNML}}; x^{n-1}) = \frac{n+1}{2} \log n - \frac{n}{2} \log(n-1) - \frac{1}{2} \log 2e + \log \frac{\Gamma((n-1)/2)}{\Gamma(n/2)},$$

and grows as  $\frac{1}{n} + O(n^{-2})$  (except the first iteration), which is optimal for  $k = 2$ .

### 5.5. Open Problem

In the paper, the SNML strategy always plays against the exponential family  $\mathcal{M}$  with parameter space restricted to a compact subset  $\Xi_0$ . As shown in Section 2, this is necessary, otherwise the strategy would suffer infinite regret already in the first iteration. However, as we saw in all the above examples, starting from the second iteration, choosing  $\Xi_0 = \Xi$  does not hurt anymore, so that the only place in which restricting the power of the reference strategies to  $\Xi_0$  matters, is the beginning of the game. A similar phenomenon was already found in the MDL community (Grünwald, 2007), in which it was noticed that some models with infinite complexity (i.e. with infinite worst-case regret) can be made finitely complex when conditioned on the first outcome(s).

Therefore we pose the following problem. Assume that the loss suffered in the first  $n_0 \geq 1$  iterations is not included in the cumulative regret of the predicting strategy. Is it then possible to achieve the worst-case regret  $\frac{k}{2} \log n + O(1)$  without restricting the parameter space of the exponential family?

## 6. SNML and Bayes with Jeffreys' Prior

Given a prior distribution  $\pi(\mu)$  over  $\Xi$ , the *Bayes* prediction strategy  $P_{\text{BAYES}}$  is defined as:

$$P_{\text{BAYES}}(x^n) = \int_{\Xi} P_{\mu}(x^n) \pi(\mu) d\mu, \quad (18)$$

a  $\pi$ -mixture of distributions from  $\Xi$ . Let  $\Xi_0$  be a compact subset of  $\Xi$ . It is known (Grünwald, 2007) that under the assumption, that the sequence  $x_1, x_2, \dots$  satisfies  $\bar{x}_n \in \Xi_0$  for all large  $n$ , the Bayes strategy achieves asymptotically optimal regret  $\mathcal{R}(P_{\text{BAYES}}, x^n) = \frac{k}{2} \log n + O(1)$ . Moreover, using the *Jeffreys' prior*  $\pi(\mu) \propto \sqrt{\det I(\mu)}$  minimizes the  $O(1)$  term in the worst case, up to an  $o(1)$  term, so that the Jeffreys' prior is in some sense a minimax prior, achieving asymptotically the same regret as NML.

Let us consider instead the case when  $\Xi_0 = \Xi$ . Although the joint distribution for Bayes with Jeffreys' prior  $P_{\text{JEF}}(x^n)$  is often undefined in this case (the prior cannot be normalized), the conditionals  $P_{\text{JEF}}(x_n | x^{n-1})$  for all  $n \geq m$ , for some  $m$ , can still be properly defined (Grünwald, 2007). This is also the case of SNML (examples in Section 5 show that SNML is properly defined for  $n \geq 2$ ). Moreover, the same story will apply to NML (minimax) strategy if, instead of using definition (6), we will define NML through the conditionals:

$$P_{\text{NML}}(x_i | x^{i-1}) = \frac{P_{\text{NML}}(x^i)}{P_{\text{NML}}(x^{i-1})} = \frac{\int_{\mathcal{X}^{n-i}} \sup_{\mu \in \Xi} P_{\mu}(x^i, z^{n-i}) dz^{n-i}}{\int_{\mathcal{X}^{n-i+1}} \sup_{\mu \in \Xi} P_{\mu}(x^{i-1}, z^{n-i+1}) dz^{n-i+1}},$$

where  $n$  is the time horizon for NML. Note that the definition coincides with the concept of *conditional NML-2* in (Grünwald, 2007).

Interestingly, in all but one of the examples from Section 5, the SNML strategy and Jeffreys' strategy coincide. Only in the case of Bernoulli, the strategies differ, with an advantage for SNML, whose worst-case regret is smaller by a constant than the one achieved by Jeffreys' strategy<sup>1</sup>. The two open questions we pose are: (1) Under what conditions are

---

1. The Jeffreys' strategy for Bernoulli is the Krichevsky-Trofimov strategy.

SNML and Bayes with Jeffreys' prior the same? (2) If SNML and Jeffreys' differ, is there any general relationship between the worst-case regrets of the two?

To shed some light on the questions above, we prove the following fact, which shows that the equivalence of SNML and Jeffreys' implies that both strategies are equal to the NML strategy. This is what actually happens in three out of four examples shown in Section 5.

**Theorem 8** *Let  $\mathcal{P}$  be an exponential family, such that, starting from some  $m$ , the conditional distributions for SNML and Jeffreys' strategies are properly defined and coincide,  $P_{\text{SNML}}(x_n|x^{n-1}) = P_{\text{JEFF}}(x_n|x^{n-1})$  for all  $x^n$ ,  $n \geq m$ . Then, both strategies are equal to the minimax (NML) strategy  $P_{\text{NML}}(x_n|x^{n-1})$ , for  $n \geq m$ , and the NML strategy does not depend on the time horizon.*

**Proof** Since the strategies might only be defined for  $n \geq m$ , let us focus on the *conditional regret* defined as  $\mathcal{R}(P; x^{m:n}|x^{m-1}) = -\log P(x^{n:m}|x^{m-1}) + \log P_{\hat{\mu}_n}(x^n)$ , where  $x^{m:n} = x_m, x_{m+1}, \dots, x_n$ . From the definition (13), the conditional regret of the SNML strategy does not depend on the last outcome:

$$\mathcal{R}(P_{\text{SNML}}; x^{m:n}|x^{m-1}) = -\log P_{\text{SNML}}(x^{m:n-1}|x^{m-1}) + \log \int_{\mathcal{X}} P_{\hat{\mu}_n}(x^{n-1}, x) dx.$$

Since  $P_{\text{JEF}}(x^{m:n}|x^{m-1}) = P_{\text{SNML}}(x^{m:n}|x^{m-1})$ , Jeffreys' strategy inherits this property. On the other hand, Jeffreys' is a particular case of the Bayes strategy, which is known to be *exchangeable*, i.e.  $P_{\text{BAYES}}(x^{m:n}|x^{m-1})$  does not depend on the order of the outcomes in  $x^{m:n}$ ; this property follows directly from the definition (18). Since the comparator  $P_{\hat{\mu}_n}(x^n)$  does not depend on the order either, the same property holds for the conditional regret. But then, if the conditional regret is invariant under changing the last outcome and under changing the order of the outcomes, it is also invariant under changing all the outcomes in  $x^{m:n}$ . This means that the strategy  $P_{\text{JEF}}$  gives equal conditional regret for all possible data sequences  $x^{m:n}$  (i.e. the strategy is an *equalizer* of conditional regret), which implies  $P_{\text{JEF}}(x^{m:n}|x^{m-1})$  must be equal  $P_{\text{NML}}(x^{m:n}|x^{m-1})$  for all  $x^n$  (Grünwald, 2007). ■

Theorem (8) would also hold if we replace Jeffreys' strategy with a general Bayes strategy. However, due to the known relationship between the Jeffreys' strategy and the minimax regret, we do not expect the conditions of the theorem to be satisfied by Bayes with any other prior than Jeffreys'.

## 7. Conclusions and Further Work

We analyzed the regret of the ML (“follow the leader”) strategy for general exponential families. The lower and upper bounds show that the ML strategy requires boundedness of the data sequence to obtain logarithmic regret, and moreover, the constant in front of the logarithm is suboptimal and can be very large. Those two drawbacks are essentially unavoidable, not only for the ML strategy, but for any plug-in strategy. However, we also showed that both problems are solved by adding the currently predicted outcome to the calculation of the maximum likelihood, followed by normalization, which leads to the SNML strategy. We proved that SNML achieves asymptotically optimal regret. We also noted an interesting relationship to the Bayes strategy with Jeffreys' prior.

In future work, we plan to work on the two open questions posed in the paper: (1) Is it possible to relax the condition that the model is constrained to the compact subset of the parameter space by conditioning the regret on the outcome from the first iteration? (2) When is SNML equal to Bayes with Jeffreys' prior and is there any general relationship between the worst-case regrets of the two?

## References

- K. Azoury and M. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Journal of Machine Learning*, 43(3):211–246, 2001.
- O.E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, Chichester, UK, 1978.
- A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.
- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- N. Cesa-Bianchi and G. Lugosi. Worst-case bounds for the logarithmic loss of predictors. *Journal of Machine Learning*, 43(3):247–264, 2001.
- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.
- Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley, 1991.
- Sanjoy Dasgupta and Daniel Hsu. On-line estimation with the multivariate gaussian distribution. In *Conference on Learning Theory (COLT '07)*, 2007.
- A.P. Dawid. Present position and potential developments: Some personal views, statistical theory, the prequential approach. *J. Royal Stat.Soc., Ser. A*, 147(2):278–292, 1984.
- Y. Freund. Predicting a binary sequence almost as well as the optimal biased coin. In *Computational Learning Theory (COLT' 96)*, pages 89–98, 1996.
- P. Grünwald and W. Kotłowski. Prequential plug-in codes that achieve optimal redundancy rates even if the model is wrong. In *The IEEE International Symposium on Information Theory (ISIT '10)*, 2010.
- P. D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.
- P. D. Grünwald and S. de Rooij. Asymptotic log-loss of prequential maximum likelihood codes. In *Conference on Learning Theory (COLT 2005)*, pages 652–667, 2005.
- E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2–3):169–192, 2007.
- W. Kotłowski, P. Grünwald, and S. de Rooij. Following the flattened leader. In *Conference on Learning Theory (COLT '10)*, 2010.

- M. Raginsky, R. F. Marcia, S. Jorge, and R. Willett. Sequential probability assignment via online convex programming using exponential families. In *IEEE International Symposium on Information Theory*, 2009.
- J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Information Theory*, IT-42(1):40–47, 1996.
- J. Rissanen and T. Roos. Conditional NML universal models. In *Information Theory and Applications Workshop (ITA-07)*, pages 337–341, 2007.
- T. Roos and J. Rissanen. On sequentially normalized maximum likelihood models. In *Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08)*, 2008.
- Y. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):175–186, 1987.
- E. Takimoto and M. Warmuth. The last-step minimax algorithm. In *Conference on Algorithmic Learning Theory (ALT '00)*, 2000.

KOTŁOWSKI GRÜNWALD

# A New Algorithm for Compressed Counting with Applications in Shannon Entropy Estimation in Dynamic Data

**Ping Li**

*Department of Statistical Science, Cornell University, Ithaca, NY 14853*

PINGLI@CORNELL.EDU

**Cun-Hui Zhang**

*Department of Statistics and Biostatistics, Rutgers University, New Brunswick, NJ 08901*

CZHANG@STAT.RUTGERS.EDU

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

Efficient estimation of the moments and Shannon entropy of data streams is an important task in modern machine learning and data mining. To estimate the Shannon entropy, it suffices to accurately estimate the  $\alpha$ -th moment with  $\Delta = |1 - \alpha| \approx 0$ . To guarantee that the error of estimated Shannon entropy is within a  $\nu$ -additive factor, the method of *symmetric stable random projections* requires  $O(\frac{1}{\nu^2 \Delta^2})$  samples, which is extremely expensive. The first paper (Li, 2009a) in *Compressed Counting (CC)*, based on *skewed-stable random projections*, supplies a substantial improvement by reducing the sample complexity to  $O(\frac{1}{\nu^2 \Delta})$ , which is still expensive. The followup work (Li, 2009b) provides a practical algorithm, which is however difficult to analyze theoretically.

In this paper, we propose a new accurate algorithm for Compressed Counting, whose sample complexity is only  $O(\frac{1}{\nu^2})$  for  $\nu$ -additive Shannon entropy estimation. The constant factor for this bound is merely about 6. In addition, we prove that our algorithm achieves an upper bound of the Fisher information and in fact it is close to 100% statistically optimal. An empirical study is conducted to verify the accuracy of our algorithm.

**Keywords:** Data Streams, Entropy Estimation, Maximally-Skewed Stable Random Projections

## 1. Introduction

The problem of “scaling up for high dimensional data and high speed data streams” is among the “10 challenging problems in data mining research” (Yang and Wu, 2006). This paper is devoted to estimating entropy of data streams. Mining data streams in (e.g.,) 100 TB scale databases has become an important area of research, e.g., (Henzinger et al., 1999; Domeniconi and Gunopulos, 2001; Aggarwal et al., 2004; Muthukrishnan, 2005), as the Web and network data can easily reach that scale (Yang and Wu, 2006).

Consider the *Turnstile* stream model (Muthukrishnan, 2005). The input stream  $a_t = (i_t, I_t)$ ,  $i_t \in [1, D]$  arriving sequentially describes the underlying signal  $A$ , meaning

$$A_t[i_t] = A_{t-1}[i_t] + I_t, \quad (1)$$

where the increment  $I_t$  can be either positive (insertion) or negative (deletion). For example, in network measurements,  $I_t$  can be the increment of the packet size at the location numbered by  $i_t$ .

In the model (1), restricting  $A_t[i] \geq 0$  results in the *strict-Turnstile* model, which suffices for describing almost all natural phenomena (Muthukrishnan, 2005). This paper focuses on efficient algorithms for estimating  $\alpha$ -th frequency moment of data streams

$$F_{(\alpha)} = \sum_{i=1}^D A_t[i]^\alpha. \quad (2)$$

We are interested in the case of  $\alpha \rightarrow 1$ , which is crucial for the estimation of *Shannon entropy*. Note that the first moment (i.e., the sum)  $F_{(1)} = \sum_{s=0}^t I_s$  can be computed using a single counter.

### 1.1. Entropy, Moments, and Estimation Complexity

A widely useful summary statistic is the *Shannon entropy*

$$H = - \sum_{i=1}^D \frac{A_t[i]}{F_{(1)}} \log \frac{A_t[i]}{F_{(1)}}. \quad (3)$$

There are various generalizations of the Shannon entropy. The Rényi entropy (Rényi, 1961), denoted by  $H_\alpha$ , and the Tsallis entropy (Havrda and Charvát, 1967; Tsallis, 1988), denoted by  $T_\alpha$ , are

$$H_\alpha = \frac{1}{1-\alpha} \log \frac{F_{(\alpha)}}{F_{(1)}^\alpha}, \quad T_\alpha = \frac{1}{1-\alpha} \left( \frac{F_{(\alpha)}}{F_{(1)}^\alpha} - 1 \right). \quad (4)$$

As  $\alpha \rightarrow 1$ , both Rényi entropy and Tsallis entropy converge to Shannon entropy:

$$\lim_{\alpha \rightarrow 1} H_\alpha = \lim_{\alpha \rightarrow 1} T_\alpha = H. \quad (5)$$

Thus, both Rényi entropy and Tsallis entropy can be computed from the  $\alpha$ -th frequency moment; and one can approximate Shannon entropy using  $\alpha \approx 1$ . While this fact is well-known, it appears that (Zhao et al., 2007) is the first study that applied (5) to Shannon entropy estimation in data streams. Later (Harvey et al., 2008b,a) proposed criteria on theoretically (and conservatively) how close to 1 the  $\alpha$  needs to be. One can numerically verify  $\Delta = |1 - \alpha| < 10^{-7}$  in (Harvey et al., 2008b) or  $\Delta < 10^{-5}$  in (Harvey et al., 2008a) are very likely.<sup>1</sup>

The difficulty in Shannon entropy estimation is reflected by the estimation variance. By the definitions of the Rényi and Tsallis entropies, we need estimators of  $F_{(\alpha)}$  with variances proportional to  $O(\Delta^2)$  in order to cancel the term  $\frac{1}{(1-\alpha)^2} = \frac{1}{\Delta^2}$  (otherwise the sample size must be proportional to  $\frac{1}{\Delta^2}$ ). In other words, the estimators of  $F_{(\alpha)}$  must be extremely accurate.

---

1. In (Harvey et al., 2008b),  $\Delta = \frac{c}{16 \log(1/c)}$ ,  $c = \frac{\nu}{4 \log(D) \log(m)}$ , where  $m$  is the number of streaming updates. If we let  $D = 2^{64}$ ,  $m = 2^{64}$ ,  $\nu = 0.1$ , then  $\Delta \approx 7 \times 10^{-8}$ . If we let  $m = 10^6$ ,  $\nu = 0.1$ , then  $\Delta \approx 2.5 \times 10^{-7}$ . Harvey et al. (2008a) provides some improvements, to allow slightly larger  $\Delta$ , which is still extremely small.

## 1.2. Some Applications of Shannon Entropy

**Real-Time Network Anomaly Detection** Network traffic is a typical example of high-rate data streams. An effective measurement of network traffic in real-time is crucial for anomaly detection and network diagnosis; and one such measurement metric is the Shannon entropy (Feinstein et al., 2003; Lakhina et al., 2005; Xu et al., 2005; Brauckhoff et al., 2006; Lall et al., 2006; Zhao et al., 2007). The *Turnstile* data stream model (1) is naturally suitable for describing network traffic, especially when the goal is to characterize the statistical distribution of the traffic. In its empirical form, a statistical distribution is described by histograms,  $A_t[i]$ ,  $i = 1$  to  $D$ . It is possible that  $D = 2^{64}$  (or larger) if one is interested in measuring the traffic streams of all unique sources or destinations.

The Distributed Denial of Service (**DDoS**) attack, as a representative example of network anomalies, attempts to make computers unavailable to intended users, either by forcing users to reset the computers or by exhausting the resources of service-hosting sites. Since a DDoS attack often changes the statistical distribution of network traffic, a common practice to detect such an attack is to monitor the network traffic using certain summary statistics. As the Shannon entropy is well-suited for characterizing a distribution, a popular detection method is to measure the time-history of entropy and alarm anomalies when the entropy becomes abnormal (Feinstein et al., 2003; Lall et al., 2006).

Entropy measurements do not have to be “perfect” for detecting attacks. It is, however crucial that the algorithm should be computationally efficient at low memory cost, because the traffic data generated by large high-speed networks are enormous and transient. Algorithms should be real-time and one-pass, as the traffic data are unlikely to be stored permanently. Many algorithms have been proposed for “sampling” the traffic streaming data for estimating entropy (Lall et al., 2006; Zhao et al., 2007; Bhuvanagiri and Ganguly, 2006; Guha et al., 2006; Chakrabarti et al., 2006, 2007; Harvey et al., 2008b,a; Zhao et al., 2010).

**Entropy of Query Logs in Web Search** Mei and Church (2008) proposed to estimate the Shannon entropy of some commercial search logs, to help answer some basic problems in Web search, such as, *how big is the web?* The search logs can be viewed as data streams, and Mei and Church (2008) analyzed several “snapshots” of a sample of the search logs, which contained 10 million  $\langle \text{Query}, \text{URL}, \text{IP} \rangle$  triples; each triple corresponded to a click from a particular IP address on a particular URL for a particular query. (Mei and Church, 2008) drew their important conclusions on this (hopefully) representative sample. Alternatively, one could apply new data stream algorithms on the entire history of the search logs.

**Entropy in Neural Computations** A workshop in NIPS’03 was devoted to entropy estimation ([www.menem.com/~ilya/pages/NIPS03](http://www.menem.com/~ilya/pages/NIPS03)), owing to the wide-spread use of entropy in neural computations (Paninski, 2003), e.g., for studying the underlying structure of spike trains.

**Graph Estimation** As demonstrated in a recent paper (Gupta et al., 2010), Shannon entropy estimation plays a crucial role in graph estimation and density estimation in high dimensions.

### 1.3. Symmetric Stable Random Projections and Prior Work on Compressed Counting

The problem of estimating  $F_{(\alpha)}$  has been heavily studied since the pioneering work of (Alon et al., 1996). For  $0 < \alpha \leq 2$ , the method of *symmetric stable random projections* (Indyk, 2006; Li, 2008; Li and Hastie, 2007) in many applications provides a practical algorithm, with a sample complexity of  $O\left(\frac{1}{\epsilon^2}\right)$  (even for  $\alpha = 1$ ), to estimate  $F_{(\alpha)}$  within a  $1 \pm \epsilon$  multiplicative factor.

*Compressed Counting (CC)* (Li, 2009a,b) is a recent breakthrough, which is based on *maximally-skewed stable random projections*. Li (2009a) provided two algorithms, using the *geometric mean* and *harmonic mean*. The *geometric mean* algorithm has the variance proportional to  $O(\Delta)$  in the neighborhood of  $\alpha = 1$ , where  $\Delta = |1 - \alpha|$ . This is the first algorithm that reflected the intuition that, in the neighborhood of  $\alpha = 1$ , the moment estimation algorithms should work better and better as  $\alpha \rightarrow 1$ , in a continuous fashion. The *geometric mean* algorithm for CC, unfortunately, did not provide an adequate mechanism for entropy estimation. It only led to an entropy estimation algorithm with a complexity of  $O\left(\frac{1}{\nu^2 \Delta}\right)$ , but (theoretically)  $\Delta$  has to be extremely small.

Based on the *geometric mean* algorithm of CC (Li, 2009a), Harvey et al. (2008a) developed a complicated multi-point method for Shannon entropy estimation, with a sample (word) complexity of  $O\left(\frac{1}{\nu^2} \log M\right)$  and a very large (like  $10^7$ ) constant<sup>2</sup>, where (e.g.,)  $M = \sum_{i=1}^D |A_t[i]|$  can be viewed as the "universe size." In comparison, our new algorithm is very simple with a sample (word) complexity of  $O\left(\frac{1}{\nu^2}\right)$  and a small constant (about 6), without the  $\log M$  term.

Li (2009b) proposed a practical algorithm based on numerical optimization and achieved very good performance. Since that estimator was complicated and implicit, Li (2009b) did not analyze the sample complexity and statistical efficiency and left them as open problems.

### 1.4. Another Perspective for Entropy Estimation

By the definition of Rényi entropy (4), instead of estimating  $F_{(\alpha)}$ , it suffices to estimate  $J_{(\alpha)}$ , where

$$J_{(\alpha)} = F_{(\alpha)}^{-1/\Delta} = \left[ \sum_{i=1}^D A_t[i]^\alpha \right]^{-1/\Delta}, \quad (6)$$

because, if  $\Delta = 1 - \alpha > 0$ , then

$$H_\alpha = \frac{1}{1 - \alpha} \log \frac{F_{(\alpha)}}{F_{(1)}^\alpha} = \frac{1}{\Delta} \log \frac{J_{(\alpha)}^{-\Delta}}{F_{(1)}^\alpha} = -\log J_{(\alpha)} - \frac{1}{\Delta} \log F_{(1)}^\alpha. \quad (7)$$

Since  $\frac{1}{\Delta} \log F_{(1)}^\alpha$  is computed exactly, we only need to estimate  $J_{(\alpha)}$ . Our new algorithm will provide a  $\nu$ -multiplicative estimate of  $J_{(\alpha)}$  with a complexity of  $O\left(\frac{1}{\nu^2}\right)$ . For small  $\nu$ , this translates into:

---

2. In Sec. 5.2 of (Harvey et al., 2008a), their sample complexity bound is  $O\left(\left[200(z+1)^3\right]^2 \frac{1}{\nu^2} \log M\right)$ , where  $z = \log(1/\nu) + \log \log M$ . The constant  $\left[200(z+1)^3\right]^2$  will exceed  $10^7$ , even just for  $z = 3$ .

1. An  $\epsilon = \nu\Delta$ -multiplicative estimate of  $F_{(\alpha)}$ , with a sample complexity of  $O\left(\frac{1}{\nu^2}\right)$ . For example, denote the estimate of  $J_{(\alpha)}$  by  $\hat{J}_{(\alpha)}$ , then

$$\begin{aligned}\Pr\left(\hat{J}_{(\alpha)} \geq (1 + \nu)J_{(\alpha)}\right) &= \Pr\left(\hat{J}_{(\alpha)}^{-\Delta} \leq (1 + \nu)^{-\Delta}F_{(\alpha)}\right) \\ \Pr\left(\hat{J}_{(\alpha)} \leq (1 - \nu)J_{(\alpha)}\right) &= \Pr\left(\hat{J}_{(\alpha)}^{-\Delta} \geq (1 - \nu)^{-\Delta}F_{(\alpha)}\right).\end{aligned}$$

For small  $\nu$ , we have  $(1 + \nu)^{-\Delta} \approx 1 - \nu\Delta = 1 - \epsilon$  and  $(1 - \nu)^{-\Delta} \approx 1 + \nu\Delta = 1 + \epsilon$ .

2. A  $\nu$ -additive estimate of  $\log J_{(\alpha)}$ , with a sample complexity of  $O\left(\frac{1}{\nu^2}\right)$ . For example

$$\begin{aligned}\Pr\left(\hat{J}_{(\alpha)} \geq (1 + \nu)J_{(\alpha)}\right) &= \Pr\left(\log \hat{J}_{(\alpha)} \geq \log(1 + \nu) + \log J_{(\alpha)}\right) \\ \Pr\left(\hat{J}_{(\alpha)} \leq (1 - \nu)J_{(\alpha)}\right) &= \Pr\left(\log \hat{J}_{(\alpha)} \leq \log(1 - \nu) + \log J_{(\alpha)}\right).\end{aligned}$$

For small  $\nu$ , we have  $\log(1 + \nu) \approx \nu$  and  $\log(1 - \nu) \approx -\nu$ .

## 2. The Proposed Algorithm

Consider the *strict-Turnstile* data stream model (1). Conceptually, we multiply the data stream vector  $A_t \in \mathbb{R}^D$  by a random matrix  $\mathbf{R} \in \mathbb{R}^{D \times k}$ , resulting in a vector  $X = A_t \times \mathbf{R} \in \mathbb{R}^k$  with entries

$$x_j = [A_t \times \mathbf{R}]_j = \sum_{i=1}^D r_{ij} A_t[i], \quad j = 1, 2, \dots, k$$

where  $r_{ij}$ 's are random variables generated as follows:

$$r_{ij} = \frac{\sin(\alpha v_{ij})}{[\sin v_{ij}]^{1/\alpha}} \left[ \frac{\sin(v_{ij}\Delta)}{w_{ij}} \right]^{\frac{\Delta}{\alpha}}, \quad \Delta = 1 - \alpha > 0, \quad (8)$$

where  $v_{ij} \sim Uniform(0, \pi)$  (i.i.d.) and  $w_{ij} \sim Exp(1)$  (i.i.d.), an exponential distribution with mean 1. In data stream computations, the matrix  $\mathbf{R}$  is not materialized. The standard procedure is to (re)generate entries of  $\mathbf{R}$  on-demand (Indyk, 2006). Whenever a stream element  $a_t = (i_t, I_t)$  arrives, one updates entries of  $X$ :

$$x_j \leftarrow x_j + I_t r_{i_t j}, \quad j = 1, 2, \dots, k.$$

The cost of (re)generating (pseudo) random numbers is proportional to the sample size  $k$ . As our work substantially reduces the sample size, it also tremendously reduces the processing time.

Here, our goal is to estimate  $J_{(\alpha)} = F_{(\alpha)}^{-1/\Delta}$  (and hence also  $F_{(\alpha)}$ ). Our proposed algorithm is

$$\hat{J}_{(\alpha)} = \frac{\Delta}{k} \sum_{j=1}^k x_j^{-\alpha/\Delta}, \quad (9)$$

from which one can estimate the Shannon entropy, for example, by the Rényi entropy as

$$\hat{H}_\alpha = -\log \hat{J}_{(\alpha)} - \frac{1}{\Delta} \log F_{(1)}^\alpha \quad (10)$$

The following Lemma provides the moments of  $\hat{J}_{(\alpha)}$ .

**Lemma 1**

$$E \left( \hat{J}_{(\alpha)} \right) = J_{(\alpha)}, \quad (11)$$

$$Var \left( \hat{J}_{(\alpha)} \right) = \frac{J_{(\alpha)}^2}{k} (3 - 2\Delta), \quad (12)$$

$$E \left( \hat{J}_{(\alpha)} - J_{(\alpha)} \right)^3 = \frac{J_{(\alpha)}^3}{k^2} (17 - 21\Delta + 6\Delta^2), \quad (13)$$

$$E \left( \hat{J}_{(\alpha)} - J_{(\alpha)} \right)^4 = 3 \frac{J_{(\alpha)}^4}{k^2} (3 - 2\Delta)^2 + \frac{J_{(\alpha)}^4}{k^3} (142 - 252\Delta + 140\Delta^2 - 24\Delta^3). \quad (14)$$

□

The first two moments immediately imply that the sample complexity of  $\hat{J}_{(\alpha)}$  is  $O \left( \frac{1}{\nu^2} \right)$  for a  $\nu$ -multiplicative approximation of  $J_{(\alpha)}$ . The higher moments in Lemma 1 are also useful for the proof of Lemma 10.

The next Lemma provides the precise tail bounds.

**Lemma 2** 1. *The right tail bound: for  $\nu > 0$ ,*

$$\Pr \left( \hat{J}_{(\alpha)} \geq (1 + \nu) J_{(\alpha)} \right) \leq \exp \left( -k \frac{\nu^2}{G_R} \right) \quad (15)$$

$$\frac{\nu^2}{G_R} = -\log \left( 1 + \sum_{n=1}^{\infty} t_R^n e^n H(n; \Delta) \right) + t_R(1 + \nu) \quad (16)$$

where  $t_R$  is the solution to

$$-\frac{\sum_{n=1}^{\infty} n t_R^{n-1} e^n H(n; \Delta)}{1 + \sum_{n=1}^{\infty} t_R^n e^n H(n; \Delta)} + (1 + \nu) = 0 \quad (17)$$

and

$$H(n; \Delta) = \prod_{i=0}^{n-1} \frac{n - i\Delta}{e(n - i)} \quad (18)$$

2. *The left tail bound: for  $0 < \nu < 1$ ,*

$$\Pr \left( \hat{J}_{(\alpha)} \leq (1 - \nu) J_{(\alpha)} \right) \leq \exp \left( -k \frac{\nu^2}{G_L} \right) \quad (19)$$

$$\frac{\nu^2}{G_L} = -\log \left( 1 + \sum_{n=1}^{\infty} (-t_L)^n e^n H(n; \Delta) \right) - t_L(1 - \nu) \quad (20)$$

where  $t_L$  is the solution to

$$\frac{\sum_{n=1}^{\infty} (-1)^n n(t_L)^{n-1} e^n H(n; \Delta)}{1 + \sum_{n=1}^{\infty} (-t_L)^n e^n H(n; \Delta)} + (1 - \nu) = 0. \quad (21)$$

□

While the expressions for the tail bounds in Lemma 2 appear sophisticated, they are carefully formulated so that they can be accurately evaluated numerically; see Figure 1. The function  $H(n; \Delta)$  in (18) approaches  $e^{-n}$  when  $\Delta \rightarrow 1$ , and it is always upper bounded by  $\frac{1}{\sqrt{2\pi n}}$  even as  $\Delta \rightarrow 0$ , since

$$\prod_{i=0}^{n-1} \frac{n - i\Delta}{n - i} \leq \frac{n^n}{n!} \leq \frac{n^n}{(n-1)!} \leq \frac{e^n}{\sqrt{2\pi n}}$$

according to Stirling's series (Gradshteyn and Ryzhik, 1994, 8.327)

$$\Gamma(n) = (n-1)! = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \left[1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{51840n^3} - \dots\right].$$

Interestingly, we can obtain closed-form expressions when  $\nu \rightarrow 0$ .

**Lemma 3** As  $\nu \rightarrow 0$ , the constants  $G_R$  and  $G_L$  in (16) and (20), respectively, become

$$G_R \rightarrow 6 - 4\Delta, \quad G_L \rightarrow 6 - 4\Delta. \quad (22)$$

□

In addition, when  $\Delta \rightarrow 1-$ , we can actually analytically express the tail bounds in closed-forms.

**Lemma 4** When  $\Delta \rightarrow 1-$ , i.e.,  $\alpha \rightarrow 0+$ ,

$$\frac{\nu^2}{G_R} = -\log(1+\nu) + \nu, \quad \nu > 0 \quad (23)$$

$$\frac{\nu^2}{G_L} = -\log(1-\nu) - \nu, \quad 0 < \nu < 1. \quad (24)$$

□

We summarize the complexity bound of our algorithm in a theorem.

**Theorem 5** The proposed algorithm  $\hat{J}_{(\alpha)}$  in (9) provides a  $\nu$ -multiplicative approximation of  $J_{(\alpha)}$  and a  $\nu$ -additive approximation of the Shannon entropy with a probability at least  $1 - \delta$ , using  $\frac{C}{\nu^2} \log 2/\delta$  samples (words). The constant  $C$  approaches  $6 - 4\Delta$  as  $\nu \rightarrow 0$ . □

### 3. More Intuition and Explanation

The proposed algorithm (9) is based on the idea of *maximally-skewed stable random projections*.

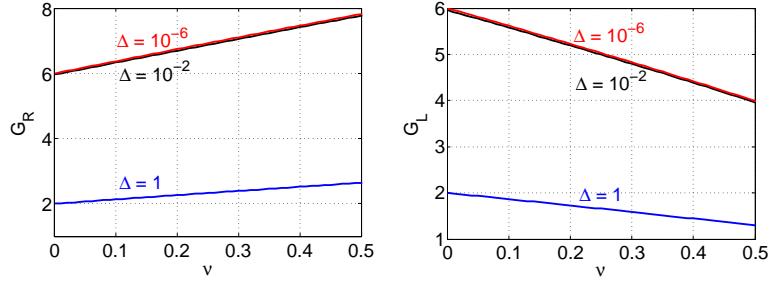


Figure 1: Numerical values of  $G_R$  (left panel) and  $G_L$  (right panel) in the tail bounds (16) and (20), for  $\Delta = 10^{-2}$  and  $\Delta = 10^{-6}$ , together with the closed-form expressions for  $\Delta = 1$  as obtained in Lemma 4. Note that as  $v \rightarrow 0$ , both  $G_R$  and  $G_L$  approach  $6 - 4\Delta$ , as proved in Lemma 3.

### 3.1. Review Maximally-Skewed Stable Random Projections and Estimators

The method for sampling from skewed stable distributions was proposed by Chambers et al. (1976). To sample from  $S(\alpha, \beta = 1, 1)$ , i.e.,  $\alpha$ -stable maximally-skewed ( $\beta = 1$ ) with unit scale, one first generates an exponential random variable with mean 1,  $W \sim Exp(1)$ , and a uniform random variable  $U \sim Uniform(-\frac{\pi}{2}, \frac{\pi}{2})$ . Then the following nonlinear transformation of  $W$  and  $U$  results in the desired random variable:

$$Z' = \frac{\sin(\alpha(U + \rho))}{[\cos U \cos(\rho\alpha)]^{1/\alpha}} \left[ \frac{\cos(U - \alpha(U + \rho))}{W} \right]^{\frac{1-\alpha}{\alpha}} \sim S(\alpha, \beta = 1, 1), \quad (25)$$

where  $\rho = \frac{\pi}{2}$  when  $\alpha < 1$  and  $\rho = \frac{\pi}{2} \frac{2-\alpha}{\alpha}$  when  $\alpha > 1$ . Note that  $\cos(\frac{\pi}{2}\alpha) \rightarrow 0$  as  $\alpha \rightarrow 1$ . For convenience (and to avoid numerical problems), we use

$$Z = Z' \cos^{1/\alpha}(\rho\alpha) \sim S(\alpha, \beta = 1, \cos(\rho\alpha)).$$

It turns out, the random variable  $Z$  with  $\alpha < 1$  has good properties. This study only considers  $\alpha = 1 - \Delta < 1$ , i.e.,  $\rho = \frac{\pi}{2}$ . After simplification, we obtain

$$Z = \frac{\sin(\alpha V)}{[\sin V]^{1/\alpha}} \left[ \frac{\sin(V\Delta)}{W} \right]^{\frac{\Delta}{\alpha}}, \quad (26)$$

where  $V = \frac{\pi}{2} + U \sim Uniform(0, \pi)$ . This explains (8).

Let  $X = A_t \times \mathbf{R}$ , where entries of  $\mathbf{R}$  are i.i.d. samples of  $S(\alpha, \beta = 1, \cos(\frac{\pi}{2}\alpha))$ . Then by properties of stable distributions, the entries of  $X$  are

$$x_j = [A_t \times \mathbf{R}]_j = \sum_{i=1}^D r_{i,j} A_t[i] \sim S\left(\alpha, \beta = 1, \cos\left(\frac{\pi}{2}\alpha\right) F_{(\alpha)}\right),$$

where  $F_{(\alpha)} = \sum_{i=1}^D A_t[i]^\alpha$  as defined in (2). Li (2009a) provided two algorithms using on the *geometric mean* and *harmonic mean* estimators, based on the following basic moment formula.

**Lemma 6** (Li, 2009a). *If  $X \sim S(\alpha, \beta = 1, F_{(\alpha)} \cos(\frac{\alpha\pi}{2}))$ , then  $X > 0$ , and for any  $-\infty < \lambda < \alpha < 1$ ,*

$$E(X^\lambda) = F_{(\alpha)}^{\lambda/\alpha} \frac{\Gamma(1 - \frac{\lambda}{\alpha})}{\Gamma(1 - \lambda)}.$$

□

### 3.1.1. THE GEOMETRIC MEAN ESTIMATOR

Assume  $x_j$ ,  $j = 1$  to  $k$ , are i.i.d. samples from  $S(\alpha, \beta = 1, F_{(\alpha)} \cos(\frac{\alpha\pi}{2}))$ . After simplifying the corresponding expression in (Li, 2009a), we obtain

$$\hat{F}_{(\alpha),gm} = \left[ \frac{\Gamma(1 - \frac{\alpha}{k})}{\Gamma(1 - \frac{1}{k})} \right]^k \prod_{j=1}^k x_j^{\alpha/k}, \quad (27)$$

which is unbiased and has asymptotic variance

$$\text{Var}(\hat{F}_{(\alpha),gm}) = \frac{F_{(\alpha)}^2}{k} \frac{\pi^2}{6} \Delta(1 + \alpha) + O\left(\frac{1}{k^2}\right) \quad (28)$$

As  $\Delta = 1 - \alpha \rightarrow 0$ , the asymptotic variance approaches zero at the rate of only  $O(\Delta)$  (not  $O(\Delta^2)$ ).

### 3.1.2. THE HARMONIC MEAN ESTIMATOR

$$\hat{F}_{(\alpha),hm} = \frac{k^{\frac{1}{\Gamma(1+\alpha)}}}{\sum_{j=1}^k x_j^{-\alpha}} \left( 1 - \frac{1}{k} \left( \frac{2\Gamma^2(1+\alpha)}{\Gamma(1+2\alpha)} - 1 \right) \right), \quad (29)$$

which is asymptotically unbiased and has variance

$$\text{Var}(\hat{F}_{(\alpha),hm}) = \frac{F_{(\alpha)}^2}{k} \left( \frac{2\Gamma^2(1+\alpha)}{\Gamma(1+2\alpha)} - 1 \right) + O\left(\frac{1}{k^2}\right). \quad (30)$$

### 3.2. Limitations of the Geometric Mean and Harmonic Mean Estimators

In order to estimate the Shannon entropy with a guaranteed  $\nu$ -additive accuracy, the variance of the estimator of  $F_{(\alpha)}$  should be  $O(\Delta^2)$ ; or equivalently, the sample complexity should be  $O(\frac{1}{\nu^2})$ .

The geometric mean estimator has variance proportional to only  $O(\Delta)$ ; or equivalently, its complexity is  $O(\frac{1}{\nu^2\Delta})$ , where  $\Delta$  needs to be extremely small (e.g.,  $< 10^{-5}$ ). For the harmonic mean estimator in Li (2009a), the following Lemma says its variance is also proportional to  $O(\Delta)$ .

**Lemma 7** *As  $\Delta = 1 - \alpha \rightarrow 0$ ,*

$$\frac{2\Gamma^2(1+\alpha)}{\Gamma(1+2\alpha)} - 1 = \Delta + \Delta^2 \left( 2 - \frac{\pi^2}{6} \right) + O(\Delta^3). \quad (31)$$

□

In other words, the harmonic mean estimator improves the geometric mean estimator by reducing the variance by a factor of  $\frac{\pi^2}{6} \cdot 2 = 3.29$ . Thus, we must develop significantly better algorithms.

### 3.3. The Distribution Function

This section provides the distribution function of  $Z \sim S(\alpha < 1, \beta = 1, \cos(\frac{\pi}{2}\alpha))$ , which will be needed for better understanding the proposed estimator (9).

**Lemma 8** *Suppose a random variable  $Z \sim S(\alpha < 1, \beta = 1, \cos(\frac{\pi}{2}\alpha))$ . The cumulative distribution function (CDF) is*

$$F_Z(t) = \mathbf{Pr}(Z \leq t) = \frac{1}{\pi} \int_0^\pi \exp\left(-t^{-\alpha/\Delta} g(\theta; \Delta)\right) d\theta. \quad (32)$$

where

$$g(\theta; \Delta) = \frac{[\sin(\alpha\theta)]^{\alpha/\Delta}}{[\sin \theta]^{1/\Delta}} \sin(\theta\Delta), \quad \theta \in (0, \pi)$$

$$g(0+; \Delta) = \lim_{\theta \rightarrow 0+} g(\theta; \Delta) = \Delta \alpha^{\alpha/\Delta}.$$

□

Note that  $g(0+; \Delta) = \Delta \alpha^{\alpha/\Delta} \approx \Delta e^{-1}$  approaches zero as  $\Delta \rightarrow 0$ . Thus, one might be wondering if we replace  $g(\theta; \Delta)$  by  $g(0+; \Delta)$ , the errors may be quite small, as seen in Figure 2.

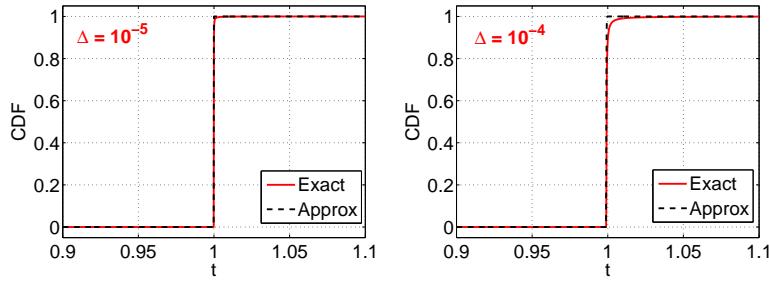


Figure 2: We plot the CDF curves as derived in Lemma 8, for  $\Delta = 10^{-5}$  and  $10^{-4}$ . As  $\Delta \rightarrow 0$ , the exact CDF (solid curves) is very close to the approximate CDF (dashed curves), which we obtain by replacing the exact  $g(\theta; \Delta)$  function in Lemma 8 with the limit  $g(0+; \Delta)$ .

### 3.4. One Intuition Behind the Proposed Algorithm

The difficulty in developing accurate algorithms lies in that  $F_Z$  in (32) has no closed-form expression. From Lemma 8 and Figure 2, it appears that if one replaces the exact  $g(\theta; \Delta)$  with its approximation  $g(0+; \Delta)$ , the error may be small. Thus, we consider a random variable  $Y$  with CDF

$$F_Y(t) = \mathbf{Pr}(y \leq t) = \exp\left(-t^{-\alpha/\Delta} \Delta \alpha^{\alpha/\Delta}\right), \quad t \in [0, \infty). \quad (33)$$

It is indeed a CDF because it is an increasing function of  $t \in [0, \infty)$ ,  $F_Y(0) = 0$ , and  $F_Y(\infty) = 1$ .

Here, we are interested in estimating  $c^\alpha$  from  $k$  i.i.d. samples  $x_j = cy_j$ ,  $j = 1$  to  $k$ . Statistics theory tells us that the maximum likelihood estimator (MLE) achieves the (asymptotic) optimality. Because  $F_Y$  has a closed-form expression, we can compute the MLE exactly.

**Lemma 9** Suppose  $y_j$ ,  $j = 1$  to  $k$ , are i.i.d. samples from a distribution whose CDF is given by (33). Let  $x_j = cy_j$ , where  $c > 0$ . Then the maximum likelihood estimator of  $c^\alpha$  is given by

$$\frac{1}{\Delta^{\Delta \alpha^\alpha}} \left[ \frac{k}{\sum_{j=1}^k x_j^{-\alpha/\Delta}} \right]^\Delta. \quad (34)$$

□

In comparison, our proposed algorithm for estimating  $J_{(\alpha)} = F_{(\alpha)}^{-1/\Delta}$  is defined in (9), which provides an estimator of  $F_{(\alpha)}$ :

$$\hat{F}_{(\alpha)} = \frac{1}{\Delta^\Delta} \left[ \frac{k}{\sum_{j=1}^k x_j^{-\alpha/\Delta}} \right]^\Delta, \quad (35)$$

which is almost identical to (34). Note that, as  $\Delta \rightarrow 0$ , the extra term in (34),  $\alpha^\alpha \rightarrow 1$ , converges much faster than  $\Delta^\Delta \rightarrow 1$ . In other words,  $\alpha^\alpha$  is negligible.

Therefore, we should expect that our proposed estimator (35) is actually very close to the true MLE, even though we can not explicitly derive the MLE. Indeed, in the next section, Lemma 11 says that our algorithm is close to be 100% statistically optimal.

## 4. Additional Technical Results

### 4.1. The Moments of $\hat{F}_{(\alpha)}$

The following Lemma analyzes the mean square error:  $MSE = E \left[ \hat{F}_{(\alpha)} - F_{(\alpha)} \right]^2 = Var \left( \hat{F}_{(\alpha)} \right) + \text{Bias}^2$ .

**Lemma 10** *The estimator  $\hat{F}_{(\alpha)}$  is asymptotically unbiased:*

$$E \left( \hat{F}_{(\alpha)} \right) = F_{(\alpha)} \left( 1 + O \left( \frac{\Delta}{k} \right) \right). \quad (36)$$

The mean square error (MSE) is

$$E \left[ \hat{F}_{(\alpha)} - F_{(\alpha)} \right]^2 = \frac{F_{(\alpha)}^2}{k} \Delta^2 \left( (3 - 2\Delta) + O \left( \frac{1}{k} \right) \right). \quad (37)$$

More precisely

$$0 \leq E \left( \hat{F}_{(\alpha)} - F_{(\alpha)} \right) \leq \frac{\Delta F_{(\alpha)}}{k} e^{2+\Delta} \left( (1 + \Delta)(3 - 2\Delta)/2 + \frac{k}{k - \Delta} \right). \quad (38)$$

and

$$\left| E \left( \frac{\hat{F}_{(\alpha)}}{F_{(\alpha)}} - 1 \right)^2 - \frac{\Delta^2}{k} (3 - 2\Delta) - \frac{\Delta^2}{k^2} C_3^*(\Delta) \right| \leq \frac{\Delta^2 C_4^*(\Delta)}{4k^2} (3 - 2\Delta)^2 + O(\Delta^2 k^{-3} \log k) \quad (39)$$

where  $C_3^*(\Delta) = (1 + \Delta)(17 - 21\Delta + 6\Delta^2) = 17 + O(\Delta)$  and  $C_4^*(\Delta) = e^{4+2\Delta}(11 + 18\Delta + 7\Delta^2) + e^{5+2\Delta}(6 + 11\Delta + 6\Delta^2 + \Delta^3) = 11e^4 + 6e^5 + O(\Delta)$ .  $\square$

## 4.2. Statistical Optimality

Recall we have  $k$  i.i.d. samples  $x_j \sim S(\alpha, \beta = 1, \cos(\frac{\pi}{2}\alpha) F_{(\alpha)})$ . The goal is to estimate  $J_{(\alpha)} = F_{(\alpha)}^{-1/\Delta}$ . The classical theory of the Cramér-Rao lower bound tells us that the variance of the estimator is lower bounded by  $\frac{1}{k} \frac{1}{I(J_{(\alpha)})}$ , where  $I(J_{(\alpha)})$  is the Fisher Information of  $J_{(\alpha)}$ .

A natural question is how much more improvement can we expect, after we have developed the estimator  $\hat{J}_{(\alpha)}$  (9), whose variance is  $\frac{J_{(\alpha)}^2}{k}(3 - 2\Delta)$ ? Lemma 11 provides the answer.

**Lemma 11** *For a distribution  $S(\alpha, \beta = 1, \cos(\frac{\pi}{2}\alpha) F_{(\alpha)})$ , the Fisher Information of  $J_{(\alpha)} = F_{(\alpha)}^{-1/\Delta}$  is given by*

$$I(J_{(\alpha)}) = \frac{1}{J_{(\alpha)}^2} (I_2 - 1), \quad I_2 = \int_0^\infty \frac{\left[ \frac{1}{\pi} \int_0^\pi s g^2 e^{-sg} d\theta \right]^2}{\frac{1}{\pi} \int_0^\pi g e^{-sg} d\theta} ds \quad (40)$$

where  $g = g(\theta; \Delta) = \frac{[\sin(\alpha\theta)]^{\alpha/\Delta}}{[\sin\theta]^{1/\Delta}} \sin(\theta\Delta)$ . The Fisher Information of  $F_{(\alpha)}$  is given by

$$I(F_{(\alpha)}) = \frac{1}{\Delta^2 F_{(\alpha)}^2} (I_2 - 1). \quad (41)$$

Furthermore,  $I_2$  is bounded by  $I_2 \leq 2$ . Therefore the following bounds hold:

$$I(J_{(\alpha)}) \leq \frac{1}{J_{(\alpha)}^2}, \quad I(F_{(\alpha)}) \leq \frac{1}{\Delta^2 F_{(\alpha)}^2}. \quad (42)$$

□

The Fisher information bounds (42) suggest that the optimal estimator (if one can find it) of  $J_{(\alpha)}$  (or  $F_{(\alpha)}$ ) exhibits variance of at least  $\frac{J_{(\alpha)}^2}{k}$  (or  $\frac{F_{(\alpha)}^2}{k} \Delta^2$ ). In this sense, our proposed estimator is statistically optimal (up to a constant factor) in the framework of CC. Furthermore, the integral  $I_2$  in (40) can be numerically evaluated. Figure 3 plots  $\frac{1}{I_2 - 1}$  (dashed curve) and  $3 - 2\Delta$  (solid curve). Our proposed estimator is close to be 100% optimal and hence there is little room for improvement.

## 5. Experiments

This section demonstrates that the proposed estimator  $\hat{J}_{(\alpha)}$  in (9) is a practical algorithm, while the previously proposed geometric mean algorithm (Li, 2009a) is inadequate for entropy estimation. We also demonstrate that algorithms based on *symmetric stable random projections* (Indyk, 2006; Li, 2009a; Li and Hastie, 2007) are not suitable for entropy estimation in practice. Note that Lemma 7 has shown that the harmonic mean algorithm proposed in (Li, 2009a) is only 3.29-fold better than the geometric mean algorithm and hence it makes no essential difference for entropy estimation.

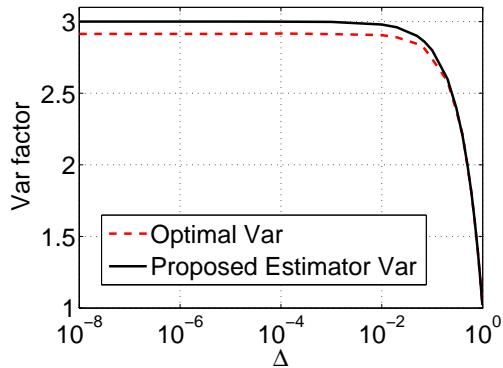


Figure 3: Dashed (red) curve:  $\frac{1}{I_2 - 1}$  as in (40). Solid (black) curve:  $3 - 2\Delta$ .

### 5.1. Data

Since the estimation accuracy is what we are interested in, we simply use static data instead of real data streams, because the projected data vector  $X = \mathbf{R}^T A_t$  is the same at the end of the stream, regardless of whether it is computed at once (i.e., static) or incrementally (i.e., dynamic). As summarized in Table 1, 8 English words are selected from a chunk of Web crawl data, i.e., 8 vectors whose entries are the numbers of word occurrences in each document. The words are selected fairly randomly, although we make sure they cover a wide range of data sparsity, from function words (e.g., “A”), to common words (e.g., “FRIDAY”) to rare words (e.g., “TWIST”).

### 5.2. Estimating Shannon Entropies

We used the estimated frequency moments to estimate the Shannon entropies. For the data vector “TWIST”, we present the results at sample sizes  $k = 3, 10, 100, 1000$ , and  $10000$ . For all other vectors, we did not use  $k = 10000$ . Figure 4 presents the normalized mean square errors (MSEs).

Using our proposed algorithm (middle panels), only  $k = 10$  samples already produces fairly accurate estimates. In fact, for some vectors (such as “A”), even  $k = 3$  may provide reasonable estimates. We believe the performance of the new estimator is remarkable. Another nice property is that the estimation errors become stable after (e.g.,)  $\Delta < 10^{-3}$  (or  $10^{-4}$ ). This essentially frees practitioners from specifying  $\Delta$ .

Table 1: The data set consists of 8 English words selected from a corpus of Web pages, forming 8 vectors whose values are the word occurrences. The table lists their fractions of non-zeros (sparsity) and the Shannon entropies ( $H$ ). The last column is the variance ratio for comparing CC with another algorithm named CRS; the details are in Section 6.

Word	Sparsity	Entropy $H$	Improvement over CRS
TWIST	0.004	5.4873	2.1
FRIDAY	0.034	7.0487	38.9
FUN	0.047	7.6519	23.1
BUSINESS	0.126	8.3995	48.7
NAME	0.144	8.5162	65.9
HAVE	0.267	8.9782	67.7
THIS	0.423	9.3893	84.4
A	0.596	9.5463	113.7

In comparison, the performance of the *geometric mean* algorithm (left panels) is not satisfactory. This is because its variance decreases only at the rate of  $O(\Delta)$ , not  $O(\Delta^2)$ . Also clearly, using *symmetric stable random projections* (right panels) would not provide good estimates of the Shannon entropy (unless the sample size is extremely large with a carefully chosen  $\Delta$ ).

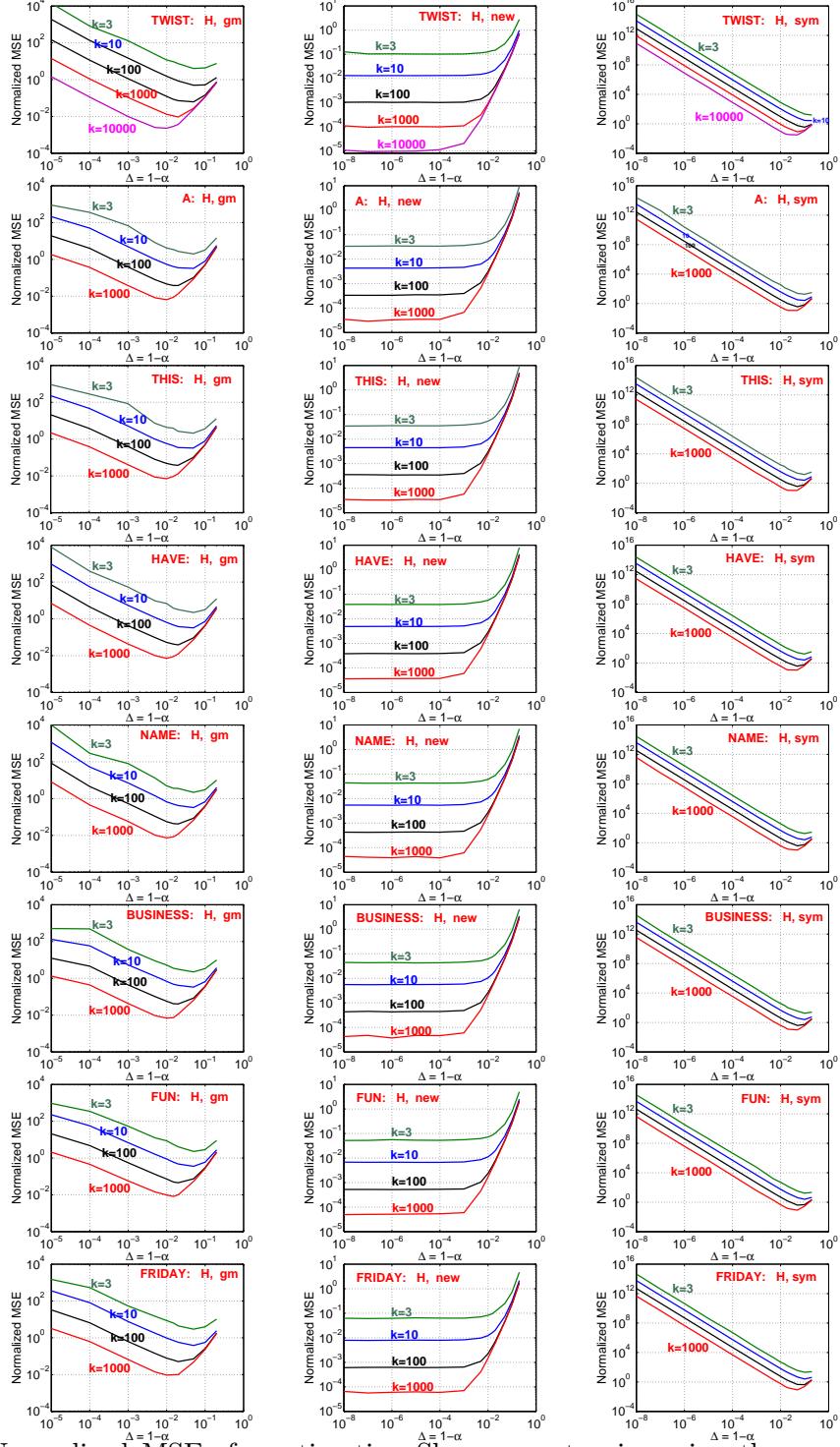


Figure 4: Normalized MSEs for estimating Shannon entropies using the geometric mean algorithm (left panels) proposed in (Li, 2009a), the proposed new algorithm  $\hat{J}_{(\alpha)}$  (9) (middle panels) in this paper, and the geometric mean algorithm for symmetric stable random projections (right panels) in (Li, 2008).

## 6. Comparisons with Conditional Random Sampling (CRS)

Conditional Random Sampling (CRS), which is applicable to data stream computations, is another randomized algorithm (Li and Church, 2007; Li et al., 2008) particularly designed for sampling from sparse data. The significant advantage of CRS is that the method is “one sketch for all,” meaning that the same set of “sketches” can be used to estimate a very wide range of summary statistics and distances including histograms, cross-entropy,  $\chi^2$  distances, inner products, general  $l_\alpha$  distances (for any  $\alpha$ ). In comparison, the methods of (symmetric and skewed)  $\alpha$ -stable random projections are generally limited to  $0 < \alpha \leq 2$  and one has to re-do the projections (and keep multiple sets of samples) if the application requires to use multiple  $\alpha$  values.<sup>3</sup> A recent manuscript (Zhao et al., 2010) compared CRS with a variety of other algorithms on the network data provided by ATT Labs.

It is interesting to compare CC (using the new estimator in this paper) with CRS for estimating Shannon entropy. Suppose we use the estimator (10) with sufficiently small  $\Delta$ . Then the estimation variance is roughly just  $\frac{3}{k}$ , essentially independent of the original data. Using the generic approximate variance formula in Li et al. (2008), the variance for entropy estimation is denoted by  $Var(\hat{H}_{CRS})$ :

$$Var(\hat{H}_\alpha) \approx \frac{3}{k} + O\left(\frac{1}{k^2}\right), \quad \text{for sufficiently small } \Delta \quad (43)$$

$$Var(\hat{H}_{CRS}) \approx \frac{|\{i | A_t[i] > 0\}|}{k} \left\{ \sum_{i=1}^D \left[ \frac{A_t[i]}{F_{(1)}} \log \frac{A_t[i]}{F_{(1)}} \right]^2 - \frac{1}{D} \left[ \sum_{i=1}^D \frac{A_t[i]}{F_{(1)}} \log \frac{A_t[i]}{F_{(1)}} \right]^2 \right\} + O\left(\frac{1}{k^2}\right). \quad (44)$$

Table 1 (last column) already presents the variance ratios:  $\frac{Var(\hat{H}_{CRS})}{Var(\hat{H}_\alpha)}$  for the data used in our experiments. The ratios range from 2.1 to 113.7. The comparison further conforms that CC is extremely accurate for entropy estimation. On the other hand, CRS is actually also pretty good for entropy estimation, considering it is “one-sketch-for-all.” Another significant advantage of CRS is that it is not limited to the strict-Turnstile data stream model, or even the general Turnstile model. It is particularly useful when applications require using nonlinearly transformed data (e.g., TF-IDF weighting in search and natural language processing) instead of the original data.

---

3. We should mention that the method of normal ( $l_2$ ) random projections was recently extended (Li et al., 2010) to estimating  $l_\alpha$  distances for  $\alpha = 4, 6, 8, \dots$  in massive (static) data matrices.

## 7. Conclusion

Many machine learning (e.g., neural computation, graph estimation) and data mining (e.g., anomaly detection) problems require estimating the Shannon entropy. When the data are dynamic (e.g., data streams), efficient estimation of the Shannon entropy using small space has been a challenging problem. It is known that we can approximate the Shannon entropy using the  $\alpha$ -th frequency moment of the stream with  $\alpha$  very close to 1, if the estimator of the moment is accurate enough with variance proportional to  $O(\Delta^2)$ , where  $\Delta = |1 - \alpha|$ . Our paper provides such a practical estimator. Our method is an ideal solution to the problem of entropy estimation when the data streams follow the strict-Turnstile model.

For  $\nu$ -additive Shannon entropy estimation, the sample complexity of the algorithm is only  $O\left(\frac{1}{\nu^2}\right)$ . The constant factor for this bound is merely about 6. In addition, we prove that our algorithm achieves an upper bound of the Fisher information and in fact it is close to 100% statistically optimal. An empirical study is also conducted to verify the accuracy of our algorithm.

**Further research:** To further reduce the processing cost in order to better accommodate high-rate data streams, it is desirable to replace the dense matrix of skewed stable variables by a sparse matrix of Pareto-type variables. This is closely related to the prior study of *very sparse symmetric stable random projections* (Li, 2007). However, the extension to CC requires further work.

## Acknowledgments

The authors thank the anonymous reviewers for their constructive comments. Ping Li's research is partially supported by the National Science Foundation (DMS-0808864), the Office of Naval Research (YIP-N000140910911), and a gift from Google. Cun-Hui Zhang's research is partially supported by the National Science Foundation (DMS-0804626, DMS-0906420) and the National Security Agency (H98230-09-1-0006, H98230-11-1-0205).

## References

- Charu C. Aggarwal, Jiawei Han, Jianyong Wang, and Philip S. Yu. On demand classification of data streams. In *KDD*, pages 503–508, Seattle, WA, 2004.
- Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *STOC*, pages 20–29, Philadelphia, PA, 1996.
- Lakshminath Bhuvanagiri and Sumit Ganguly. Estimating entropy over data streams. In *ESA*, pages 148–159, Zurich, Switzerland, 2006.
- Daniela Brauckhoff, Bernhard Tellenbach, Arno Wagner, Martin May, and Anukool Lakhina. Impact of packet sampling on anomaly detection metrics. In *IMC*, pages 159–164, Rio de Janeiro, Brazil, 2006.
- Amit Chakrabarti, Khanh Do Ba, and S. Muthukrishnan. Estimating entropy and entropy norm on data streams. *Internet Mathematics*, 3(1):63–78, 2006.

Amit Chakrabarti, Graham Cormode, and Andrew McGregor. A near-optimal algorithm for computing the entropy of a stream. In *SODA*, pages 328–335, New Orleans, Louisiana, 2007.

John M. Chambers, C. L. Mallows, and B. W. Stuck. A method for simulating stable random variables. *Journal of the American Statistical Association*, 71(354):340–344, 1976.

Carlotta Domeniconi and Dimitrios Gunopulos. Incremental support vector machine construction. In *ICDM*, pages 589–592, San Jose, CA, 2001.

Laura Feinstein, Dan Schnackenberg, Ravindra Balupari, and Darrell Kindred. Statistical approaches to DDoS attack detection and response. In *DARPA Information Survivability Conference and Exposition*, pages 303–314, 2003.

Izrail S. Gradshteyn and Iosif M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, New York, fifth edition, 1994.

Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian. Streaming and sub-linear approximation of entropy and information distances. In *SODA*, pages 733 – 742, Miami, FL, 2006.

Anupam Gupta, John D. Lafferty, Han Liu, Larry A. Wasserman, and Min Xu. Forest density estimation. In *COLT*, pages 394–406, Haifa, Israel, 2010.

Nicholas J. A. Harvey, Jelani Nelson, and Krzysztof Onak. Sketching and streaming entropy via approximation theory. In *FOCS*, 2008a.

Nicholas J. A. Harvey, Jelani Nelson, and Krzysztof Onak. Streaming algorithms for estimating entropy. In *ITW*, 2008b.

M E. Havrda and F. Charvát. Quantification methods of classification processes: Concept of structural  $\alpha$ -entropy. *Kybernetika*, 3:30–35, 1967.

Monika R. Henzinger, Prabhakar Raghavan, and Sridhar Rajagopalan. *Computing on Data Streams*. American Mathematical Society, Boston, MA, USA, 1999.

Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of ACM*, 53(3):307–323, 2006.

Anukool Lakhina, Mark Crovella, and Christophe Diot. Mining anomalies using traffic feature distributions. In *SIGCOMM*, pages 217–228, Philadelphia, PA, 2005.

Ashwin Lall, Vyas Sekar, Mitsunori Ogihara, Jun Xu, and Hui Zhang. Data streaming algorithms for estimating entropy of network traffic. In *SIGMETRICS*, pages 145–156, Saint Malo, France, 2006.

Ping Li. Very sparse stable random projections for dimension reduction in  $l_\alpha$  ( $0 < \alpha \leq 2$ ) norm. In *KDD*, San Jose, CA, 2007.

Ping Li. Estimators and tail bounds for dimension reduction in  $l_\alpha$  ( $0 < \alpha \leq 2$ ) using stable random projections. In *SODA*, pages 10 – 19, San Francisco, CA, 2008.

- Ping Li. Compressed counting. In *SODA*, New York, NY, 2009a.
- Ping Li. Improving compressed counting. In *UAI*, Montreal, CA, 2009b.
- Ping Li and Kenneth W. Church. A sketch algorithm for estimating two-way and multi-way associations. *Computational Linguistics (Preliminary results appeared in HLT/EMNLP 2005)*, 33(3):305–354, 2007.
- Ping Li and Trevor J. Hastie. A unified near-optimal estimator for dimension reduction in  $l_\alpha$  ( $0 < \alpha \leq 2$ ) using stable random projections. In *NIPS*, Vancouver, BC, Canada, 2007.
- Ping Li, Kenneth W. Church, and Trevor J. Hastie. One sketch for all: Theory and applications of conditional random sampling. In *NIPS (Preliminary results appeared in NIPS 2006)*, Vancouver, BC, Canada, 2008.
- Ping Li, Michael Mahoney, and Yiyuan She. Approximating higher-order distances using random projections. In *UAI*, 2010.
- Qiaozhu Mei and Kenneth Church. Entropy of search logs: How hard is search? with personalization? with backoff? In *WSDM*, pages 45 – 54, Palo Alto, CA, 2008.
- S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1:117–236, 2 2005.
- Liam Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6): 1191–1253, 2003.
- Alfred Rényi. On measures of information and entropy. In *The 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960*, pages 547–561, 1961.
- Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.
- Kuai Xu, Zhi-Li Zhang, and Supratik Bhattacharyya. Profiling internet backbone traffic: behavior models and applications. In *SIGCOMM '05: Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 169–180, Philadelphia, Pennsylvania, USA, 2005.
- Qiang Yang and Xindong Wu. 10 challeng problems in data mining research. *International Journal of Information Technology and Decision Making*, 5(4):597–604, 2006.
- Haiquan Zhao, Ashwin Lall, Mitsunori Ogihara, Oliver Spatscheck, Jia Wang, and Jun Xu. A data streaming algorithm for estimating entropies of od flows. In *IMC*, San Diego, CA, 2007.
- Haiquan Zhao, Nan Hua, Ashwin Lall, Ping Li, Jia Wang, and Jun Xu. Towards a universal sketch for origin-destination network measurements. Technical report, 2010.

# A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences

**Odalric-Ambrym Maillard**

**Rémi Munos**

*INRIA Lille Nord Europe*

*SequeL Project*

*40 avenue Halley*

*59650 Villeneuve d'Ascq, France*

ODALRIC.MAILLARD@INRIA.FR

REMI.MUNOS@INRIA.FR

**Gilles Stoltz**

GILLES.STOLTZ@ENS.FR

*Ecole Normale Supérieure, CNRS, INRIA*

*45 rue d'Ulm*

*75005 Paris, France*

&

*HEC Paris, CNRS*

*1 rue de la Libération*

*78351 Jouy-en-Josas, France*

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We consider a Kullback-Leibler-based algorithm for the stochastic multi-armed bandit problem in the case of distributions with finite supports (not necessarily known beforehand), whose asymptotic regret matches the lower bound of Burnetas and Katehakis (1996). Our contribution is to provide a finite-time analysis of this algorithm; we get bounds whose main terms are smaller than the ones of previously known algorithms with finite-time analyses (like UCB-type algorithms).

**Keywords:** Multi-armed bandit problem, finite-time analysis, Kullback-Leibler divergence, Sanov's lemma

## 1. Introduction

The *stochastic* multi-armed bandit problem, introduced by Robbins (1952), formalizes the problem of decision-making under uncertainty, and illustrates the fundamental tradeoff that appears between *exploration*, i.e., making decisions in order to improve the knowledge of the environment, and *exploitation*, i.e., maximizing the payoff.

**Setting.** In this paper, we consider a multi-armed bandit problem with *finitely* many arms indexed by  $\mathcal{A}$ , for which each arm  $a \in \mathcal{A}$  is associated with an unknown and fixed probability distribution  $\nu_a$  over  $[0, 1]$ . The game is *sequential* and goes as follows: at each round  $t \geq 1$ , the player first picks an arm  $A_t \in \mathcal{A}$  and then receives a stochastic payoff  $Y_t$  drawn at random according to  $\nu_{A_t}$ . He only gets to see the payoff  $Y_t$ .

For each arm  $a \in \mathcal{A}$ , we denote by  $\mu_a$  the expectation of its associated distribution  $\nu_a$  and we let  $a^*$  be any optimal arm, i.e.,  $a^* \in \operatorname{argmax}_{a \in \mathcal{A}} \mu_a$ .

We write  $\mu^*$  as a short-hand notation for the largest expectation  $\mu_{a^*}$  and denote the gap of the expected payoff  $\mu_a$  of an arm  $a \in \mathcal{A}$  to  $\mu^*$  as  $\Delta_a = \mu^* - \mu_a$ . In addition, the number of times each arm  $a \in \mathcal{A}$  is pulled between the rounds 1 and  $T$  is referred to as  $N_T(a)$ ,

$$N_T(a) \stackrel{\text{def}}{=} \sum_{t=1}^T \mathbb{I}_{\{A_t=a\}}.$$

The quality of a strategy will be evaluated through the standard notion of *expected regret*, which we recall now. The expected regret (or simply regret) at round  $T \geq 1$  is defined as

$$R_T \stackrel{\text{def}}{=} \mathbb{E}\left[T\mu^* - \sum_{t=1}^T Y_t\right] = \mathbb{E}\left[T\mu^* - \sum_{t=1}^T \mu_{A_t}\right] = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[N_T(a)], \quad (1)$$

where we used the tower rule for the first equality. Note that the expectation is with respect to the random draws of the  $Y_t$  according to the  $\nu_{A_t}$  and also to the possible auxiliary randomizations that the decision-making strategy is resorting to.

The regret measures the cumulative loss resulting from pulling sub-optimal arms, and thus quantifies the amount of exploration required by an algorithm in order to find a best arm, since, as (1) indicates, the regret scales with the expected number of pulls of sub-optimal arms. Since the formulation of the problem by Robbins (1952) the regret has been a popular criterion for assessing the quality of a strategy.

**Known lower bounds.** Lai and Robbins (1985) showed that for some (one-dimensional) parametric classes of distributions, any consistent strategy (i.e., any strategy not pulling sub-optimal arms more than in a polynomial number of rounds) will despite all asymptotically pull in expectation any sub-optimal arm  $a$  at least

$$\mathbb{E}[N_T(a)] \geq \left( \frac{1}{\mathcal{K}(\nu_a, \nu^*)} + o(1) \right) \log(T)$$

times, where  $\mathcal{K}(\nu_a, \nu^*)$  is the Kullback-Leibler (KL) divergence between  $\nu_a$  and  $\nu^*$ ; it measures how close distributions  $\nu_a$  and  $\nu^*$  are from a theoretical information perspective.

Later, Burnetas and Katehakis (1996) extended this result to some classes of multi-dimensional parametric distributions and proved the following generic lower bound: for a given family  $\mathcal{P}$  of possible distributions over the arms,

$$\mathbb{E}[N_T(a)] \geq \left( \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^*)} + o(1) \right) \log(T), \quad \text{where } \mathcal{K}_{\inf}(\nu_a, \mu^*) \stackrel{\text{def}}{=} \inf_{\nu \in \mathcal{P}: E(\nu) > \mu^*} \mathcal{K}(\nu_a, \nu),$$

with the notation  $E(\nu)$  for the expectation of a distribution  $\nu$ . The intuition behind this improvement is to be related to the goal that we want to achieve in bandit problems; it is not detecting whether a distribution is optimal or not (for this goal, the relevant quantity would be  $\mathcal{K}(\nu_a, \nu^*)$ ), but rather achieving the optimal rate of reward  $\mu^*$  (i.e., one needs to measure how close  $\nu_a$  is to any distribution  $\nu \in \mathcal{P}$  whose expectation is at least  $\mu^*$ ).

**Known upper bounds.** Lai and Robbins (1985) provided an algorithm based on the KL divergence, which has been extended by Burnetas and Katehakis (1996) to an algorithm based on  $\mathcal{K}_{\inf}$ ; it is asymptotically optimal since the number of pulls of any sub-optimal arm  $a$  satisfies

$$\mathbb{E}[N_T(a)] \leq \left( \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^*)} + o(1) \right) \log(T).$$

This result holds for finite-dimensional parametric distributions under some assumptions, e.g., the distributions having a finite and known support or belonging to a set of Gaussian distributions with known variance. Recently Honda and Takemura (2010a) extended this asymptotic result to the case of distributions  $\mathcal{P}$  with support in  $[0, 1]$  and such that  $\mu^* < 1$ ; the key ingredient in this case is that  $\mathcal{K}_{\inf}(\nu_a, \mu^*)$  is equal to

$$\mathcal{K}_{\min}(\nu_a, \mu^*) \stackrel{\text{def}}{=} \inf_{\nu \in \mathcal{P}: E(\nu) \geq \mu^*} \mathcal{K}(\nu_a, \nu).$$

**Motivation.** All the results mentioned above provide asymptotic bounds only. However, any algorithm is only used for a finite number of rounds and it is thus essential to provide a finite-time analysis of its performance. Auer et al. (2002) initiated this work by providing an algorithm (UCB1) based on a Chernoff-Hoeffding bound; it pulls any sub-optimal arm, till any time  $T$ , at most  $(8/\Delta_a^2) \log T + 1 + \pi^2/3$  times, in expectation. Although this yields a logarithmic regret, the multiplicative constant depends on the gap  $\Delta_a^2 = (\mu^* - \mu_a)^2$  but not on  $\mathcal{K}_{\inf}(\nu_a, \mu^*)$ , which can be seen to be larger than  $\Delta_a^2/2$  by Pinsker's inequality; that is, this non-asymptotic bound does not have the right dependence in the distributions. (How much is gained of course depends on the specific families of distributions at hand.) Audibert et al. (2009) provided an algorithm (UCB-V) that takes into account the empirical variance of the arms and exhibited a strategy such that  $\mathbb{E}[N_T(a)] \leq 10(\sigma_a^2/\Delta_a^2 + 2/\Delta_a) \log T$  for any time  $T$  (where  $\sigma_a^2$  is the variance of arm  $a$ ); it improves over UCB1 in case of arms with small variance. Other variants include the MOSS algorithm by Audibert and Bubeck (2010) and Improved UCB by Auer and Ortner (2010).

However, all these algorithms only rely on one moment (for UCB1) or two moments (for UCB-V) of the empirical distributions of the obtained rewards; they do not fully exploit the empirical distributions. As a consequence, the resulting bounds are expressed in terms of the means  $\mu_a$  and variances  $\sigma_a^2$  of the sub-optimal arms and not in terms of the quantity  $\mathcal{K}_{\inf}(\nu_a, \mu^*)$  appearing in the lower bounds. The numerical experiments reported in Filippi (2010) confirm that these algorithms are less efficient than those based on  $\mathcal{K}_{\inf}$ .

**Our contribution.** In this paper we analyze a  $\mathcal{K}_{\inf}$ -based algorithm inspired by the ones studied in Lai and Robbins (1985); Burnetas and Katehakis (1996); Filippi (2010); it indeed takes into account the full empirical distribution of the observed rewards. The analysis is performed (with explicit bounds) in the case of Bernoulli distributions over the arms. Less explicit but finite-time bounds are obtained in the case of finitely supported distributions (whose supports do not need to be known in advance). Finally, we pave the way for handling the case of general finite-dimensional parametric distributions. These results improve on the ones by Burnetas and Katehakis (1996); Honda and Takemura (2010a) since finite-time bounds (implying their asymptotic results) are obtained; and on Auer et al. (2002); Audibert et al. (2009) as the dependency of the main term scales with  $\mathcal{K}_{\inf}(\nu_a, \mu^*)$ . The proposed  $\mathcal{K}_{\inf}$ -based algorithm is also more natural and more appealing than the one presented in Honda and Takemura (2010a).

**Recent related works.** Since our initial submission of the present paper, we got aware of two papers that tackle problems similar to ours. First, a revised version of Honda and Takemura (2010b, personal communication) obtains finite-time regret bounds (with prohibitively large constants) for a *randomized* (less natural) strategy in the case of distributions with finite supports (also not known in advance). Second, another paper at this conference (Garivier and Cappé, 2011) also deals with the  $\mathcal{K}$ -strategy which we study in Theorem 3; they however do not obtain second-order terms in closed forms as we do and later extend their strategy to exponential families of distributions (while we extend our strategy to the case of distributions with finite supports). On the other hand, they show how the  $\mathcal{K}$ -strategy can be extended in a straightforward manner to guarantee bounds with respect to the family of all bounded distributions on a known interval; these bounds are suboptimal but improve on the ones of UCB-type algorithms.

## 2. Definitions and tools

Let  $\mathcal{X}$  be a Polish space; in the next sections, we will consider  $\mathcal{X} = \{0, 1\}$  or  $\mathcal{X} = [0, 1]$ . We denote by  $\mathcal{P}(\mathcal{X})$  the set of probability distributions over  $\mathcal{X}$  and equip  $\mathcal{P}(\mathcal{X})$  with the distance  $d$  induced by the norm  $\|\cdot\|$  defined by  $\|\nu\| = \sup_{f \in \mathcal{L}} |\int_{\mathcal{X}} f d\nu|$ , where  $\mathcal{L}$  is the set of Lipschitz functions over  $\mathcal{X}$ , taking values in  $[-1, 1]$  and with Lipschitz constant smaller than 1.

**Kullback-Leibler divergence:** For two elements  $\nu, \kappa \in \mathcal{P}(\mathcal{X})$ , we write  $\nu \ll \kappa$  when  $\nu$  is absolutely continuous with respect to  $\kappa$  and denote in this case by  $d\nu/d\kappa$  the density of  $\nu$  with respect to  $\kappa$ . We recall that the Kullback-Leibler divergence between  $\nu$  and  $\kappa$  is defined as

$$\mathcal{K}(\nu, \kappa) = \int_{[0,1]} \frac{d\nu}{d\kappa} \log \frac{d\nu}{d\kappa} d\kappa \quad \text{if } \nu \ll \kappa; \quad \text{and} \quad \mathcal{K}(\nu, \kappa) = +\infty \quad \text{otherwise.} \quad (2)$$

**Empirical distribution:** We consider a sequence  $X_1, X_2, \dots$  of random variables taking values in  $\mathcal{X}$ , independent and identically distributed according to a distribution  $\nu$ . For all integers  $t \geq 1$ , we denote the empirical distribution corresponding to the first  $t$  elements of the sequence by

$$\widehat{\nu}_t = \frac{1}{t} \sum_{s=1}^t \delta_{X_t}.$$

**Non-asymptotic Sanov's Lemma:** The following lemma follows from a straightforward adaptation of Dinwoodie (1992, Theorem 2.1 and comments on page 372). Details of the proof are provided in the extended version (Maillard et al., 2011) of the present paper.

**Lemma 1** *Let  $\mathcal{C}$  be an open convex subset of  $\mathcal{P}(\mathcal{X})$  such that  $\Lambda(\mathcal{C}) = \inf_{\kappa \in \mathcal{C}} \mathcal{K}(\kappa, \nu) < \infty$ .*

*Then, for all  $t \geq 1$ , one has  $\mathbb{P}_{\nu}\{\widehat{\nu}_t \in \overline{\mathcal{C}}\} \leq e^{-t\Lambda(\overline{\mathcal{C}})}$  where  $\overline{\mathcal{C}}$  is the closure of  $\mathcal{C}$ .*

This lemma should be thought of as a deviation inequality. The empirical distribution converges (in distribution) to  $\nu$ . Now, if (and only if)  $\nu$  is not in the closure of  $\mathcal{C}$ , then  $\Lambda(\mathcal{C}) > 0$  and the lemma indicates how unlikely it is that  $\widehat{\nu}_t$  is in this set  $\overline{\mathcal{C}}$  not containing the limit  $\nu$ . The probability of interest decreases at a geometric rate, which depends on  $\Lambda(\mathcal{C})$ .

### 3. Finite-time analysis for Bernoulli distributions

In this section, we start with the case of Bernoulli distributions. Although this case is a special case of the general results of Section 4, we provide here a complete and self-contained analysis of this case, where, in addition, we are able to provide closed forms for all the terms in the regret bound. Note however that the resulting bound is slightly worse than what could be derived from the general case (for which more sophisticated tools are used). This result is mainly provided as a warm-up.

#### 3.1. Reminder of some useful results for Bernoulli distributions

We denote by  $\mathcal{B}$  the subset of  $\mathcal{P}([0, 1])$  formed by the Bernoulli distributions; it corresponds to  $\mathcal{B} = \mathcal{P}(\{0, 1\})$ . A generic element of  $\mathcal{B}$  will be denoted by  $\beta(p)$ , where  $p \in [0, 1]$  is the probability mass put on 1. We consider a sequence  $X_1, X_2, \dots$  of independent and identically distributed random variables, with common distribution  $\beta(p)$ ; for the sake of clarity we will index, in this subsection only, all probabilities and expectations with  $p$ .

For all integers  $t \geq 1$ , we denote by  $\hat{p}_t = \frac{1}{t} \sum_{s=1}^t X_t$  the empirical average of the first  $t$  elements of the sequence.

The lemma below follows from an adaptation of Garivier and Leonardi (2011, Proposition 2).

**Lemma 2** *For all  $p \in [0, 1]$ , all  $\varepsilon > 1$ , and all  $t \geq 1$ ,*

$$\mathbb{P}_p \left( \bigcup_{s=1}^t \left\{ s \mathcal{K}(\beta(\hat{p}_s), \beta(p)) \geq \varepsilon \right\} \right) \leq 2e \lceil \varepsilon \log t \rceil e^{-\varepsilon}.$$

*In particular, for all random variables  $N_t$  taking values in  $\{1, \dots, t\}$ ,*

$$\mathbb{P}_p \left\{ N_t \mathcal{K}(\beta(\hat{p}_{N_t}), \beta(p)) \geq \varepsilon \right\} \leq 2e \lceil \varepsilon \log t \rceil e^{-\varepsilon}.$$

Another immediate fact about Bernoulli distributions is that for all  $p \in (0, 1)$ , the mappings  $\mathcal{K}_{\cdot, p} : q \in (0, 1) \mapsto \mathcal{K}(\beta(p), \beta(q))$  and  $\mathcal{K}_{p, \cdot} : q \in [0, 1] \mapsto \mathcal{K}(\beta(q), \beta(p))$  are continuous and take finite values. In particular, we have, for instance, that for all  $\varepsilon > 0$  and  $p \in (0, 1)$ , the set

$$\left\{ q \in [0, 1] : \mathcal{K}(\beta(p), \beta(q)) \leq \varepsilon \right\}$$

is a closed interval containing  $p$ . This property still holds when  $p \in \{0, 1\}$ , as in this case, the interval is reduced to  $\{p\}$ .

#### 3.2. Strategy and analysis

We consider the so-called  $\mathcal{K}$ -strategy of Figure 1, which was already considered in the literature, see Burnetas and Katehakis (1996); Filippi (2010). The numerical computation of the quantities  $B_{a,t}^+$  is straightforward (by convexity of  $\mathcal{K}$  in its second argument, by using iterative methods) and is detailed therein.

Before proceeding, we denote by  $\sigma_a^2 = \mu_a(1 - \mu_a)$  the variance of each arm  $a \in \mathcal{A}$  (and take the short-hand notation  $\sigma^{*,2}$  for the variance of an optimal arm).

---

*Parameters:* A non-decreasing function  $f : \mathbb{N} \rightarrow \mathbb{R}$

*Initialization:* Pull each arm of  $\mathcal{A}$  once

*For* rounds  $t + 1$ , where  $t \geq |\mathcal{A}|$ ,

- compute for each arm  $a \in \mathcal{A}$  the quantity

$$B_{a,t}^+ = \max \left\{ q \in [0, 1] : N_t(a) \mathcal{K}(\beta(\hat{\mu}_{a,N_t(a)}), \beta(q)) \leq f(t) \right\},$$

$$\text{where } \hat{\mu}_{a,N_t(a)} = \frac{1}{N_t(a)} \sum_{s \leq t: A_s=a} Y_s;$$

- in case of a tie, pick an arm with largest value of  $\hat{\mu}_{a,N_t(a)}$ ;
  - pull any arm  $A_{t+1} \in \operatorname{argmax}_{a \in \mathcal{A}} B_{a,t}^+$ .
- 

Figure 1: The  $\mathcal{K}$ -strategy.

**Theorem 3** When  $\mu^* \in (0, 1)$ , for all non-decreasing functions  $f : \mathbb{N} \rightarrow \mathbb{R}_+$  such that  $f(1) \geq 1$ , the expected regret  $R_T$  of the strategy of Figure 1 is upper bounded by the infimum, as the  $(c_a)_{a \in \mathcal{A}}$  describe  $(0, +\infty)$ , of the quantities

$$\sum_{a \in \mathcal{A}} \Delta_a \left( \frac{(1 + c_a) f(T)}{\mathcal{K}(\beta(\mu_a), \beta(\mu^*))} + 4e \sum_{t=|\mathcal{A}|}^{T-1} \lceil f(t) \log t \rceil e^{-f(t)} + \frac{(1 + c_a)^2}{8 c_a^2 \Delta_a^2 \min\{\sigma_a^4, \sigma^{*,4}\}} \mathbb{I}_{\{\mu_a \in (0,1)\}} + 3 \right).$$

For  $\mu^* = 0$ , its regret is null. For  $\mu^* = 1$ , it satisfies  $R_T \leq 2(|\mathcal{A}| - 1)$ .

A possible choice for the function  $f$  is  $f(t) = \log((et) \log^3(et))$ , which is non decreasing, satisfies  $f(1) \geq 1$ , and is such that the second term in the sum above is bounded (by a basic result about so-called Bertrand's series). Now, as the constants  $c_a$  in the bound are parameters of the analysis (and not of the strategy), they can be optimized. For instance, with the choice of  $f(t)$  mentioned above, taking each  $c_a$  proportional to  $(\log T)^{-1/3}$  (up to a multiplicative constant that depends on the distributions  $\nu_a$ ) entails the regret bound

$$\sum_{a \in \mathcal{A}} \Delta_a \frac{\log T}{\mathcal{K}(\beta(\mu_a), \beta(\mu^*))} + \varepsilon_T,$$

where it is easy to give an explicit and closed-form expression of  $\varepsilon_T$ ; in this conference version, we only indicate that  $\varepsilon_T$  is of order of  $(\log T)^{2/3}$  but we do not know whether the order of magnitude of this second-order term is optimal.

**Proof** We first deal with the case where  $\mu^* \notin \{0, 1\}$  and introduce an additional notation. In view of the remark at the end of Section 3.1, for all arms  $a$  and rounds  $t$ , we let  $B_{a,t}^-$  be the element in  $[0, 1]$  such that

$$\left\{ q \in [0, 1] : N_t(a) \mathcal{K}(\beta(\hat{\mu}_{a,N_t(a)}), \beta(q)) \leq f(t) \right\} = [B_{a,t}^-, B_{a,t}^+]. \quad (3)$$

As (1) indicates, it suffices to bound  $N_T(a)$  for all suboptimal arms  $a$ , i.e., for all arms such that  $\mu_a < \mu^*$ . We will assume in addition that  $\mu_a > 0$  (and we also have  $\mu_a \leq \mu^* < 1$ ); the case where  $\mu_a = 0$  will be handled separately.

**Step 1: A decomposition of the events of interest.** For  $t \geq |\mathcal{A}|$ , when  $A_{t+1} = a$ , we have in particular, by definition of the strategy, that  $B_{a,t}^+ \geq B_{a^*,t}^+$ . On the event

$$\{A_{t+1} = a\} \cap \left\{\mu^* \in [B_{a^*,t}^-, B_{a^*,t}^+]\right\} \cap \left\{\mu_a \in [B_{a,t}^-, B_{a,t}^+]\right\},$$

we therefore have, on the one hand,  $\mu^* \leq B_{a^*,t}^+ \leq B_{a,t}^+$  and on the other hand,  $B_{a,t}^- \leq \mu_a \leq \mu^*$ , that is, the considered event is included in  $\{\mu^* \in [B_{a,t}^-, B_{a,t}^+]\}$ . We thus proved that

$$\{A_{t+1} = a\} \subseteq \left\{\mu^* \notin [B_{a^*,t}^-, B_{a^*,t}^+]\right\} \cup \left\{\mu_a \notin [B_{a,t}^-, B_{a,t}^+]\right\} \cup \left\{\mu^* \in [B_{a,t}^-, B_{a,t}^+]\right\}.$$

Going back to the definition (3), we get in particular the inclusion

$$\begin{aligned} \{A_{t+1} = a\} &\subseteq \left\{N_t(a^*) \mathcal{K}\left(\beta(\widehat{\mu}_{a^*,N_t(a^*)}), \beta(\mu^*)\right) > f(t)\right\} \\ &\cup \left\{N_t(a) \mathcal{K}\left(\beta(\widehat{\mu}_{a,N_t(a)}), \beta(\mu_a)\right) > f(t)\right\} \\ &\cup \left(\left\{N_t(a) \mathcal{K}\left(\beta(\widehat{\mu}_{a,N_t(a)}), \beta(\mu^*)\right) \leq f(t)\right\} \cap \{A_{t+1} = a\}\right). \end{aligned}$$

**Step 2: Bounding the probabilities of two elements of the decomposition.** We consider the filtration  $(\mathcal{F}_t)$ , where for all  $t \geq 1$ , the  $\sigma$ -algebra  $\mathcal{F}_t$  is generated by  $A_1, Y_1, \dots, A_t, Y_t$ . In particular,  $A_{t+1}$  and thus all  $N_{t+1}(a)$  are  $\mathcal{F}_t$ -measurable. We denote by  $\tau_{a,1}$  the deterministic round at which  $a$  was pulled for the first time and by  $\tau_{a,2}, \tau_{a,3}, \dots$  the rounds  $t \geq |\mathcal{A}| + 1$  at which  $a$  was then played; since for all  $k \geq 2$ ,

$$\tau_{a,k} = \min\{t \geq |\mathcal{A}| + 1 : N_t(a) = k\},$$

we see that  $\{\tau_{a,k} = t\}$  is  $\mathcal{F}_{t-1}$ -measurable. Therefore, for each  $k \geq 1$ , the random variable  $\tau_{a,k}$  is a (predictable) stopping time. Hence, by a well-known fact in probability theory (see, e.g., Chow and Teicher 1988, Section 5.3), the random variables  $\widetilde{X}_{a,k} = Y_{\tau_{a,k}}$ , where  $k = 1, 2, \dots$  are independent and identically distributed according to  $\nu_a$ . Since on  $\{N_t(a) = k\}$ , we have the rewriting

$$\widehat{\mu}_{a,N_t(a)} = \widetilde{\mu}_{a,k} \quad \text{where} \quad \widetilde{\mu}_{a,k} = \frac{1}{k} \sum_{j=1}^k \widetilde{X}_{a,j},$$

and since for  $t \geq |\mathcal{A}| + 1$ , one has  $N_t(a) \geq 1$  with probability 1, we can apply the second statement in Lemma 2 and get, for all  $t \geq |\mathcal{A}| + 1$ ,

$$\mathbb{P}\left\{N_t(a) \mathcal{K}\left(\beta(\widehat{\mu}_{a,N_t(a)}), \beta(\mu_a)\right) > f(t)\right\} \leq 2e^{\lceil f(t) \log t \rceil} e^{-f(t)}.$$

A similar argument shows that for all  $t \geq |\mathcal{A}| + 1$ ,

$$\mathbb{P}\left\{N_t(a^*) \mathcal{K}\left(\beta(\widehat{\mu}_{a^*, N_t(a^*)}), \beta(\mu^*)\right) > f(t)\right\} \leq 2e^{\lceil f(t) \log t \rceil} e^{-f(t)}.$$

**Step 3: Rewriting the remaining terms.** We therefore proved that

$$\begin{aligned}\mathbb{E}[N_T(a)] &\leq 1 + 4e \sum_{t=|\mathcal{A}|}^{T-1} \lceil f(t) \log t \rceil e^{-f(t)} \\ &\quad + \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\left(\left\{N_t(a) \mathcal{K}\left(\beta(\widehat{\mu}_{a, N_t(a)}), \beta(\mu^*)\right) \leq f(t)\right\} \cap \{A_{t+1} = a\}\right)\end{aligned}$$

and deal now with the last sum. Since  $f$  is non decreasing, it is bounded by

$$\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}(K_t \cap \{A_{t+1} = a\}) \quad \text{where} \quad K_t = \left\{N_t(a) \mathcal{K}\left(\beta(\widehat{\mu}_{a, N_t(a)}), \beta(\mu^*)\right) \leq f(T)\right\}.$$

$$\text{Now, } \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}(K_t \cap \{A_{t+1} = a\}) = \mathbb{E}\left[\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{I}_{\{A_{t+1}=a\}} \mathbb{I}_{K_t}\right] = \mathbb{E}\left[\sum_{k \geq 2} \mathbb{I}_{\{\tau_{a,k} \leq T\}} \mathbb{I}_{K_{\tau_{a,k}-1}}\right].$$

We note that, since  $N_{\tau_{a,k}-1}(a) = k - 1$ , we have that

$$K_{\tau_{a,k}-1} = \left\{(k-1) \mathcal{K}\left(\beta(\widetilde{\mu}_{a,k-1}), \beta(\mu^*)\right) \leq f(T)\right\}.$$

All in all, since  $\tau_{a,k} \leq T$  implies  $k \leq T - |\mathcal{A}| + 1$  (as each arm is played at least once during the first  $|\mathcal{A}|$  rounds), we have

$$\mathbb{E}\left[\sum_{k \geq 2} \mathbb{I}_{\{\tau_{a,k} \leq T\}} \mathbb{I}_{K_{\tau_{a,k}-1}}\right] \leq \mathbb{E}\left[\sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{I}_{K_{\tau_{a,k}-1}}\right] = \sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{P}\left\{(k-1) \mathcal{K}\left(\beta(\widetilde{\mu}_{a,k-1}), \beta(\mu^*)\right) \leq f(T)\right\}. \quad (4)$$

**Step 4: Bounding the probabilities of the latter sum via Sanov's lemma.** For each  $\gamma > 0$ , we define the convex open set  $\mathcal{C}_\gamma = \left\{\beta(q) \in \mathcal{B} : \mathcal{K}(\beta(q), \beta(\mu^*)) < \gamma\right\}$ , which is a non-empty set (since  $\mu^* < 1$ ); by continuity of the mapping  $\mathcal{K}_{\cdot, \mu^*}$  defined after the statement of Lemma 2 when  $\mu^* \in (0, 1)$ , its closure is  $\overline{\mathcal{C}}_\gamma = \left\{\beta(q) \in \mathcal{B} : \mathcal{K}(\beta(q), \beta(\mu^*)) \leq \gamma\right\}$ .

In addition, since  $\mu_a \in (0, 1)$ , we have that  $\mathcal{K}(\beta(q), \beta(\mu_a)) < \infty$  for all  $q \in [0, 1]$ . In particular, for all  $\gamma > 0$ , the condition  $\Lambda(\mathcal{C}_\gamma) < \infty$  of Lemma 1 is satisfied. Denoting this value by

$$\theta_a(\gamma) = \inf\left\{\mathcal{K}(\beta(q), \beta(\mu_a)) : \beta(q) \in \mathcal{B} \text{ such that } \mathcal{K}(\beta(q), \beta(\mu^*)) \leq \gamma\right\},$$

we get by the indicated lemma that for all  $k \geq 1$ ,

$$\mathbb{P}\left\{\mathcal{K}\left(\beta(\widetilde{\mu}_{a,k}), \beta(\mu^*)\right) \leq \gamma\right\} = \mathbb{P}\left\{\beta(\widetilde{\mu}_{a,k}) \in \overline{\mathcal{C}}_\gamma\right\} \leq e^{-k \theta_a(\gamma)}.$$

Now, since (an open neighborhood of)  $\beta(\mu_a)$  is not included in  $\bar{\mathcal{C}}_\gamma$  as soon as  $0 < \gamma < \mathcal{K}(\beta(\mu_a), \beta(\mu^*))$ , we have that  $\theta_a(\gamma) > 0$  for such values of  $\gamma$ . To apply the obtained inequality to the last sum in (4), we fix a constant  $c_a > 0$  and denote by  $k_0$  the following upper integer part,  $k_0 = \left\lceil \frac{(1+c_a) f(T)}{\mathcal{K}(\beta(\mu_a), \beta(\mu^*))} \right\rceil$ , so that  $f(T)/k \leq \mathcal{K}(\beta(\mu_a), \beta(\mu^*))/(1+c_a) < \mathcal{K}(\beta(\mu_a), \beta(\mu^*))$  for  $k \geq k_0$ , hence,

$$\begin{aligned} \sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{P}\left\{ (k-1) \mathcal{K}(\beta(\tilde{\mu}_{a,k-1}), \beta(\mu^*)) \leq f(T) \right\} &\leq \sum_{k=1}^T \mathbb{P}\left\{ \mathcal{K}(\beta(\tilde{\mu}_{a,k}), \beta(\mu^*)) \leq \frac{f(T)}{k} \right\} \\ &\leq k_0 - 1 + \sum_{k=k_0}^T \exp\left(-k \theta_a(f(T)/k)\right). \end{aligned}$$

Since  $\theta_a$  is a non-increasing function,

$$\begin{aligned} \sum_{k=k_0}^T \exp\left(-k \theta_a(f(T)/k)\right) &\leq \sum_{k=k_0}^T \exp\left(-k \theta_a(\mathcal{K}(\beta(\mu_a), \beta(\mu^*))/(1+c_a))\right) \\ &\leq \Gamma_a(c_a) \exp\left(-k_0 \theta_a(\mathcal{K}(\beta(\mu_a), \beta(\mu^*))/(1+c_a))\right) \leq \Gamma_a(c_a), \end{aligned}$$

where  $\Gamma_a(c_a) = \left[1 - \exp\left(-\theta_a(\mathcal{K}(\beta(\mu_a), \beta(\mu^*))/(1+c_a))\right)\right]^{-1}$ .

Putting all pieces together, we thus proved so far that

$$\mathbb{E}[N_T(a)] \leq 1 + \frac{(1+c_a) f(T)}{\mathcal{K}(\beta(\mu_a), \beta(\mu^*))} + 4e \sum_{t=|\mathcal{A}|}^{T-1} \lceil f(t) \log t \rceil e^{-f(t)} + \Gamma_a(c_a)$$

and it only remains to deal with  $\Gamma_a(c_a)$ .

**Step 5: Getting an upper bound in closed form for  $\Gamma_a(c_a)$ .** We will make repeated uses of Pinsker's inequality: for  $p, q \in [0, 1]$ , one has  $\mathcal{K}(\beta(p), \beta(q)) \geq 2(p-q)^2$ . In what follows, we use the short-hand notation  $\Theta_a = \theta_a(\mathcal{K}(\beta(\mu_a), \beta(\mu^*))/(1+c_a))$  and therefore need to upper bound  $1/(1-e^{-\Theta_a})$ . Since for all  $u \geq 0$ , one has  $e^{-u} \leq 1-u+u^2/2$ , we get  $\Gamma_a(c_a) \leq \frac{1}{\Theta_a(1-\Theta_a/2)} \leq \frac{2}{\Theta_a}$  for  $\Theta_a \leq 1$ , and  $\Gamma_a(c_a) \leq \frac{1}{1-e^{-1}} \leq 2$  for  $\Theta_a \geq 1$ . It thus only remains to lower bound  $\Theta_a$  in the case when it is smaller than 1.

By the continuity properties of the Kullback-Leibler divergence, the infimum in the definition of  $\theta_a$  is always achieved; we therefore let  $\tilde{\mu}$  be an element in  $[0, 1]$  such that

$$\Theta_a = \mathcal{K}(\beta(\tilde{\mu}), \beta(\mu_a)) \quad \text{and} \quad \mathcal{K}(\beta(\tilde{\mu}), \beta(\mu^*)) = \frac{\mathcal{K}(\beta(\mu_a), \beta(\mu^*))}{1+c};$$

it is easy to see that we have the ordering  $\mu_a < \tilde{\mu} < \mu^*$ . By Pinsker's inequality,  $\Theta_a \geq 2(\tilde{\mu} - \mu_a)^2$  and we now lower bound the latter quantity. We use the short-hand notation  $f(p) = \mathcal{K}(\beta(p), \beta(\mu^*))$  and note that the thus defined mapping  $f$  is convex and differentiable on  $(0, 1)$ ; its derivative equals  $f'(p) = \log((1-\mu^*)/(\mu^*)) + \log(p/(1-p))$  for all  $p \in (0, 1)$  and

is therefore non positive for  $p \leq \mu^*$ . By the indicated convexity of  $f$ , using a sub-gradient inequality, we get  $f(\tilde{\mu}) - f(\mu_a) \geq f'(\mu_a)(\tilde{\mu} - \mu_a)$ , which entails, since  $f'(\mu_a) < 0$ ,

$$\tilde{\mu} - \mu_a \geq \frac{f(\tilde{\mu}) - f(\mu_a)}{f'(\mu_a)} = \frac{c_a}{1 + c_a} \frac{f(\mu_a)}{-f'(\mu_a)}, \quad (5)$$

where the equality follows from the fact that by definition of  $\mu$ , we have  $f(\tilde{\mu}) = f(\mu_a)/(1 + c_a)$ . Now, since  $f'$  is differentiable as well on  $(0, 1)$  and takes the value 0 at  $\mu^*$ , a Taylor's equality entails that there exists a  $\xi \in (\mu_a, \mu^*)$  such that

$$-f'(\mu_a) = f'(\mu^*) - f'(\mu_a) = f''(\xi)(\mu^* - \mu_a) \text{ where } f''(\xi) = 1/\xi + 1/(1 - \xi) = 1/(\xi(1 - \xi)).$$

Therefore, by convexity of  $\tau \mapsto \tau(1 - \tau)$ , we get that

$$\frac{1}{-f'(\mu_a)} \geq \frac{\min\{\mu_a(1 - \mu_a), \mu^*(1 - \mu^*)\}}{\mu^* - \mu_a}.$$

Substituting this into (5) and using again Pinsker's inequality to lower bound  $f(\mu_a)$ , we have proved

$$\tilde{\mu} - \mu_a \geq 2 \frac{c_a}{1 + c_a} (\mu^* - \mu_a) \min\{\mu_a(1 - \mu_a), \mu^*(1 - \mu^*)\}.$$

Putting all pieces together, we thus proved that

$$\Gamma_a(c_a) \leq 2 \max \left\{ \frac{(1 + c_a)^2}{8 c_a^2 (\mu^* - \mu_a)^2 \left( \min\{\mu_a(1 - \mu_a), \mu^*(1 - \mu^*)\} \right)^2}, 1 \right\};$$

bounding the maximum of the two quantities by their sum concludes the main part of the proof.

**Step 6: For  $\mu^* \in \{0, 1\}$  and/or  $\mu_a = 0$ .** When  $\mu^* = 1$ , then  $\hat{\mu}_{a^*, N_t(a^*)} = 1$  for all  $t \geq |\mathcal{A}| + 1$ , so that  $B_{a^*, t}^+ = 1$  for all  $t \geq |\mathcal{A}| + 1$ . Thus, the arm  $a$  is played after round  $t \geq |\mathcal{A}| + 1$  only if  $B_{a, t}^+ = 1$  and  $\hat{\mu}_{a, N_t(a)} = 1$  (in view of the tie-breaking rule of the considered strategy). But this means that  $a$  is played as long as it gets payoffs equal to 1 and is stopped being played when it receives the payoff 0 for the first time. Hence, in this case, we have that the sum of payoffs equals at least  $T - 2(|\mathcal{A}| - 1)$  and the regret  $R_T = \mathbb{E}[T\mu^* - (Y_1 + \dots + Y_t)]$  is therefore bounded by  $2(|\mathcal{A}| - 1)$ .

When  $\mu^* = 0$ , a Dirac mass over 0 is associated with all arms and the regret of all strategies is equal to 0.

We consider now the case  $\mu^* \in (0, 1)$  and  $\mu_a = 0$ , for which the first three steps go through; only in the upper bound of step 4 we used the fact that  $\mu_a > 0$ . But in this case, we have a deterministic bound on (4). Indeed, since  $\mathcal{K}(\beta(0), \beta(\mu^*)) = -\log \mu^*$ , we have  $k \mathcal{K}(\beta(0), \beta(\mu^*)) \leq f(T)$  if and only if

$$k \leq \frac{f(T)}{-\log \mu^*} = \frac{f(T)}{\mathcal{K}(\beta(\mu_a), \beta(\mu^*))},$$

which improves on the general bound exhibited in step 4. ■

**Remark 4** Note that Step 5 in the proof is specifically designed to provide an upper bound on  $\Gamma_a(c_a)$  in the case of Bernoulli distributions. In the general case, getting such an explicit bound seems more involved.

#### 4. A finite-time analysis in the case of distributions with finite support

Before stating and proving our main result, Theorem 10, we introduce the quantity  $\mathcal{K}_{\inf}$  and list some of its properties.

##### 4.1. Some useful properties of $\mathcal{K}_{\inf}$ and its level sets

We now introduce the key quantity in order to generalize the previous algorithm to handle the case of distributions with finite support. To that end, we introduce  $\mathcal{P}_F([0, 1])$ , the subset of  $\mathcal{P}([0, 1])$  that consists of distributions with finite support.

**Definition 5** For all distributions  $\nu \in \mathcal{P}_F([0, 1])$  and  $\mu \in [0, 1]$ , we define

$$\mathcal{K}_{\inf}(\nu, \mu) = \inf \left\{ \mathcal{K}(\nu, \nu') : \nu' \in \mathcal{P}_F([0, 1]) \text{ s.t. } E(\nu') > \mu \right\},$$

where  $E(\nu') = \int_{[0,1]} x d\nu'(x)$  denotes the expectation of the distribution  $\nu'$ .

We now remind some useful properties of  $\mathcal{K}_{\inf}$ . Honda and Takemura (2010b, Lemma 6) can be reformulated in our context as follows.

**Lemma 6** For all  $\nu \in \mathcal{P}_F([0, 1])$ , the mapping  $\mathcal{K}_{\inf}(\nu, \cdot)$  is continuous and non decreasing in its argument  $\mu \in [0, 1]$ . Moreover, the mapping  $\mathcal{K}_{\inf}(\cdot, \mu)$  is lower semi-continuous on  $\mathcal{P}_F([0, 1])$  for all  $\mu \in [0, 1]$ .

The next two lemmas bound the variation of  $\mathcal{K}_{\inf}$ , respectively in its first and second arguments. (For clarity, we denote the expectations with respect to  $\nu$  by  $\mathbb{E}_\nu$ .) Their proofs can be found in the extended version of the present conference paper (Maillard et al., 2011). We denote by  $\|\cdot\|_1$  the  $\ell^1$ -norm on  $\mathcal{P}([0, 1])$  and recall that the  $\ell^1$ -norm of  $\nu - \nu'$  corresponds to twice the distance in variation between  $\nu$  and  $\nu'$ .

**Lemma 7** For all  $\mu \in (0, 1)$  and for all  $\nu, \nu' \in \mathcal{P}_F([0, 1])$ , the following holds true.

- In the case when  $\mathbb{E}_\nu[(1-\mu)/(1-X)] > 1$ , then  $\mathcal{K}_{\inf}(\nu, \mu) - \mathcal{K}_{\inf}(\nu', \mu) \leq M_{\nu, \mu} \|\nu - \nu'\|_1$ , for some constant  $M_{\nu, \mu} > 0$ .
- In the case when  $\mathbb{E}_\nu[(1-\mu)/(1-X)] \leq 1$ , the fact that  $\mathcal{K}_{\inf}(\nu, \mu) - \mathcal{K}_{\inf}(\nu', \mu) \geq \alpha \mathcal{K}_{\inf}(\nu, \mu)$  for some  $\alpha \in (0, 1)$  entails that

$$\|\nu - \nu'\|_1 \geq \frac{1 - \mu}{(2/\alpha)((2/\alpha) - 1)}.$$

**Lemma 8** We have that for any  $\nu \in \mathcal{P}_F([0, 1])$ , provided that  $\mu \geq \mu - \varepsilon > E(\nu)$ , the following inequalities hold true:

$$\varepsilon/(1 - \mu) \geq \mathcal{K}_{\inf}(\nu, \mu) - \mathcal{K}_{\inf}(\nu, \mu - \varepsilon) \geq 2\varepsilon^2$$

Moreover, the first inequality is also valid when  $E(\nu) \geq \mu > \mu - \varepsilon$  or  $\mu > E(\nu) \geq \mu - \varepsilon$ .

**Level sets of  $\mathcal{K}_{\inf}$ :** For each  $\gamma > 0$  and  $\mu \in (0, 1)$ , we consider the set

$$\begin{aligned} \mathcal{C}_{\mu, \gamma} &= \left\{ \nu' \in \mathcal{P}_F([0, 1]) : \mathcal{K}_{\inf}(\nu', \mu) < \gamma \right\} \\ &= \left\{ \nu' \in \mathcal{P}_F([0, 1]) : \exists \nu'_\mu \in \mathcal{P}_F([0, 1]) \text{ s.t. } E(\nu'_\mu) > \mu \text{ and } \mathcal{K}(\nu', \nu'_\mu) < \gamma \right\}. \end{aligned}$$

We detail a property in the following lemma, whose proof can be found in the extended version of the present conference paper (Maillard et al., 2011).

**Lemma 9** *For all  $\gamma > 0$  and  $\mu \in (0, 1)$ , the set  $\mathcal{C}_{\mu, \gamma}$  is a non-empty open convex set. Moreover,*

$$\bar{\mathcal{C}}_{\mu, \gamma} \supseteq \left\{ \nu' \in \mathcal{P}_F([0, 1]) : \mathcal{K}_{\inf}(\nu', \mu) \leq \gamma \right\}.$$

#### 4.2. The $\mathcal{K}_{\inf}$ -strategy and a general performance guarantee

For each arm  $a \in \mathcal{A}$  and round  $t$  with  $N_t(a) > 0$ , we denote by  $\widehat{\nu}_{a, N_t(a)}$  the empirical distribution of the payoffs obtained till round  $t$  when picking arm  $a$ , that is,

$$\widehat{\nu}_{a, N_t(a)} = \frac{1}{N_t(a)} \sum_{s \leq t: A_s=a} \delta_{Y_s},$$

where for all  $x \in [0, 1]$ , we denote by  $\delta_x$  the Dirac mass on  $x$ . We define the corresponding empirical averages as

$$\widehat{\mu}_{a^*, N_t(a^*)} = E(\widehat{\nu}_{a^*, N_t(a^*)}) = \frac{1}{N_t(a)} \sum_{s \leq t: A_s=a} Y_s.$$

We then consider the  $\mathcal{K}_{\inf}$ -strategy defined in Figure 2. Note that the use of maxima in the definitions of the  $B_{a,t}^+$  is justified by Lemma 6.

As explained in Honda and Takemura (2010b), the computation of the quantities  $\mathcal{K}_{\inf}$  can be done efficiently in this case, i.e., when we consider only distributions with finite supports. This is because in the computation of  $\mathcal{K}_{\inf}$ , it is sufficient to consider only distributions with the same support as the empirical distributions (up to one point). Note that the knowledge of the support of the distributions associated with the arms is not required.

**Theorem 10** *Assume that  $\nu^*$  is finitely supported, with expectation  $\mu^* \in (0, 1)$  and with support denoted by  $\mathcal{S}^*$ . Let  $a \in \mathcal{A}$  be a suboptimal arm such that  $\mu_a > 0$  and  $\nu_a$  is finitely supported. Then, for all  $c_a > 0$  and all*

$$0 < \varepsilon < \min \left\{ \Delta_a, \frac{c_a/2}{1+c_a} (1 - \mu^*) \mathcal{K}_{\inf}(\nu_a, \mu^*) \right\},$$

*the expected number of times the  $\mathcal{K}_{\inf}$ -strategy, run with  $f(t) = \log t$ , pulls arm  $a$  satisfies*

$$\begin{aligned} \mathbb{E}[N_T(a)] &\leq 1 + \frac{(1 + c_a) \log T}{\mathcal{K}_{\inf}(\nu_a, \mu^*)} + \frac{1}{1 - e^{-\Theta_a(c_a, \varepsilon)}} + \frac{1}{\varepsilon^2} \log \left( \frac{1}{1 - \mu^* + \varepsilon} \right) \sum_{k=1}^T (k+1)^{|\mathcal{S}^*|} e^{-k\varepsilon^2} \\ &\quad + \frac{1}{(\Delta_a - \varepsilon)^2}, \end{aligned}$$

---

*Parameters:* A non-decreasing function  $f : \mathbb{N} \rightarrow \mathbb{R}$

*Initialization:* Pull each arm of  $\mathcal{A}$  once

*For* rounds  $t + 1$ , where  $t \geq |\mathcal{A}|$ ,

- compute for each arm  $a \in \mathcal{A}$  the quantity

$$B_{a,t}^+ = \max \left\{ q \in [0, 1] : N_t(a) \mathcal{K}_{\inf}(\widehat{\nu}_{a,N_t(a)}, q) \leq f(t) \right\},$$

where  $\widehat{\nu}_{a,N_t(a)} = \frac{1}{N_t(a)} \sum_{s \leq t: A_s=a} \delta_{Y_s}$ ;

- in case of a tie, pick an arm with largest value of  $\widehat{\mu}_{a,N_t(a)}$ ;
  - pull any arm  $A_{t+1} \in \operatorname{argmax}_{a \in \mathcal{A}} B_{a,t}^+$ .
- 

Figure 2: The strategy  $\mathcal{K}_{\inf}$ .

where

$$\Theta_a(c_a, \varepsilon) = \theta_a \left( \frac{\log T}{k_0} + \frac{\varepsilon}{1 - \mu^*} \right) \quad \text{with} \quad k_0 = \left\lceil \frac{(1 + c_a) \log T}{\mathcal{K}_{\inf}(\nu_a, \mu^*)} \right\rceil.$$

and for all  $\gamma > 0$ ,

$$\theta_a(\gamma) = \inf \left\{ \mathcal{K}(\nu', \nu_a) : \nu' \text{ s.t. } \mathcal{K}_{\inf}(\nu', \mu^*) < \gamma \right\}.$$

As a corollary, we get (by taking some common value for all  $c_a$ ) that for all  $c > 0$ ,

$$\overline{R}_T \leq \sum_{a \in \mathcal{A}} \Delta_a \frac{(1 + c) \log T}{\mathcal{K}_{\inf}(\nu_a, \mu^*)} + h(c),$$

where  $h(c) < \infty$  is a function of  $c$  (and of the distributions associated with the arms), which is however independent of  $T$ . As a consequence, we recover the asymptotic results of Burnetas and Katehakis (1996); Honda and Takemura (2010a), i.e., the guarantee that

$$\limsup_{T \rightarrow \infty} \frac{\overline{R}_T}{\log T} \leq \sum_{a \in \mathcal{A}} \frac{\Delta_a}{\mathcal{K}_{\inf}(\nu_a, \mu^*)}.$$

Of course, a sharper optimization can be performed by carefully choosing the constants  $c_a$ , that are parameters of the analysis; similarly to the comments after the statement of Theorem 3, we would then get a dominant term with a constant factor 1 instead of  $1 + c$  as above, plus an additional second-order term. Details are left to a journal version of this paper.

**Proof** By arguments similar to the ones used in the first step of the proof of Theorem 3, we have

$$\{A_{t+1} = a\} \subseteq \left\{ \mu^* - \varepsilon < \widehat{\mu}_{a,N_t(a)} \right\} \cup \left\{ \mu^* - \varepsilon > B_{a^*,t}^+ \right\} \cup \left\{ \mu^* - \varepsilon \in [\widehat{\mu}_{a,N_t(a)}, B_{a,t}^+] \right\};$$

indeed, on the event  $\{A_{t+1} = a\} \cap \{\mu^* - \varepsilon \geq \widehat{\mu}_{a,N_t(a)}\} \cap \{\mu^* - \varepsilon \leq B_{a^*,t}^+\}$ , we have,  $\widehat{\mu}_{a,N_t(a)} \leq \mu^* - \varepsilon \leq B_{a^*,t}^+ \leq B_{a,t}^+$  (where the last inequality is by definition of the strategy). Before proceeding, we note that

$$\left\{ \mu^* - \varepsilon \in [\widehat{\mu}_{a,N_t(a)}, B_{a,t}^+] \right\} \subseteq \left\{ N_t(a) \mathcal{K}_{\inf}(\widehat{\nu}_{a,N_t(a)}, \mu^* - \varepsilon) \leq f(t) \right\},$$

since  $\mathcal{K}_{\inf}$  is a non-decreasing function in its second argument and  $\mathcal{K}_{\inf}(\nu, E(\nu)) = 0$  for all distributions  $\nu$ . Therefore,

$$\begin{aligned} \mathbb{E}[N_T(a)] &\leq 1 + \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\left\{ \mu^* - \varepsilon < \widehat{\mu}_{a,N_t(a)} \text{ and } A_{t+1} = a \right\} + \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\left\{ \mu^* - \varepsilon > B_{a^*,t}^+ \right\} \\ &\quad + \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\left\{ N_t(a) \mathcal{K}_{\inf}(\widehat{\nu}_{a,N_t(a)}, \mu^* - \varepsilon) \leq f(t) \text{ and } A_{t+1} = a \right\}; \end{aligned}$$

now, the two sums with the events “and  $A_{t+1} = a$ ” can be rewritten by using the stopping times  $\tau_{a,k}$  introduced in the proof of Theorem 3; more precisely, by mimicking the transformations performed in its step 3, we get the simpler bound

$$\begin{aligned} \mathbb{E}[N_T(a)] &\leq 1 + \sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{P}\left\{ \mu^* - \varepsilon < \widetilde{\mu}_{a,k-1} \right\} + \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\left\{ \mu^* - \varepsilon > B_{a^*,t}^+ \right\} \\ &\quad + \sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{P}\left\{ (k-1) \mathcal{K}_{\inf}(\widetilde{\nu}_{a,k-1}, \mu^* - \varepsilon) \leq f(t) \right\}, \quad (6) \end{aligned}$$

where the  $\widetilde{\nu}_{a,s}$  and  $\widetilde{\mu}_{a,s}$  are respectively the empirical distributions and empirical expectations computed on the first  $s$  elements of the sequence of the random variables  $\widetilde{X}_{a,j} = Y_{\tau_{a,j}}$ , which are i.i.d. according to  $\nu_a$ .

**Step 1: The first sum in (6)** is bounded by resorting to Hoeffding’s inequality, whose application is legitimate since  $\mu^* - \mu_a - \varepsilon > 0$ ;

$$\begin{aligned} \sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{P}\left\{ \mu^* - \varepsilon < \widetilde{\mu}_{a,k-1} \right\} &= \sum_{k=1}^{T-|\mathcal{A}|} \mathbb{P}\left\{ \mu^* - \mu_a - \varepsilon < \widetilde{\mu}_{a,k} - \mu_a \right\} \\ &\leq \sum_{k=1}^{T-|\mathcal{A}|} e^{-2k(\mu^* - \mu_a - \varepsilon)^2} \leq \frac{1}{1 - e^{-2(\mu^* - \mu_a - \varepsilon)^2}} \leq \frac{1}{(\mu^* - \mu_a - \varepsilon)^2} \end{aligned}$$

where we used for the last inequality the general upper bounds provided at the beginning of step 5 in the proof of Theorem 3.

**Step 2: The second sum in (6)** is bounded by first using the definition of  $B_{a^*,t}^+$ , then, decomposing the event depending on the values taken by  $N_t(a^*)$ ; and finally using the fact that on  $\{N_t(a^*) = k\}$ , we have the rewriting  $\widehat{\nu}_{a,N_t(a)} = \widetilde{\nu}_{a,k}$  and  $\widehat{\mu}_{a,N_t(a)} = \widetilde{\mu}_{a,k}$ ;

more precisely,

$$\begin{aligned}
\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\left\{\mu^* - \varepsilon > B_{a^*, t}^+\right\} &\leq \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\left\{N_t(a^*) \mathcal{K}_{\inf}(\widehat{\nu}_{a^*, N_t(a^*)}, \mu^* - \varepsilon) > f(t)\right\} \\
&= \sum_{t=|\mathcal{A}|}^{T-1} \sum_{k=1}^t \mathbb{P}\left\{N_t(a^*) = k \text{ and } k \mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^* - \varepsilon) > f(t)\right\} \\
&\leq \sum_{k=1}^T \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\left\{k \mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^* - \varepsilon) > f(t)\right\}.
\end{aligned}$$

Since  $f = \log$  is increasing, we can rewrite the bound, using a Fubini-Tonelli argument, as

$$\begin{aligned}
\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\left\{\mu^* - \varepsilon > B_{a^*, t}^+\right\} &\leq \sum_{k=1}^T \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\left\{f^{-1}(k \mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^* - \varepsilon)) > t\right\} \\
&\leq \sum_{k=1}^T \mathbb{E}\left[f^{-1}(k \mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^* - \varepsilon)) \mathbb{I}_{\{\mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^* - \varepsilon) > 0\}}\right].
\end{aligned}$$

Now, Honda and Takemura (2010a, Lemma 13) indicates that, since  $\mu^* - \varepsilon \in [0, 1]$ ,

$$\sup_{\nu \in \mathcal{P}_F([0, 1])} \mathcal{K}_{\inf}(\nu, \mu^* - \varepsilon) \leq \log(1/(1 - \mu^* + \varepsilon)) \stackrel{\text{def}}{=} K_{\max};$$

we define  $Q = K_{\max}/\varepsilon^2$  and introduce the following sets  $(V_q)_{1 \leq q \leq Q}$ :

$$V_q = \left\{\nu \in \mathcal{P}_F([0, 1]) : (q-1)\varepsilon^2 < \mathcal{K}_{\inf}(\nu, \mu^* - \varepsilon) \leq q\varepsilon^2\right\}.$$

A peeling argument (and by using that  $f^{-1} = \exp$  is increasing as well) entails, for all  $k \geq 1$ ,

$$\begin{aligned}
&\mathbb{E}\left[f^{-1}(k \mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^* - \varepsilon)) \mathbb{I}_{\{\mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^* - \varepsilon) > 0\}}\right] \tag{7} \\
&= \sum_{q=1}^Q \mathbb{E}\left[f^{-1}(k \mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^* - \varepsilon)) \mathbb{I}_{\{\widetilde{\nu}_{a^*, k} \in V_q\}}\right] \\
&\leq \sum_{q=1}^Q \mathbb{P}\{\widetilde{\nu}_{a^*, k} \in V_q\} f^{-1}(kq\varepsilon^2) \leq \sum_{q=1}^Q \mathbb{P}\left\{\mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^* - \varepsilon) > (q-1)\varepsilon^2\right\} f^{-1}(kq\varepsilon^2) \tag{8}
\end{aligned}$$

where we used the definition of  $V_q$  to obtain each of the two inequalities. Now, by Lemma 8, when  $E(\widetilde{\nu}_{a^*, k}) < \mu^* - \varepsilon$ , which is satisfied whenever  $\mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^* - \varepsilon) > 0$ , we have

$$\mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^* - \varepsilon) \leq \mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^*) - 2\varepsilon^2 \leq \mathcal{K}(\widetilde{\nu}_{a^*, k}, \nu^*) - 2\varepsilon^2,$$

where the last inequality is by mere definition of  $\mathcal{K}_{\inf}$ . Therefore,

$$\mathbb{P}\left\{\mathcal{K}_{\inf}(\widetilde{\nu}_{a^*, k}, \mu^* - \varepsilon) > (q-1)\varepsilon^2\right\} \leq \mathbb{P}\left\{\mathcal{K}(\widetilde{\nu}_{a^*, k}, \nu^*) > (q+1)\varepsilon^2\right\}.$$

We note that for all  $k \geq 1$ ,  $\mathbb{P}\left\{\mathcal{K}(\tilde{\nu}_{a^*,k}, \nu^*) > (q+1)\varepsilon^2\right\} \leq (k+1)^{|\mathcal{S}^*|} e^{-k(q+1)\varepsilon^2}$ , where we recall that  $\mathcal{S}^*$  denotes the finite support of  $\nu^*$  and where we applied the method of types; see, e.g., the extended version of the present paper (Maillard et al., 2011) for more details about this standard inequality. Now, (8) then yields, via the choice  $f = \log$  and thus  $f^{-1} = \exp$ , that

$$\mathbb{E}\left[f^{-1}\left(k \mathcal{K}_{\inf}(\tilde{\nu}_{a^*,k}, \mu^* - \varepsilon)\right) \mathbb{I}_{\{\mathcal{K}_{\inf}(\tilde{\nu}_{a^*,k}, \mu^* - \varepsilon) > 0\}}\right] \leq \underbrace{\sum_{q=1}^Q (k+1)^{|\mathcal{S}^*|} e^{-k(q+1)\varepsilon^2} e^{kq\varepsilon^2}}_{=Q(k+1)^{|\mathcal{S}^*|} e^{-k\varepsilon^2}}.$$

Substituting the value of  $Q$ , we therefore have proved that

$$\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\left\{\mu^* - \varepsilon > B_{a^*,t}^+\right\} \leq \frac{1}{\varepsilon^2} \log\left(\frac{1}{1 - \mu^* + \varepsilon}\right) \sum_{k=1}^T (k+1)^{|\mathcal{S}^*|} e^{-k\varepsilon^2}.$$

**Step 3: The third sum in (6)** is first upper bounded by Lemma 8, which states that

$$\mathcal{K}_{\inf}(\tilde{\nu}_{a,k-1}, \mu^*) - \varepsilon/(1 - \mu^*) \leq \mathcal{K}_{\inf}(\tilde{\nu}_{a,k-1}, \mu^* - \varepsilon)$$

for all  $k \geq 1$ , and by using  $f(t) \leq f(T)$ ; this gives

$$\begin{aligned} & \sum_{k=1}^{T-|\mathcal{A}|} \mathbb{P}\left\{k \mathcal{K}_{\inf}(\tilde{\nu}_{a,k}, \mu^* - \varepsilon) \leq f(T)\right\} \\ & \leq \sum_{k=1}^{T-|\mathcal{A}|} \mathbb{P}\left\{k \mathcal{K}_{\inf}(\tilde{\nu}_{a,k}, \mu^*) \leq f(T) + \frac{k\varepsilon}{1 - \mu^*}\right\} = \sum_{k=1}^{T-|\mathcal{A}|} \mathbb{P}\left\{\tilde{\nu}_{a,k} \in \bar{\mathcal{C}}_{\mu^*, \gamma_k}\right\}, \end{aligned}$$

where  $\gamma_k = f(T)/k + \varepsilon/(1 - \mu^*)$  and where the set  $\bar{\mathcal{C}}_{\mu^*, \gamma_k}$  was defined in Section 4.1. For all  $\gamma > 0$ , we then introduce

$$\theta_a(\gamma) = \inf\left\{\mathcal{K}(\nu', \nu_a) : \nu' \in \mathcal{C}_{\mu^*, \gamma}\right\} = \inf\left\{\mathcal{K}(\nu', \nu_a) : \nu' \in \bar{\mathcal{C}}_{\mu^*, \gamma}\right\},$$

(where the second equality follows from the lower semi-continuity of  $\mathcal{K}$ ) and aim at bounding  $\mathbb{P}\left\{\tilde{\nu}_{a,k} \in \bar{\mathcal{C}}_{\mu^*, \gamma}\right\}$ .

As shown in Section 4.1, the set  $\mathcal{C}_{\mu^*, \gamma}$  is a non-empty open convex set. If we prove that  $\theta_a(\gamma)$  is finite for all  $\gamma > 0$ , then all the conditions will be required to apply Lemma 1 and get the upper bound

$$\sum_{k=1}^{T-|\mathcal{A}|} \mathbb{P}\left\{\tilde{\nu}_{a,k} \in \bar{\mathcal{C}}_{\mu^*, \gamma_k}\right\} \leq \sum_{k=1}^{T-|\mathcal{A}|} e^{-k\theta_a(\gamma_k)}.$$

To that end, we use the fact that  $\nu_a$  is finitely supported. Now, either the probability of interest is null and we are done; or, it is not null, which implies that there exists a possible value of  $\tilde{\nu}_{a,k}$  that is in  $\bar{\mathcal{C}}_{\mu^*, \gamma}$ ; since this value is a distribution with a support included in

the one of  $\nu_a$ , it is absolutely continuous with respect to  $\nu_a$  and hence, the Kullback-Leibler divergence between this value and  $\nu_a$  is finite; in particular,  $\theta_a(\gamma)$  is finite.

Finally, we bound the  $\theta_a(\gamma_k)$  for values of  $k$  larger than  $k_0 = \left\lceil \frac{(1+c_a)f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^*)} \right\rceil$ ; we have that for all  $k \geq k_0$ , in view of the bound put on  $\varepsilon$ ,

$$\gamma_k \leq \gamma_{k_0} = \frac{f(T)}{k_0} + \frac{\varepsilon}{1-\mu^*} < \frac{\mathcal{K}_{\inf}(\nu_a, \mu^*)}{1+c_a} + \frac{c_a/2}{1+c_a} \mathcal{K}_{\inf}(\nu_a, \mu^*) = \frac{1+c_a/2}{1+c_a} \mathcal{K}_{\inf}(\nu_a, \mu^*). \quad (9)$$

Since  $\theta_a$  is non increasing, we have

$$\sum_{k=1}^{T-|\mathcal{A}|} e^{-k\theta_a(\gamma_k)} \leq k_0 - 1 + \sum_{k=k_0}^{T-|\mathcal{A}|} e^{-k\theta_a(\gamma_{k_0})} \leq k_0 - 1 + \frac{1}{1-e^{-\Theta_a(c_a, \varepsilon)}},$$

provided that the quantity  $\Theta_a(c_a, \varepsilon) = \theta_a(\gamma_{k_0})$  is positive, which we prove now.

Indeed for all  $\nu' \in \mathcal{C}_{\mu^*, \gamma_{k_0}}$ , we have by definition and by (9) that

$$\mathcal{K}_{\inf}(\nu', \mu^*) - \mathcal{K}_{\inf}(\nu_a, \mu^*) < \gamma_{k_0} - \mathcal{K}_{\inf}(\nu_a, \mu^*) < -((c_a/2)/(1+c_a)) \mathcal{K}_{\inf}(\nu_a, \mu^*).$$

Now, in the case where  $\mathbb{E}_{\nu_a}[(1-\mu^*)/(1-X)] > 1$ , we have, first by application of Pinsker's inequality and then by Lemma 7, that

$$\mathcal{K}(\nu', \nu_a) \geq \frac{\|\nu' - \nu_a\|_1^2}{2} \geq \frac{1}{2M_{\nu_a, \mu^*}^2} (\mathcal{K}_{\inf}(\nu_a, \mu^*) - \mathcal{K}_{\inf}(\nu', \mu^*))^2 > \frac{c_a^2 (\mathcal{K}_{\inf}(\nu_a, \mu^*))^2}{8(1+c_a)^2 M_{\nu_a, \mu^*}^2};$$

since, again by Pinsker's inequality,  $\mathcal{K}_{\inf}(\nu_a, \mu^*) \geq (\mu_a - \mu^*)^2/2 > 0$ , we have exhibited a lower bound independent of  $\nu'$  in this case. In the case where  $\mathbb{E}_{\nu_a}[(1-\mu^*)/(1-X)] \leq 1$ , we apply the second part of Lemma 7, with  $\alpha_a = (c_a/2)/(1+c_a)$ , and get

$$\mathcal{K}(\nu', \nu_a) \geq \frac{\|\nu' - \nu_a\|_1^2}{2} \geq \frac{1}{2} \left( \frac{1-\mu^*}{(2/\alpha_a)((2/\alpha_a)-1)} \right)^2 > 0.$$

Thus, in both cases we found a positive lower bound independent of  $\nu'$ , so that the infimum over  $\nu' \in \mathcal{C}_{\mu^*, \gamma_{k_0}}$  of the quantities  $\mathcal{K}_{\inf}(\nu', \mu^*)$ , which precisely equals  $\theta_a(\gamma_{k_0})$ , is also positive. This concludes the proof.  $\blacksquare$

**Conclusion.** We provided a finite-time analysis of the (asymptotically optimal)  $\mathcal{K}_{\inf}$ -strategy in the case of finitely supported distributions. The extension to the case of general distributions (e.g., by histogram-based approximations of such general distributions) is left for future work.

## Acknowledgments

The authors wish to thank *Peter Auer* and *Daniil Ryabko* for insightful discussions. They acknowledge support from the French National Research Agency (ANR-08-COSI-004 project EXPLO-RA), the European Community's Seventh Framework Programme (project COM-PLACS, grant agreement no. 231495), and the PASCAL2 Network of Excellence (EC grant no. 506778).

## References

- J-Y. Audibert, R. Munos, and C. Szepesvari. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.
- J.Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2635–2686, 2010.
- P. Auer and R. Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.
- P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.
- A.N. Burnetas and M.N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.
- Y. Chow and H. Teicher. *Probability Theory*. Springer, 1988.
- I.H. Dinwoodie. Mesures dominantes et théorème de Sanov. *Annales de l’Institut Henri Poincaré – Probabilités et Statistiques*, 28(3):365–373, 1992.
- S. Filippi. *Stratégies optimistes en apprentissage par renforcement*. PhD thesis, Télécom ParisTech, 2010.
- A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of COLT*, 2011.
- A. Garivier and F. Leonardi. Context tree selection: A unifying view. *Stochastic Processes and their Applications*, 2011. In press.
- J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of COLT*, pages 67–79, 2010a.
- J. Honda and A. Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. arXiv:0905.2776, 2010b.
- T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.
- O.-A. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. 2011. URL <http://hal.archives-ouvertes.fr/inria-00574987/>.
- H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

# Robust approachability and regret minimization in games with partial monitoring

**Shie Mannor**

*Technion, Haifa  
Israel*

SHIE@EE.TECHNION.AC.IL

**Vianney Perchet**

*Ecole normale supérieure, Cachan  
France*

VIANNEY.PERCHET@NORMALESUP.ORG

**Gilles Stoltz**

*Ecole Normale Supérieure – CNRS – INRIA, Paris  
France  
&  
HEC Paris – CNRS, Jouy-en-Josas  
France*

GILLES.STOLTZ@ENS.FR

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

Approachability has become a standard tool in analyzing learning algorithms in the adversarial online learning setup. We develop a variant of approachability for games where there is ambiguity in the obtained reward that belongs to a set, rather than being a single vector. Using this variant we tackle the problem of approachability in games with partial monitoring and develop simple and efficient algorithms (i.e., with constant per-step complexity) for this setup. We finally consider external and internal regret in repeated games with partial monitoring, for which we derive regret-minimizing strategies based on approachability theory.

**Keywords:** Approachability, partial monitoring, regret, adversarial learning

## 1. Introduction

Blackwell’s approachability theory and its variants has become a standard and useful tool in analyzing online learning algorithms (Cesa-Bianchi and Lugosi, 2006) and algorithms for learning in games (Hart and Mas-Colell, 2000, 2001). The first application of Blackwell’s approachability to learning in the online setup is due to Blackwell himself in Blackwell (1956b). Numerous other contributions are summarized in Cesa-Bianchi and Lugosi (2006). Blackwell’s approachability theory enjoys a clear geometric interpretation that allows it to be used in situations where online convex optimization or exponential weights do not seem to be easily applicable and, in some sense, to go beyond the minimization of the regret and/or to control quantities of a different flavor; e.g., in Mannor et al. (2009), to minimize the regret together with path constraints, and in Mannor and Shimkin (2008), to minimize

the regret in games whose stage duration is not fixed. Recently, it has been shown that approachability and low regret learning are equivalent in the sense that efficient reductions exist from one to the other (Abernethy et al., 2011). Another recent paper (Rakhlin et al., 2011) showed that approachability can be analyzed from the perspective of learnability using tools from learning theory.

In this paper we consider approachability and online learning with partial monitoring in games against Nature. In partial monitoring the decision maker does not know how much reward was obtained and only gets a (random) signal whose distribution depends on the action of the decision maker and the action of Nature. There are two extremes of this setup that are well studied. On the one extreme we have the case where the signal includes the reward itself (or a signal that can be used to unbiasedly estimate the reward), which is essentially the celebrated bandits setup. The other extreme is the case where the signal is not informative (i.e., it tells the decision maker nothing about the actual reward obtained); this setting then essentially consists of repeating the same situation over and over again, as no information is gained over time. We consider a setup encompassing these situations and more general ones, in which the signal is indicative of the actual reward, but is not necessarily a sufficient statistics thereof. The difficulty is that the decision maker cannot compute the actual reward he obtained nor the actions of Nature.

Regret minimization with partial monitoring has been studied in several papers in the learning theory community. Piccolboni and Schindelhauer (2001); Mannor and Shimkin (2003); Cesa-Bianchi et al. (2006) study special cases where an accurate estimation of the rewards (or worst-case rewards) of the decision maker is possible thanks to some extra structure. A general policy with vanishing regret is presented in Lugosi et al. (2008). This policy is based on exponential weights and a specific estimation procedure for the (worst-case) obtained rewards. In contrast, we provide approachability-based results for the problem of regret minimization. On route, we define a new type of approachability setup, with enables to re-derive the extension of approachability to the partial monitoring vector-valued setting proposed by Perchet (2011a). More importantly, we provide concrete algorithms for this approachability problem that are more efficient in the sense that, unlike previous works in the domain, their complexity is constant over all steps. Moreover, their rates of convergence are, as in Blackwell (1956b) but for the first time in this general framework, independent of the game at hand. The paper is organized as follows. In Section 2 we recall some basic facts from approachability theory. In Section 3 we propose a novel setup for approachability, termed “robust approachability,” where instead of obtaining a vector-valued reward, the decision maker obtains a set, that represents the ambiguity concerning his reward. We provide a simple characterization of approachable convex sets and an algorithm for the set-valued reward setup. In Section 4 we show how to apply the robust approachability framework to the repeated vector-valued games with partial monitoring. We start in Section 4.1 with the case where the signaling structure is bi-piecewise linear. For this important special case, we provide a simple and constructive algorithm. Previous results for approachability in this setup were either non-constructive (Rustichini, 1999) or were highly inefficient as they relied on some sort of lifting to the space of probability measures on mixed actions (Perchet, 2011a) and typically required a grid that is progressively refined (leading to a step complexity that is exponential in the number  $T$  of past steps). In Section 4.2 we apply our results for both external and internal regret minimization with partial monitoring.

In both cases our proofs are simple, lead to algorithms with constant complexity at each step, and are accompanied with rates. Our results for external regret have rates similar to Lugosi et al. (2008), but our proof is direct and simpler. For internal regret minimization we present the first algorithm not relying on a grid being refined over time and the first convergence rates. In Section 4.3 we mention the general signaling case and explain how it is possible to approach certain special sets such as polytopes efficiently and general convex sets inefficiently.

## 2. Some basic facts from approachability theory

In this section we recall the most basic versions of Blackwell's approachability theorem for vector-valued payoff functions.

We consider a vector-valued game between two players, a decision maker (first player) and Nature (second player), with respective finite action sets  $\mathcal{A}$  and  $\mathcal{B}$ , whose cardinalities are referred to as  $N_{\mathcal{A}}$  and  $N_{\mathcal{B}}$ . We denote by  $d$  the dimension of the reward vector and equip  $\mathbb{R}^d$  with the  $\ell^2$ -norm  $\|\cdot\|_2$ . The payoff function of the first player is given by a mapping  $m : \mathcal{A} \times \mathcal{B} \rightarrow \mathbb{R}^d$ , which is multi-linearly extended to  $\Delta(\mathcal{A}) \times \Delta(\mathcal{B})$ , the set of product-distributions over  $\mathcal{A} \times \mathcal{B}$ .

We consider two frameworks, depending on whether pure or mixed actions are taken.

**Pure actions taken and observed.** We denote by  $A_1, A_2, \dots$  and  $B_1, B_2, \dots$  the actions in  $\mathcal{A}$  and  $\mathcal{B}$  sequentially taken by each player; they are possibly given by randomized strategies, i.e., the actions  $A_t$  and  $B_t$  were obtained by random draws according to respective probability distributions denoted by  $\mathbf{x}_t \in \Delta(\mathcal{A})$  and  $\mathbf{y}_t \in \Delta(\mathcal{B})$ . For now, we assume that the first player has a full monitoring of the pure actions taken by the opponent player: at the end of round  $t$ , when receiving the payoff  $m(A_t, B_t)$ , the pure action  $B_t$  is revealed to him.

**Definition 1** A set  $\mathcal{C} \subseteq \mathbb{R}^d$  is  $m$ -approachable with pure actions if there exists a strategy<sup>1</sup> of the first player such that for all strategies of the second player,

$$\limsup_{T \rightarrow \infty} \inf_{c \in \mathcal{C}} \left\| c - \frac{1}{T} \sum_{t=1}^T m(A_t, B_t) \right\|_2 = 0 \quad a.s.$$

That is, the first player has a strategy that ensures that the average of his vector-valued payoffs converges to the set  $\mathcal{C}$ .

**Mixed actions taken and observed.** In this case, we denote by  $\mathbf{x}_1, \mathbf{x}_2, \dots$  and  $\mathbf{y}_1, \mathbf{y}_2, \dots$  the actions in  $\Delta(\mathcal{A})$  and  $\Delta(\mathcal{B})$  sequentially taken by each player. We also assume a full monitoring for the first player: at the end of round  $t$ , when receiving the payoff  $m(\mathbf{x}_t, \mathbf{y}_t)$ , the mixed action  $\mathbf{y}_t$  is revealed to him.

---

1. The original definition given by Blackwell requires uniformity w.r.t. the strategy set of the opponent. We ignore the uniformity to avoid excessive nomenclature.

**Definition 2** In this context, a set  $\mathcal{C} \subseteq \mathbb{R}^d$  is  $m$ -approachable with mixed actions if there exists a strategy of the first player such that for all strategies of the second player,

$$\limsup_{T \rightarrow \infty} \inf_{c \in \mathcal{C}} \left\| c - \frac{1}{T} \sum_{t=1}^T m(\mathbf{x}_t, \mathbf{y}_t) \right\|_2 = 0 \quad a.s.$$

**Necessary and sufficient condition for approachability.** For closed convex sets there is a simple characterization of approachability that is a direct consequence of the minimax theorem; the condition is the same for the two settings, whether pure or mixed actions are taken and observed.

**Theorem 3 (Blackwell 1956a, Theorem 3)** A closed convex set  $\mathcal{C} \subseteq \mathbb{R}^d$  is approachable (with pure or mixed actions) if and only if

$$\forall \mathbf{y} \in \Delta(\mathcal{B}), \quad \exists \mathbf{x} \in \Delta(\mathcal{A}), \quad m(\mathbf{x}, \mathbf{y}) \in \mathcal{C}.$$

In the latter case, an explicit strategy achieves the following convergence rates. We denote by  $M$  a bound in norm over  $m$ , i.e.,

$$\max_{(a,b) \in \mathcal{A} \times \mathcal{B}} \|m(a,b)\|_2 \leq M.$$

With mixed actions taken and observed, for all strategies of the second player, with probability 1,

$$\inf_{c \in \mathcal{C}} \left\| c - \frac{1}{T} \sum_{t=1}^T m(\mathbf{x}_t, \mathbf{y}_t) \right\|_2 \leq \frac{2M}{\sqrt{T}}.$$

With pure actions taken and observed, for all  $\delta \in (0, 1)$  and for all strategies of the second player, with probability at least  $1 - \delta$ ,

$$\inf_{c \in \mathcal{C}} \left\| c - \frac{1}{T} \sum_{t=1}^T m(A_t, B_t) \right\|_2 \leq \frac{2M}{\sqrt{T}} \left( 1 + 2\sqrt{\ln(2/\delta)} \right).$$

The proof is standard and is omitted from this article; it is detailed in the extended version of this paper (Mannor et al., 2011a).

**An associated strategy (that is efficient depending on the geometry of  $\mathcal{C}$ ).** Blackwell suggested a simple strategy with a geometric flavor.

Play an arbitrary  $\mathbf{x}_1$ . For  $t \geq 1$ , given the vector-valued quantities

$$\hat{m}_t = \frac{1}{t} \sum_{\tau=1}^t m(\mathbf{x}_\tau, B_\tau) \quad \text{or} \quad \hat{m}_t = \frac{1}{t} \sum_{\tau=1}^t m(\mathbf{x}_\tau, \mathbf{y}_\tau),$$

depending on whether pure or mixed actions are taken and observed, compute the projection  $c_t$  (in  $\ell^2$ -norm) of  $\hat{m}_t$  on  $\mathcal{C}$ . Find a mixed action  $\mathbf{x}_{t+1}$  that solves the minimax equation

$$\min_{\mathbf{x} \in \Delta(\mathcal{A})} \max_{\mathbf{y} \in \Delta(\mathcal{B})} \langle \hat{m}_t - c_t, m(\mathbf{x}, \mathbf{y}) \rangle, \quad (1)$$

where  $\langle \cdot, \cdot \rangle$  is the Euclidian inner product in  $\mathbb{R}^d$ . The minimax problem above is easily seen to be a (scalar) zero-sum game and is therefore efficiently solvable using, e.g., linear programming: the associated complexity is polynomial in  $N_{\mathcal{A}}$  and  $N_{\mathcal{B}}$ . All in all, this strategy is efficient as soon as the computations of the required projections onto  $\mathcal{C}$  in  $\ell^2$ -norm can be performed efficiently.

In the case when pure actions are taken and observed, it only remains to draw  $A_{t+1}$  at random according to  $\mathbf{x}_{t+1}$ .

### 3. Robust approachability

In this section we extend the results of the previous section to set-valued payoff functions. To this end, we denote by  $\mathcal{S}(\mathbb{R}^d)$  the set of all subsets of  $\mathbb{R}^d$  and consider a set-valued payoff function  $\overline{m} : \mathcal{A} \times \mathcal{B} \rightarrow \mathcal{S}(\mathbb{R}^d)$ .

**Pure actions taken and observed.** At each round  $t$ , the players choose simultaneously respective actions  $A_t \in \mathcal{A}$  and  $B_t \in \mathcal{B}$ , possibly at random according to mixed distributions  $\mathbf{x}_t$  and  $\mathbf{y}_t$ . Full monitoring still takes place for the first player: he observes  $B_t$  at the end of round  $t$ . However, as a result, the first player gets the subset  $\overline{m}(A_t, B_t)$  as a payoff. This models the ambiguity or uncertainty associated with some true underlying payoff gained.

We extend  $\overline{m}$  multi-linearly to  $\Delta(\mathcal{A}) \times \Delta(\mathcal{B})$  and even to  $\Delta(\mathcal{A} \times \mathcal{B})$ , the set of joint probability distributions on  $\mathcal{A} \times \mathcal{B}$ , as follows. Let

$$\mu = (\mu_{a,b})_{(a,b) \in \mathcal{A} \times \mathcal{B}}$$

be such a joint probability distribution; then  $\overline{m}(\mu)$  is defined as a finite convex combination<sup>2</sup> of subsets of  $\mathbb{R}^d$ ,

$$\overline{m}(\mu) = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \mu_{a,b} \overline{m}(a, b).$$

When  $\mu$  is the product-distribution of some  $\mathbf{x} \in \Delta(\mathcal{A})$  and  $\mathbf{y} \in \Delta(\mathcal{B})$ , we use the notation  $\overline{m}(\mu) = \overline{m}(\mathbf{x}, \mathbf{y})$ .

We denote by

$$\pi_T = \frac{1}{T} \sum_{t=1}^T \delta_{(A_t, B_t)}$$

the empirical distribution of the pairs  $(A_t, B_t)$  of actions taken during the first  $T$  rounds and will be interested in the behavior of

$$\frac{1}{T} \sum_{t=1}^T \overline{m}(A_t, B_t),$$

which can also be rewritten here in a compact way as  $\overline{m}(\pi_T)$ , by linearity of the extension of  $\overline{m}$ .

---

2. For two sets  $S, T$  and  $\alpha \in [0, 1]$ , the convex combination  $\alpha S + (1 - \alpha)T$  is defined as

$$\{\alpha s + (1 - \alpha)t, \quad s \in S \text{ and } t \in T\}.$$

**Definition 4** Let  $\mathcal{C} \subseteq \mathbb{R}^d$  be some set;  $\mathcal{C}$  is  $\overline{m}$ -approachable with pure actions if there exists a strategy of the first player such that for all strategies of the second player,

$$\limsup_{T \rightarrow \infty} \sup_{d \in \overline{m}(\pi_T)} \inf_{c \in \mathcal{C}} \|c - d\|_2 = 0 \quad a.s.$$

That is, when  $\mathcal{C}$  is  $\overline{m}$ -approachable with pure actions, the first player has a strategy that ensures that the average of the sets of payoffs converges to the set  $\mathcal{C}$ : the sets  $\overline{m}(\pi_T)$  are included in  $\varepsilon_T$ -neighborhoods of  $\mathcal{C}$ , where the sequence of  $\varepsilon_T$  tends almost-surely to 0.

**Mixed actions taken and observed.** At each round  $t$ , the players choose simultaneously respective mixed actions  $\mathbf{x}_t \in \Delta(\mathcal{A})$  and  $\mathbf{y}_t \in \Delta(\mathcal{B})$ . Full monitoring still takes place for the first player: he observes  $\mathbf{y}_t$  at the end of round  $t$ ; he however gets the subset  $\overline{m}(\mathbf{x}_t, \mathbf{y}_t)$  as a payoff (which, again, accounts for the uncertainty).

The product-distribution of two elements  $\mathbf{x} = (x_a)_{a \in \mathcal{A}} \in \Delta(\mathcal{A})$  and  $\mathbf{y} = (y_b)_{b \in \mathcal{B}} \in \Delta(\mathcal{B})$  will be denoted by  $\mathbf{x} \otimes \mathbf{y}$ ; it gives a probability mass of  $x_a y_b$  to each pair  $(a, b) \in \mathcal{A} \times \mathcal{B}$ . We consider the empirical joint distribution of mixed actions taken during the first  $T$  rounds,

$$\nu_T = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \otimes \mathbf{y}_t,$$

and will be interested in the behavior of

$$\frac{1}{T} \sum_{t=1}^T \overline{m}(\mathbf{x}_t, \mathbf{y}_t),$$

which can also be rewritten here in a compact way as  $\overline{m}(\nu_T)$ , by linearity of the extension of  $\overline{m}$ .

**Definition 5** Let  $\mathcal{C} \subseteq \mathbb{R}^d$  be some set;  $\mathcal{C}$  is  $\overline{m}$ -approachable with mixed actions if there exists a strategy of the first player such that for all strategies of the second player,

$$\limsup_{T \rightarrow \infty} \sup_{d \in \overline{m}(\nu_T)} \inf_{c \in \mathcal{C}} \|c - d\|_2 = 0 \quad a.s.$$

**A useful continuity lemma.** Before proceeding we provide a continuity lemma. It can be reformulated as indicating that for all joint distributions  $\mu$  and  $\nu$  over  $\mathcal{A} \times \mathcal{B}$ , the set  $\overline{m}(\mu)$  is contained in a  $M \|\mu - \nu\|_1$ -neighborhood of  $\overline{m}(\nu)$ , where  $M$  is a bound in  $\ell^2$ -norm on  $\overline{m}$ ; this is a fact that we will use repeatedly below.

**Lemma 6** Let  $\mu$  and  $\nu$  be two probability distributions over  $\mathcal{A} \times \mathcal{B}$ . We assume that the set-valued function  $\overline{m}$  is bounded in norm by  $M$ , i.e., that there exists a real number  $M > 0$  such that

$$\forall (a, b) \in \mathcal{A} \times \mathcal{B}, \quad \sup_{d \in \overline{m}(a, b)} \|d\|_2 \leq M.$$

Then

$$\sup_{d \in \overline{m}(\mu)} \inf_{c \in \overline{m}(\nu)} \|d - c\|_2 \leq M \|\mu - \nu\|_1 \leq M \sqrt{N_{\mathcal{A}} N_{\mathcal{B}}} \|\mu - \nu\|_2,$$

where the norms in the right-hand side are respectively the  $\ell^1$  and  $\ell^2$ -norms between probability distributions.

**Proof** Let  $d$  be an element of  $\overline{m}(\mu)$ ; it can be written as

$$d = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \mu_{a,b} \theta_{a,b}$$

for some elements  $\theta_{a,b} \in \overline{m}(a, b)$ . We consider

$$c = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \nu_{a,b} \theta_{a,b},$$

which is an element of  $\overline{m}(\nu)$ . Then by the triangle inequality,

$$\|d - c\|_2 = \left\| \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} (\mu_{a,b} - \nu_{a,b}) \theta_{a,b} \right\|_2 \leq \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} |\mu_{a,b} - \nu_{a,b}| \|\theta_{a,b}\|_2 \leq M \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} |\mu_{a,b} - \nu_{a,b}|.$$

This entails the first claimed inequality. The second one follows from an application of the Cauchy-Schwarz inequality. ■

**Necessary and sufficient condition for approachability.** We state the condition in the theorem below, as well as the associated convergence rates. Explicit strategies can be deduced from the proof, which is based on Theorem 3; these strategies are efficient as soon as projections in  $\ell^2$ -norm onto the set  $\tilde{\mathcal{C}}$  defined in (3) can be computed efficiently. The latter fact depends on the respective geometries of  $\overline{m}$  and  $\mathcal{C}$ .

**Theorem 7** Suppose that the set-valued function  $\overline{m}$  is bounded in norm by  $M$ . A closed convex set  $\mathcal{C} \subseteq \mathbb{R}^d$  is approachable (with pure or mixed actions) if and only if the following robust approachability condition is satisfied,

$$\forall \mathbf{y} \in \Delta(\mathcal{B}), \quad \exists \mathbf{x} \in \Delta(\mathcal{A}), \quad \overline{m}(\mathbf{x}, \mathbf{y}) \subseteq \mathcal{C}. \quad (\text{RAC})$$

In the latter case, the following convergence rates are achieved by a strategy constructed in the proof. With mixed actions taken and observed, for all strategies of the second player, with probability 1,

$$\sup_{d \in \overline{m}(\nu_T)} \inf_{c \in \mathcal{C}} \|c - d\|_2 \leq \frac{2M}{\sqrt{T}} \sqrt{N_{\mathcal{A}} N_{\mathcal{B}}}.$$

With pure actions taken and observed, for all  $\delta \in (0, 1)$  and for all strategies of the second player, with probability at least  $1 - \delta$ ,

$$\sup_{d \in \overline{m}(\pi_T)} \inf_{c \in \mathcal{C}} \|c - d\|_2 \leq \frac{2M}{\sqrt{T}} \sqrt{N_{\mathcal{A}} N_{\mathcal{B}}} \left(1 + 2\sqrt{\ln(2/\delta)}\right).$$

**Proof that Condition (RAC) is necessary.** If the condition does not hold, then there exists  $\mathbf{y}_0 \in \Delta(\mathcal{B})$  such that for every  $\mathbf{x} \in \mathcal{A}$ , the set  $\overline{m}(\mathbf{x}, \mathbf{y}_0)$  is not included in  $\mathcal{C}$ , i.e., it contains at least one point not in  $\mathcal{C}$ . We then define a mapping  $D : \Delta(\mathcal{A}) \rightarrow \mathbb{R}$  by

$$\forall \mathbf{x} \in \Delta(\mathcal{A}), \quad D(\mathbf{x}) = \sup_{d \in \overline{m}(\mathbf{x}, \mathbf{y}_0)} \inf_{c \in \mathcal{C}} \|c - d\|_2.$$

Since  $\mathcal{C}$  is closed, distances of given individual points to  $\mathcal{C}$  are achieved; therefore, by the choice of  $\mathbf{y}_0$ , we get that  $D(\mathbf{x}) > 0$  for all  $\mathbf{x} \in \Delta(\mathcal{A})$ .

We now show that  $D$  is continuous on the compact set  $\Delta(\mathcal{A})$ ; it thus attains its minimum, whose value we denote by  $D_{\min} > 0$ . More precisely, it suffices to show that for all  $\mathbf{x}, \mathbf{x}' \in \Delta(\mathcal{A})$ , the condition  $\|\mathbf{x}' - \mathbf{x}\|_1 \leq \varepsilon$  implies that  $D(\mathbf{x}) - D(\mathbf{x}') \leq M\varepsilon$ . Indeed, fix  $\delta > 0$  and let  $d_{\delta, \mathbf{x}} \in \overline{m}(\mathbf{x}, \mathbf{y}_0)$  be such that

$$D(\mathbf{x}) \leq \inf_{c \in \mathcal{C}} \|c - d_{\delta, \mathbf{x}}\|_2 + \delta. \quad (2)$$

By Lemma 6 (with the choices  $\mu = \mathbf{x} \otimes \mathbf{y}_0$  and  $\nu = \mathbf{x}' \otimes \mathbf{y}_0$ ) there exists  $d_{\delta, \mathbf{x}'} \in \overline{m}(\mathbf{x}', \mathbf{y}_0)$  such that  $\|d_{\delta, \mathbf{x}} - d_{\delta, \mathbf{x}'}\|_2 \leq M\varepsilon + \delta$ . The triangle inequality entails that

$$\inf_{c \in \mathcal{C}} \|c - d_{\delta, \mathbf{x}}\|_2 \leq \inf_{c \in \mathcal{C}} \|c - d_{\delta, \mathbf{x}'}\|_2 + M\varepsilon + \delta.$$

Substituting in (2), we get that

$$D(\mathbf{x}) \leq M\varepsilon + 2\delta + \inf_{c \in \mathcal{C}} \|c - d_{\delta, \mathbf{x}'}\|_2 \leq M\varepsilon + 2\delta + D(\mathbf{x}'),$$

which, letting  $\delta \rightarrow 0$ , proves our continuity claim.

Assume now that the second player chooses at each round  $\mathbf{y}_t = \mathbf{y}_0$  as his mixed action. In the case of mixed actions taken and observed, denoting

$$\bar{\mathbf{x}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t,$$

we get that  $\nu_t = \bar{\mathbf{x}}_T \otimes \mathbf{y}_0$ , and hence, for all strategies of the first player and for all  $T \geq 1$ ,

$$\sup_{d \in \overline{m}(\nu_T)} \inf_{c \in \mathcal{C}} \|c - d\|_2 = D(\bar{\mathbf{x}}_T) \geq D_{\min} > 0,$$

which shows that  $\mathcal{C}$  is not approachable. The case of pure actions taken and observed is treated similarly, with the sole addition of a concentration argument. By repeated uses of the Hoeffding-Azuma inequality together with an application of the Borel-Cantelli lemma,  $\delta_T = \|\pi_T - \nu_T\|_1 \rightarrow 0$  almost surely as  $T \rightarrow \infty$ . By applying Lemma 6 as above, we get

$$\sup_{d \in \overline{m}(\pi_T)} \inf_{c \in \mathcal{C}} \|c - d\|_2 \geq \sup_{d \in \overline{m}(\nu_T)} \inf_{c \in \mathcal{C}} \|c - d\|_2 - M\delta_T \geq D_{\min} - M\delta_T;$$

we simply take the liminf in the above inequalities to conclude the argument.  $\blacksquare$

**Proof that Condition (RAC) is sufficient.** We first show that there exists a strategy of the first player such that, for all strategies of the opponent player, the sequences  $(\pi_T)$  or  $(\nu_T)$  of the empirical distributions of actions converge to the set

$$\tilde{\mathcal{C}} = \{\mu \in \Delta(\mathcal{A} \times \mathcal{B}) : \overline{m}(\mu) \subseteq \mathcal{C}\} \quad (3)$$

in  $\ell^2$ -norm, at the rates prescribed by Theorem 3.

To do so, we identify probability distributions over  $\mathcal{A} \times \mathcal{B}$  with vectors in  $\mathbb{R}^{\mathcal{A} \times \mathcal{B}}$  and consider the vector-valued payoff function

$$m : (a, b) \in \mathcal{A} \times \mathcal{B} \mapsto \delta_{(a,b)} \in \mathbb{R}^{\mathcal{A} \times \mathcal{B}},$$

which we extend multi-linearly to  $\Delta(\mathcal{A}) \times \Delta(\mathcal{B})$ . We have that

$$\pi_T = \frac{1}{T} \sum_{t=1}^T m(A_t, B_t) \quad \text{and} \quad \nu_T = \frac{1}{T} \sum_{t=1}^T m(\mathbf{x}_t, \mathbf{y}_t)$$

and we therefore only need to show that  $\tilde{\mathcal{C}}$  is  $m$ -approachable (with pure or mixed actions).

Since  $\bar{m}$  is a linear function on  $\Delta(\mathcal{A} \times \mathcal{B})$  and  $\mathcal{C}$  is convex, the set  $\tilde{\mathcal{C}}$  is convex as well. In addition, since  $\mathcal{C}$  is closed,  $\tilde{\mathcal{C}}$  is also closed. We can therefore apply the original version of the approachability theorem (stated in Theorem 3). The desired existence result follows therefore from the fact that by assumption, for all  $\mathbf{y} \in \Delta(\mathcal{B})$ , there exists some  $\mathbf{x} \in \Delta(\mathcal{A})$  such that  $\mu = m(\mathbf{x}, \mathbf{y})$ , the product-distribution between  $\mathbf{x}$  and  $\mathbf{y}$ , belongs to  $\tilde{\mathcal{C}}$ , as it satisfies  $\bar{m}(\mu) = \bar{m}(\mathbf{x}, \mathbf{y}) \subseteq \mathcal{C}$ .

Let  $P_{\tilde{\mathcal{C}}}$  denote the projection operator onto  $\tilde{\mathcal{C}}$ . We therefore have proved the existence of explicit (and possibly efficient) strategies—along the lines of the ones presented around (1)—such that, for all strategies of the second player, with probability  $1 - \delta$ ,

$$\varepsilon_T := \left\| \pi_T - P_{\tilde{\mathcal{C}}}(\pi_T) \right\|_2 = \inf_{\mu \in \tilde{\mathcal{C}}} \|\pi_T - \mu\|_2 \leq \frac{2}{\sqrt{T}} \left( 1 + \sqrt{2 \ln(2/\delta)} \right),$$

and with probability 1,  $\varepsilon'_T := \left\| \nu_T - P_{\tilde{\mathcal{C}}}(\nu_T) \right\|_2 = \inf_{\mu \in \tilde{\mathcal{C}}} \|\nu_T - \mu\|_2 \leq \frac{2}{\sqrt{T}}$ .

Lemma 6 entails that the sets  $\bar{m}(\pi_T)$  are included in  $M\sqrt{N_{\mathcal{A}}N_{\mathcal{B}}}\varepsilon_T$ -neighborhoods of  $\bar{m}(P_{\tilde{\mathcal{C}}}(\pi_T))$ , and thus, by definition of  $\tilde{\mathcal{C}}$ , in  $M\sqrt{N_{\mathcal{A}}N_{\mathcal{B}}}\varepsilon_T$ -neighborhoods of  $\mathcal{C}$ . A similar statement holds for the sets the sets  $\bar{m}(\nu_T)$  and this completes the proof. ■

#### 4. Application to games with partial monitoring

A repeated vector-valued game with partial monitoring is described as follows (see, e.g., Mertens et al., 1994; Rustichini, 1999 and the references therein). The players have respective finite action sets  $\mathcal{I}$  and  $\mathcal{J}$ . We denote by  $r : \mathcal{I} \times \mathcal{J} \rightarrow \mathbb{R}^d$  the vector-valued payoff function of the first player and extend it multi-linearly to  $\Delta(\mathcal{I}) \times \Delta(\mathcal{J})$ . At each round, players simultaneously choose their actions  $I_t \in \mathcal{I}$  and  $J_t \in \mathcal{J}$ , possibly at random according to probability distributions denoted by  $\mathbf{p}_t \in \Delta(\mathcal{I})$  and  $\mathbf{q}_t \in \Delta(\mathcal{J})$ . At the end of a round, the first player does not observe  $J_t$  or  $r(I_t, J_t)$  but only a signal. There is a finite set  $\mathcal{H}$  of possible signals; the feedback  $S_t$  that is given to the first player is drawn at random according to the distribution  $H(I_t, J_t)$ , where the mapping  $H : \mathcal{I} \times \mathcal{J} \rightarrow \Delta(\mathcal{H})$  is known by the first player.

Some additional notation will be useful. We denote by  $R$  the norm of (the linear extension of)  $r$ ,

$$R = \max_{(i,j) \in \mathcal{I} \times \mathcal{J}} \|r(i, j)\|_2.$$

The cardinalities of the finite sets  $\mathcal{I}$ ,  $\mathcal{J}$ , and  $\mathcal{H}$  will be referred to as  $N_{\mathcal{I}}$ ,  $N_{\mathcal{J}}$ , and  $N_{\mathcal{H}}$ .

Definition 1 can be extended as follows in this setting; the only new ingredient is the signaling structure, the aim is unchanged.

**Definition 8** Let  $\mathcal{C} \subseteq \mathbb{R}^d$  be some set;  $\mathcal{C}$  is  $r$ -approachable for the signaling structure  $H$  if there exists a strategy of the first player such that for all strategies of the second player,

$$\limsup_{T \rightarrow \infty} \inf_{c \in \mathcal{C}} \left\| c - \frac{1}{T} \sum_{t=1}^T r(I_t, J_t) \right\|_2 = 0 \quad \text{a.s.}$$

That is, the first player has a strategy that ensures that the sequence of his average vector-valued payoffs converges to the set  $\mathcal{C}$ , even if he only observes the random signals  $S_t$  as a feedback.

A necessary and sufficient condition for  $r$ -approachability with the signaling structure  $H$  was stated and proved by Perchet (2011a); we therefore need to indicate where our contribution lies. First, both proofs are constructive but our strategy can be efficient (as soon as some projection operator can be efficiently implemented) whereas the one of Perchet (2011a) relies on auxiliary strategies that are calibrated and that require a grid that is progressively refined to be so (leading to a step complexity that is exponential in the number  $T$  of past steps). Second, we are able to exhibit convergence rates. Third, as far as elegance is concerned, our proof is short, compact, and more direct than the one of Perchet (2011a), which relied on several layers of complicated notions (internal regret in games with partial monitoring, calibration of auxiliary strategies, etc.).

To recall the mentioned approachability condition of Perchet (2011a) we need some additional notation: for all  $\mathbf{q} \in \Delta(\mathcal{J})$ , we denote by  $\tilde{H}(\mathbf{q})$  the element in  $\Delta(\mathcal{H})^{\mathcal{I}}$  defined as follows. For all  $i \in \mathcal{I}$ , its  $i$ -th component is given by the following convex combination of probability distributions over  $\mathcal{H}$ ,

$$\tilde{H}(\mathbf{q})_i = H(i, \mathbf{q}) = \sum_{j \in \mathcal{J}} q_j H(i, j).$$

Finally, we denote by  $\mathcal{F}$  the set of feasible vectors of probability distributions over  $\mathcal{H}$ :

$$\mathcal{F} = \left\{ \tilde{H}(\mathbf{q}) : \mathbf{q} \in \Delta(\mathcal{J}) \right\}.$$

A generic element of  $\mathcal{F}$  will be denoted by  $\sigma \in \mathcal{F}$ . The necessary and sufficient condition exhibited by Perchet (2011a) for the  $r$ -approachability of  $\mathcal{C}$  for the signaling structure  $H$  can now be recalled.

**Condition 1** The signaling structure  $H$ , the vector-payoff function  $r$ , and the set  $\mathcal{C}$  satisfy

$$\forall \mathbf{q} \in \Delta(\mathcal{J}), \exists \mathbf{p} \in \Delta(\mathcal{I}), \forall \mathbf{q}' \in \Delta(\mathcal{J}), \quad \tilde{H}(\mathbf{q}) = \tilde{H}(\mathbf{q}') \Rightarrow r(\mathbf{p}, \mathbf{q}') \in \mathcal{C}.$$

Defining the set-valued function  $\bar{m}$ , for all  $\mathbf{p} \in \Delta(\mathcal{I})$  and  $\sigma \in \mathcal{F}$ , by

$$\bar{m}(\mathbf{p}, \sigma) = \left\{ r(\mathbf{p}, \mathbf{q}') : \mathbf{q}' \in \Delta(\mathcal{J}) \text{ such that } \tilde{H}(\mathbf{q}') = \sigma \right\},$$

the condition can be equivalently reformulated as

$$\forall \sigma \in \mathcal{F}, \exists \mathbf{p} \in \Delta(\mathcal{I}), \quad \bar{m}(\mathbf{p}, \sigma) \subseteq \mathcal{C}.$$

**This condition is necessary.** The subsequent sections show (in a constructive way and by constructing strategies) that Condition 1 is sufficient for  $r$ -approachability of closed convex sets  $\mathcal{C}$  given the signaling structure  $H$ . That this condition is necessary was already proved in Perchet (2011a); a slightly simpler argument can however be found in the extended version of this paper (Mannor et al., 2011a).

#### 4.1. Approachability in bi-piecewise linear games

In this section we consider the case where the signaling structure has some special property described below; the case of general signaling structures is considered in Section 4.3.

To define bi-piecewise linearity of a game, we start from a technical lemma that is needed to show that  $\bar{m}(\mathbf{p}, \sigma)$  can be written as a *finite* convex combination of sets of the form  $\bar{m}(\mathbf{p}, b)$ , where  $b$  belongs to some finite set  $\mathcal{B} \subseteq \mathcal{F}$  that depends on the game. Under the additional assumption of piecewise-linearity of the thus defined mappings  $\bar{m}(\cdot, b)$ , we then describe a (possibly) efficient strategy for approachability followed by convergence rate guarantees.

##### 4.1.1. BI-PIECEWISE LINEARITY OF A GAME – A PRELIMINARY TECHNICAL RESULT

With general signaling structures,  $\bar{m}$  is not linear, it only satisfies that for all  $\mathbf{p} \in \Delta(\mathcal{I})$ , all pairs  $\sigma, \sigma' \in \mathcal{F}$ , and all  $\alpha \in [0, 1]$ ,

$$\alpha \bar{m}(\mathbf{p}, \sigma) + (1 - \alpha) \bar{m}(\mathbf{p}, \sigma') \subseteq \bar{m}(\mathbf{p}, \alpha\sigma + (1 - \alpha)\sigma') ,$$

with a strict inclusion in general. (Specific examples can be provided.) Therefore, a direct appeal to Theorem 7 is not possible.

However, a suitable linearity property on a lifted finite set is almost given by the geometric lemma stated below. It follows from an application of Rambau and Ziegler (1996, Proposition 2.4), which entails that since  $\tilde{H}$  is linear on the polytope  $\Delta(\mathcal{J})$ , its inverse application  $\tilde{H}^{-1}$  is a piecewise linear mapping of  $\mathcal{F}$  into the subsets of  $\Delta(\mathcal{J})$ ; the detailed proof can be found in the extended version of this paper (Mannor et al., 2011a,b).

**Lemma 9** *For any game of partial monitoring, there exists a finite set  $\mathcal{B} \subset \mathcal{F}$  and a piecewise-linear (injective) mapping  $\Phi : \mathcal{F} \rightarrow \Delta(\mathcal{B})$  such that*

$$\forall \sigma \in \mathcal{F}, \quad \forall \mathbf{p} \in \Delta(\mathcal{I}), \quad \bar{m}(\mathbf{p}, \sigma) = \sum_{b \in \mathcal{B}} \Phi_b(\sigma) \bar{m}(\mathbf{p}, b) ,$$

where we denoted the convex weight vector  $\Phi(\sigma) \in \Delta(\mathcal{B})$  by  $(\Phi_b(\sigma))_{b \in \mathcal{B}}$ .

The results of this subsection will rely on the following assumption.

**Assumption 1** *A game is bi-piecewise linear if  $\bar{m}(\cdot, b)$  is piecewise linear on  $\Delta(\mathcal{I})$  for every  $b \in \mathcal{B}$ .*

Assumption 1 means that for all  $b \in \mathcal{B}$  there exists a decomposition of  $\Delta(\mathcal{I})$  into polytopes each on which  $\bar{m}(\cdot, b)$  is linear. Since  $\mathcal{B}$  is finite, there exists a finite number of such decompositions, and thus there exists a polytopial decomposition that refines all of

them. (The latter is generated by the intersection of all considered polytopes as  $b$  varies.) By construction, every  $\bar{m}(\cdot, b)$  is linear on any of the polytopes of this common decomposition. We denote by  $\mathcal{A} \subset \Delta(\mathcal{I})$  the finite subset of all their vertices: a construction similar to the one used in the proof of Lemma 9 then leads to a piecewise linear (injective) mapping  $\Theta : \Delta(\mathcal{I}) \rightarrow \Delta(\mathcal{A})$ , where  $\Theta(\mathbf{p})$  is the decomposition of  $\mathbf{p}$  on the vertices of the polytope(s) of the decomposition to which it belongs, satisfying

$$\forall b \in \mathcal{B}, \quad \forall \mathbf{p} \in \Delta(\mathcal{I}), \quad \bar{m}(\mathbf{p}, b) = \sum_{a \in \mathcal{A}} \Theta_a(\mathbf{p}) \bar{m}(a, b),$$

where we denoted the convex weight vector  $\Theta(\mathbf{p}) \in \Delta(\mathcal{B})$  by  $(\Theta_a(\mathbf{p}))_{a \in \mathcal{A}}$ . Therefore, on a lifted space,  $\bar{m}$  is seen to coincide with a bi-linear mapping.

**Definition 10** *We denote by  $\bar{\bar{m}}$  the linear extension to  $\Delta(\mathcal{A} \times \mathcal{B})$  of the restriction of  $\bar{m}$  to  $\mathcal{A} \times \mathcal{B}$ , so that for all  $\mathbf{p} \in \Delta(\mathcal{I})$  and  $\sigma \in \mathcal{F}$ ,*

$$\bar{m}(\mathbf{p}, \sigma) = \bar{\bar{m}}(\Theta(\mathbf{p}), \Phi(\sigma)).$$

#### 4.1.2. CONSTRUCTION OF A STRATEGY TO APPROACH $\mathcal{C}$

The approaching strategy for the original problem is based on a strategy  $\Psi$  for  $\bar{\bar{m}}$ -approachability of  $\mathcal{C}$ , provided by Theorem 7 and thus solving repeatedly minimax problems of the form (1). We therefore first need to prove the existence of such a  $\Psi$ .

**Lemma 11** *Under Condition 1, the closed convex set  $\mathcal{C}$  is  $\bar{\bar{m}}$ -robust approachable.*

**Proof** We show that Condition (RAC) in Theorem 7 is satisfied, that is, that for all  $\mathbf{y} \in \Delta(\mathcal{B})$ , there exists some  $\mathbf{x} \in \Delta(\mathcal{A})$  such that  $\bar{\bar{m}}(\mathbf{x}, \mathbf{y}) \subseteq \mathcal{C}$ . With a given such  $\mathbf{y} \in \Delta(\mathcal{B})$ , we associate the feasible vector of signals  $\sigma = \sum_{b \in \mathcal{B}} y_b b$  and let  $\mathbf{p}$  be given by Condition 1, so<sup>3</sup> that  $\bar{m}(\mathbf{p}, \sigma) \subset \mathcal{C}$ . By linearity of  $\bar{m}$  (for the first equality), by definition of  $\bar{m}$  (for the first inclusion), by Lemma 9 (for the second and fourth equalities), by construction of  $\mathcal{A}$  (for the third equality),

$$\begin{aligned} \bar{\bar{m}}(\Theta(\mathbf{p}), \mathbf{y}) &= \sum_{a \in \mathcal{A}} \Theta_a(\mathbf{p}) \sum_{b \in \mathcal{B}} y_b \bar{m}(a, b) \subseteq \sum_{a \in \mathcal{A}} \Theta_a(\mathbf{p}) \bar{m}(a, \sigma) = \sum_{a \in \mathcal{A}} \Theta_a(\mathbf{p}) \sum_{b \in \mathcal{B}} \Phi_b(\sigma) \bar{m}(a, b) \\ &= \sum_{b \in \mathcal{B}} \Phi_b(\sigma) \bar{m}(\mathbf{p}, b) = \bar{m}(\mathbf{p}, \sigma) \subset \mathcal{C}, \end{aligned}$$

which concludes the proof. ■

We consider the strategy described in Figure 1. It forces exploration at a  $\gamma$  rate, as is usual in situations with partial monitoring. One of its key ingredient, that conditionally unbiased estimators are available, is extracted from Lugosi et al. (2008, Section 6): in block  $n$  we consider

$$\hat{H}_t = \frac{\mathbb{I}_{\{S_t=s\}} \mathbb{I}_{\{I_t=i\}}}{p_{I_t,n}} \in \mathbb{R}^{\mathcal{H} \times \mathcal{I}},$$

---

3. Note however that we do not necessarily have that  $\Phi(\sigma)$  and  $\mathbf{y}$  are equal, as  $\Phi$  is not a one-to-one mapping.

---

*Parameters:* an integer block length  $L \geq 1$ , an exploration parameter  $\gamma \in [0, 1]$ , a strategy  $\Psi$  for  $\overline{m}$ -approachability of  $\mathcal{C}$

*Notation:*  $\mathbf{u} \in \Delta(\mathcal{I})$  is the uniform distribution over  $\mathcal{I}$ ,  $P_{\mathcal{F}}$  denotes the projection operator in  $\ell^2$ -norm of  $\mathbb{R}^{\mathcal{H} \times \mathcal{I}}$  onto  $\mathcal{F}$

*Initialization:* compute the finite set  $\mathcal{B}$  and the mapping  $\Phi : \mathcal{F} \rightarrow \Delta(\mathcal{B})$  of Lemma 9, pick an arbitrary  $\boldsymbol{\theta}_1 \in \Delta(\mathcal{A})$

For all blocks  $n = 1, 2, \dots$ ,

1. define  $\mathbf{x}_n = \sum_{a \in \mathcal{A}} \theta_{n,a} a$  and  $\mathbf{p}_n = (1 - \gamma) \mathbf{x}_n + \gamma \mathbf{u}$ ;
  2. for rounds  $t = (n-1)L + 1, \dots, nL$ ,
    - 2.1 drawn an action  $I_t \in \mathcal{I}$  at random according to  $\mathbf{p}_n$ ;
    - 2.2 get the signal  $S_t$ ;
  3. form the estimated vector of probability distributions over signals,
- $$\tilde{\sigma}_n = \left( \frac{1}{L} \sum_{t=(n-1)L+1}^{nL} \frac{\mathbb{I}_{\{S_t=s\}} \mathbb{I}_{\{I_t=i\}}}{p_{I_t,n}} \right)_{(i,s) \in \mathcal{I} \times \mathcal{H}};$$
4. compute the projection  $\hat{\sigma}_n = P_{\mathcal{F}}(\tilde{\sigma}_n)$ ;
  5. choose  $\boldsymbol{\theta}_{n+1} = \Psi(\boldsymbol{\theta}_1, \Phi(\hat{\sigma}_1), \dots, \boldsymbol{\theta}_n, \Phi(\hat{\sigma}_n))$ .
- 

Figure 1: The proposed strategy, which plays in blocks.

averaging over the respective random draws of  $I_t$  and  $S_t$  according to  $\mathbf{p}_n$  and  $H(I_t, J_t)$ , i.e., taking the conditional expectation  $\mathbb{E}_t$  with respect to  $\mathbf{p}_n$  and  $J_t$ , we get

$$\mathbb{E}_t[\hat{H}_t] = \tilde{H}(\delta_{J_t}). \quad (4)$$

This is why, by concentration-of-the-measure argument, we will be able to show that for  $L$  large enough,  $\tilde{\sigma}_n$  is close to  $\tilde{H}(\hat{\mathbf{q}}_n)$ , where

$$\hat{\mathbf{q}}_n = \frac{1}{L} \sum_{t=(n-1)L+1}^{nL} \delta_{J_t}. \quad (5)$$

Actually, since  $\mathcal{F} \subseteq \Delta(\mathcal{H})^{\mathcal{I}}$ , we have a natural embedding of  $\mathcal{F}$  into  $\mathbb{R}^{\mathcal{H} \times \mathcal{I}}$  and we can define  $P_{\mathcal{F}}$ , the convex projection operator onto  $\mathcal{F}$  (in  $\ell^2$ -norm). Instead of using directly  $\tilde{\sigma}_n$ , we consider in our strategy  $\hat{\sigma}_n = P_{\mathcal{F}}(\tilde{\sigma}_n)$ , which is even closer to  $H(\hat{\mathbf{q}}_n)$ .

#### 4.1.3. PERFORMANCE GUARANTEE

We provide a performance bound for fixed parameters  $\gamma$  and  $L$  tuned as functions of  $T$ . The proof is provided in the extended version of this paper (Mannor et al., 2011a,b). Adaptation to  $T \rightarrow \infty$  can be performed either by resorting to a standard doubling trick (see, e.g., Cesa-Bianchi and Lugosi 2006, page 17) or by taking time-varying parameters  $\gamma_t$  and  $L_t$ .

**Theorem 12** Consider a closed convex set  $\mathcal{C}$  and a game  $(r, H)$  for which Condition 1 is satisfied and that is bi-piecewise linear in the sense of Assumption 1. Then, for all  $T \geq 1$ , the strategy of Figure 1, run with parameters  $L = \lceil T^{3/5} \rceil$  and  $\gamma = T^{-1/5}$  and fed with a strategy  $\Psi$  for  $\bar{m}$ -approachability of  $\mathcal{C}$  (provided by Lemma 11) is such that, with probability at least  $1 - \delta$ ,

$$\inf_{c \in \mathcal{C}} \left\| c - \frac{1}{T} \sum_{t=1}^T r(I_t, J_t) \right\|_2 \leq \square \left( T^{-1/5} \sqrt{\ln \frac{T}{\delta}} + T^{-2/5} \ln \frac{T}{\delta} \right)$$

for some constant  $\square$  depending only on  $\mathcal{C}$  and on the game  $(r, H)$  at hand.

The efficiency of the strategy of Figure 1 depends on whether it can be fed with an efficient approachability strategy  $\Psi$ , which in turn depends on the respective geometries of  $\bar{m}$  and  $\mathcal{C}$ , as was indicated before the statement of Theorem 7. (Note that the projection onto  $\mathcal{F}$  can be performed in polynomial time, as the latter closed convex set is defined by finitely many linear constraints, and that the computation of  $\bar{m}$  can be performed beforehand.)

#### 4.2. Application to regret minimization

In this section we analyze external and internal regret minimization in repeated games with partial monitoring from the approachability perspective. Using the results developed for vector-valued games with partial monitoring, we show how to—in particular—minimize regret in both setups.

##### 4.2.1. EXTERNAL REGRET

We consider in this section the framework and aim introduced by Rustichini (1999) and studied, sometimes in special cases, by Piccolboni and Schindelhauer (2001); Mannor and Shimkin (2003); Cesa-Bianchi et al. (2006); Lugosi et al. (2008). We show that our general strategy can be used for regret minimization.

Scalar payoffs are obtained (but not observed) by the first player: the payoff function  $r$  is a mapping  $\mathcal{I} \times \mathcal{J} \rightarrow \mathbb{R}$ ; we still denote by  $R$  a bound on  $|r|$ . We define in this section

$$\hat{\mathbf{q}}_T = \frac{1}{T} \sum_{t=1}^T \delta_{J_t}$$

as the empirical distribution of the actions taken by the second player. The external regret of the first player at round  $T$  equals by definition

$$R_T^{\text{ext}} = \max_{\mathbf{p} \in \Delta(\mathcal{I})} \rho(\mathbf{p}, \tilde{H}(\hat{\mathbf{q}}_T)) - \frac{1}{T} \sum_{t=1}^T r(I_t, J_t),$$

where  $\rho : \Delta(\mathcal{I}) \times \mathcal{F}$  is defined as follows: for all  $\mathbf{p} \in \Delta(\mathcal{I})$  and  $\sigma \in \mathcal{F}$ ,

$$\rho(\mathbf{p}, \sigma) = \min \left\{ r(\mathbf{p}, \mathbf{q}) : \mathbf{q} \text{ such that } \tilde{H}(\mathbf{q}) = \sigma \right\}.$$

The function  $\rho$  is continuous in its first argument and therefore the supremum in the defining expression of  $R_T^{\text{ext}}$  is a maximum.

We recall briefly why, intuitively, this is the natural notion of external regret to consider in this case. Indeed, the first term in the definition of  $R_T^{\text{ext}}$  is (close to) the worst-case average payoff obtained by the first player when playing consistently a mixed action  $\mathbf{p}$  against a sequence of mixed actions inducing the same laws on the signals.

The following result is an easy consequence of Theorem 12, as is explained below; it corresponds to the main result of Lugosi et al. (2008), with the same convergence rate but with a different strategy. (However, Perchet 2011b, Section 2.3 exhibited an efficient strategy achieving a convergence rate of order  $T^{-1/3}$ , which is optimal; a question is thus whether the rates exhibited in Theorem 12 could be improved.)

**Corollary 13** *For all  $T$ , the first player has a strategy such that, for all strategies of the second player and with probability at least  $1 - \delta$ ,*

$$R_T^{\text{ext}} \leq \square \left( T^{-1/5} \sqrt{\ln \frac{T}{\delta}} + T^{-2/5} \ln \frac{T}{\delta} \right)$$

for some constant  $\square$  depending only on the game  $(r, H)$  at hand.

The proof below is an extension to the setting of partial monitoring of the original proof and strategy of Blackwell (1956b) for the case of external regret under full monitoring: in the case of full monitoring the vector-payoff function  $\underline{r}$  and the set  $\mathcal{C}$  considered in our proof are equal to the ones considered by Blackwell.

**Proof** We embed  $\mathcal{F}$  into  $\mathbb{R}^{\mathcal{H} \times \mathcal{I}}$  so that in this proof we will be working in the vector space  $\mathbb{R} \times \mathbb{R}^{\mathcal{H} \times \mathcal{I}}$ . We consider the closed convex set  $\mathcal{C}$  and the vector-valued payoff function  $\underline{r}$  respectively defined by

$$\mathcal{C} = \left\{ (z, \sigma) \in \mathbb{R} \times \mathcal{F} : z \geq \max_{\mathbf{p} \in \Delta(\mathcal{I})} \rho(\mathbf{p}, \sigma) \right\} \quad \text{and} \quad \underline{r}(i, j) = \begin{bmatrix} r(i, j) \\ \tilde{H}(\delta_j) \end{bmatrix},$$

for all  $(i, j) \in \mathcal{I} \times \mathcal{J}$ .

We now first show that Assumption 1 is satisfied. To do so, we will actually prove the stronger property that the mappings  $\bar{m}(\cdot, \sigma)$  are piecewise linear for all  $\sigma \in \mathcal{F}$ ; we fix such a  $\sigma$  in the sequel. Only the first coordinate of  $\underline{r}$  depends on  $\mathbf{p}$ , so the desired property is true if and only if the mapping  $\bar{m}_1(\cdot, \sigma)$  defined by

$$\mathbf{p} \in \Delta(\mathcal{I}) \longmapsto \bar{m}_1(\mathbf{p}, \sigma) = \left\{ r(\mathbf{p}, \mathbf{q}') : \mathbf{q}' \in \Delta(\mathcal{J}) \text{ such that } \tilde{H}(\mathbf{q}') = \sigma \right\}$$

is piecewise linear. Since  $\tilde{H}$  is linear, the set

$$\left\{ \mathbf{q}' \in \Delta(\mathcal{J}) \text{ such that } \tilde{H}(\mathbf{q}') = \sigma \right\}$$

is a polytope, thus, the convex hull of some finite set  $\{\mathbf{q}_{\sigma,1}, \dots, \mathbf{q}_{\sigma,M}\} \subset \Delta(\mathcal{J})$ . Therefore, for every  $\mathbf{p} \in \mathcal{I}$ , by linearity of  $r$  (and by the fact that it takes one-dimensional values),

$$\bar{m}_1(\mathbf{p}, \sigma) = \text{co} \left\{ r(\mathbf{p}, \mathbf{q}_{\sigma,1}), \dots, r(\mathbf{p}, \mathbf{q}_{\sigma,M}) \right\} = \left[ \min_{k \in \{1, \dots, M\}} r(\mathbf{p}, \mathbf{q}_{\sigma,k}), \max_{k' \in \{1, \dots, M\}} r(\mathbf{p}, \mathbf{q}_{\sigma,k'}) \right],$$

where  $\text{co}$  stands for the convex hull. Since all applications  $r(\cdot, \mathbf{q}_{\sigma,k})$  are linear, their minimum and their maximum are piecewise linear functions, thus  $\bar{m}_1(\cdot, \sigma)$  is also piecewise linear.

We then show that Condition 1 is satisfied for the considered convex set  $\mathcal{C}$  and game  $(\underline{r}, H)$ . To do so, we associate with each  $\mathbf{q} \in \Delta(\mathcal{J})$  an element  $\phi(\mathbf{q}) \in \Delta(\mathcal{I})$  such that

$$\phi(\mathbf{q}) \in \operatorname{argmax}_{\mathbf{p} \in \Delta(\mathcal{I})} \rho(\mathbf{p}, \tilde{H}(\mathbf{q})).$$

Then, given any  $\mathbf{q} \in \Delta(\mathcal{J})$ , we note that for all  $\mathbf{q}'$  satisfying  $\tilde{H}(\mathbf{q}') = \tilde{H}(\mathbf{q})$ , we have, by definition of  $\rho$ ,

$$r(\phi(\mathbf{q}), \mathbf{q}') \geq \rho(\phi(\mathbf{q}), \tilde{H}(\mathbf{q}')) = \max_{\mathbf{p} \in \Delta(\mathcal{I})} \rho(\mathbf{p}, \tilde{H}(\mathbf{q}')),$$

which shows that  $r(\phi(\mathbf{q}), \mathbf{q}') \in \mathcal{C}$ . The required condition is thus satisfied.

Theorem 12 can therefore be applied to exhibit the convergence rates; we simply need to relate the quantity of interest here to the one considered therein. To that end we use the fact that the mapping

$$\sigma \in \mathcal{F} \longmapsto \max_{\mathbf{p} \in \Delta(\mathcal{I})} \rho(\mathbf{p}, \sigma)$$

is Lipschitz, with Lipschitz constant in  $\ell^2$ -norm denoted by  $L_\rho$ ; the proof of this fact is detailed in the extended version of this paper (Mannor et al., 2011a,b). Now, the regret is non positive as soon as  $\sum_{t=1}^T r(I_t, J_t)/T$  belongs to  $\mathcal{C}$ ; we therefore only need to consider the case when this average is not in  $\mathcal{C}$ . In the latter case, we denote by  $(\tilde{r}_T, \tilde{\sigma}_T)$  its projection in  $\ell^2$ -norm onto  $\mathcal{C}$ . We have first that the defining inequality of  $\mathcal{C}$  is an equality on its border, so that

$$\tilde{r}_T = \max_{\mathbf{p} \in \Delta(\mathcal{I})} \rho(\mathbf{p}, \tilde{\sigma}_T);$$

and second, that

$$\begin{aligned} R_T^{\text{ext}} &= \max_{\mathbf{p} \in \Delta(\mathcal{I})} \rho(\mathbf{p}, \tilde{H}(\hat{\mathbf{q}}_T)) - \frac{1}{T} \sum_{t=1}^T r(I_t, J_t) \\ &\leq \left| \max_{\mathbf{p} \in \Delta(\mathcal{I})} \rho(\mathbf{p}, \tilde{H}(\hat{\mathbf{q}}_T)) - \max_{\mathbf{p} \in \Delta(\mathcal{I})} \rho(\mathbf{p}, \tilde{\sigma}_T) \right| + \left| \tilde{r}_T - \frac{1}{T} \sum_{t=1}^T r(I_t, J_t) \right| \\ &\leq L_\rho \left\| \tilde{H}(\hat{\mathbf{q}}_T) - \tilde{\sigma}_T \right\|_2 + \left| \tilde{r}_T - \frac{1}{T} \sum_{t=1}^T r(I_t, J_t) \right| \\ &\leq \sqrt{2} \max\{L_\rho, 1\} \left\| \begin{bmatrix} \tilde{r}_T \\ \tilde{\sigma}_T \end{bmatrix} - \frac{1}{T} \sum_{t=1}^T \underline{r}(I_t, J_t) \right\|_2 \\ &= \sqrt{2} \max\{L_\rho, 1\} \inf_{c \in \mathcal{C}} \left\| c - \frac{1}{T} \sum_{t=1}^T \underline{r}(I_t, J_t) \right\|_2. \end{aligned}$$

As claimed, the rates are now seen to follow from the ones indicated in Theorem 12. ■

#### 4.2.2. INTERNAL / SWAP REGRET

Foster and Vohra (1999) defined internal regret with full monitoring as follows. A player has no internal regret if, for every action  $i \in \mathcal{I}$ , he has no external regret on the stages when this specific action  $i$  was played. In other words,  $i$  is the best response to the empirical distribution of action of the other player on these stages.

With partial monitoring, the first player evaluates his payoffs in some pessimistic way through the function  $\rho$  defined above. This function is not linear over  $\Delta(\mathcal{I})$  in general (it is concave), so that the best responses are not necessarily pure actions  $i \in \mathcal{I}$  but mixed actions, i.e., elements of  $\Delta(\mathcal{I})$ . Following Lehrer and Solan (2007) we therefore should partition the stages not depending on the pure actions actually played but on the mixed actions  $\mathbf{p}_t \in \Delta(\mathcal{I})$  used to draw them. To this end, it is convenient to assume that the strategies of the first player need to pick these mixed actions in a finite (but possibly thin) grid of  $\Delta(\mathcal{I})$ , which we denote by  $\{\mathbf{p}_g, g \in \mathcal{G}\}$ , where  $\mathcal{G}$  is a finite set. At each round, the first player picks an index  $G_t \in \mathcal{G}$  and uses the distribution  $\mathbf{p}_{G_t}$  to draw his action  $I_t$ . Up to a standard concentration-of-the-measure argument, we will measure the payoff at round  $t$  with  $r(\mathbf{p}_{G_t}, J_t)$  rather than with  $r(I_t, J_t)$ .

For each  $g \in \mathcal{G}$ , we denote by  $N_T(g)$  the number of stages in  $\{1, \dots, T\}$  for which we had  $G_t = g$  and, whenever  $N_T(g) > 0$ ,

$$\hat{\mathbf{q}}_{T,g} = \frac{1}{N_T(g)} \sum_{t:G_t=g} \delta_{J_t}.$$

We define  $\hat{\mathbf{q}}_{T,g}$  is an arbitrary way when  $N_T(g) = 0$ . The internal regret of the first player at round  $T$  is measured as

$$R_T^{\text{int}} = \max_{g, g' \in \mathcal{G}} \frac{N_T(g)}{T} \left( \rho(\mathbf{p}_{g'}, \tilde{H}(\hat{\mathbf{q}}_{T,g})) - r(\mathbf{p}_g, \hat{\mathbf{q}}_{T,g}) \right).$$

Actually, our proof technique rather leads to the minimization of some swap regret (see Blum and Mansour, 2007 for the definition of swap regret in full monitoring):

$$R_T^{\text{swap}} = \sum_{g \in \mathcal{G}} \frac{N_T(g)}{T} \left( \max_{g' \in \mathcal{G}} \rho(\mathbf{p}_{g'}, \tilde{H}(\hat{\mathbf{q}}_{T,g})) - r(\mathbf{p}_g, \hat{\mathbf{q}}_{T,g}) \right).$$

Again, the following bound on the swap regret easily follows from Theorem 12; the latter constructs a simple and direct strategy to control the swap regret, thus also the internal regret. It therefore improves on the results of Lehrer and Solan (2007); Perchet (2009), two articles which presented complicated strategies to do so (strategies based on auxiliary strategies using a grid that needs to be refined over time and whose complexities is exponential in the size of these grids). Moreover, we provide convergence rates.

**Corollary 14** *For all  $T$ , the first player has an explicit strategy such that, for all strategies of the second player and with probability at least  $1 - \delta$ ,*

$$R_T^{\text{swap}} \leq \square \left( T^{-1/5} \sqrt{\ln \frac{T}{\delta}} + T^{-2/5} \ln \frac{T}{\delta} \right)$$

for some constant  $\square$  depending only on the game  $(r, H)$  at hand and on the size of the finite grid  $\mathcal{G}$ .

The proof of this corollary is based on ideas similar to the ones used in the proof of Corollary 13; it can be found in the extended version of this paper (Mannor et al., 2011a,b).

### 4.3. Approachability in the case of general games

Unfortunately, as is illustrated in the extended version of this paper (Mannor et al., 2011b), there exist games with partial monitoring that are not bi-piecewise linear.

However, we will show that if Condition 1 holds there exist strategies with a constant per-round complexity to approach polytopes even when the game is not bi-piecewise linear. That is, by considering simpler closed convex sets  $\mathcal{C}$ , no assumption is needed on the pair  $(r, H)$ . We will conclude this subsection by indicating that thanks to a doubling trick, Condition 1 is still seen to be sufficient for approachability in the most general case when no assumption is made neither on  $(r, H)$  nor on  $\mathcal{C}$ , at the cost however of inefficiency.

#### 4.3.1. APPROACHABILITY OF THE NEGATIVE ORTHANT IN THE CASE OF GENERAL GAMES

For the sake of simplicity, we start with the case of the negative orthant  $\mathbb{R}_-^d$ . Our argument will be based on Lemma 9; we use in the sequel the objects and notation introduced therein. We denote by  $r = (r_k)_{1 \leq k \leq d}$  the components of the  $d$ -dimensional payoff function  $r$  and introduce, for all  $k \in \{1, \dots, d\}$ , the set-valued mapping  $\tilde{m}_k$  defined by

$$\tilde{m}_k : (\mathbf{p}, b) \in \Delta(\mathcal{I}) \times \mathcal{B} \mapsto \tilde{m}_k(\mathbf{p}, b) = \left\{ r_k(\mathbf{p}, \mathbf{q}) : \mathbf{q} \in \Delta(\mathcal{J}) \text{ such that } \tilde{H}(\mathbf{q}) = b \right\}.$$

The mapping  $\tilde{m}$  is then defined as the Cartesian product of the  $\tilde{m}_k$ ; formally, for all  $\mathbf{p} \in \Delta(\mathcal{I})$  and  $b \in \mathcal{B}$ ,

$$\tilde{m}(\mathbf{p}, b) = \left\{ (z_1, \dots, z_d) : \forall k \in \{1, \dots, d\}, z_k \in \tilde{m}_k(\mathbf{p}, b) \right\}.$$

We then linearly extend this mapping into a set-valued mapping  $\check{m}$  defined on  $\Delta(\mathcal{I}) \times \Delta(\mathcal{B})$  and finally consider the set-valued mapping  $\check{m}$  defined on  $\Delta(\mathcal{I}) \times \mathcal{F}$  by

$$\forall b \in \mathcal{B}, \quad \forall \mathbf{p} \in \Delta(\mathcal{I}), \quad \check{m}(\mathbf{p}, \sigma) = \tilde{m}(\mathbf{p}, \Phi(\sigma)) = \sum_{b \in \mathcal{B}} \Phi_b(\sigma) \tilde{m}(\mathbf{p}, b),$$

where  $\Phi$  refers to the mapping defined in Lemma 9. The lemma below indicates why  $\check{m}$  is an excellent substitute to  $\overline{m}$  in the case of the approachability of the orthant  $\mathbb{R}_-^d$ .

**Lemma 15** *The set-valued mappings  $\check{m}$  and  $\overline{m}$  are linked by the following two properties: for all  $p \in \Delta(\mathcal{I})$  and  $\sigma \in \mathcal{F}$ ,*

1. *the inclusion  $\overline{m}(\mathbf{p}, \sigma) \subseteq \check{m}(\mathbf{p}, \sigma)$  holds;*
2. *if  $\overline{m}(\mathbf{p}, \sigma) \subseteq \mathbb{R}_-^d$ , then one also has  $\check{m}(\mathbf{p}, \sigma) \subseteq \mathbb{R}_-^d$ .*

The interpretations of these two properties are that 1.  $\check{m}$ -robust approaching a set  $\mathcal{C}$  is more difficult than  $\overline{m}$ -robust approaching it; and 2. that if Condition 1 holds for  $\overline{m}$  and

$\mathbb{R}_-^d$ , it also holds for  $\check{m}$  and  $\mathbb{R}_-^d$ .

**Proof** For property 1., note that by construction of  $\check{m}$ ,

$$\forall b \in \mathcal{B}, \quad \forall \mathbf{p} \in \Delta(\mathcal{I}), \quad \overline{m}(\mathbf{p}, b) \subseteq \check{m}(\mathbf{p}, b);$$

Lemma 9 and the linear extension of  $\check{m}$  then show that

$$\forall \sigma \in \mathcal{F}, \quad \forall \mathbf{p} \in \Delta(\mathcal{I}), \quad \overline{m}(\mathbf{p}, \sigma) \subseteq \check{m}(\mathbf{p}, \Phi(\sigma)) = \check{m}(\mathbf{p}, \sigma).$$

As for property 2., it suffices to note that (by Lemma 9 again) the stated assumption exactly means that  $\sum_{b \in \mathcal{B}} \Phi_b(\sigma) \overline{m}(\mathbf{p}, b) \subset \mathbb{R}_-^d$ . In particular, rewriting the non-positivity constraint for each of the  $d$  components of the payoff vectors, we get

$$\sum_{b \in \mathcal{B}} \Phi_b(\sigma) \check{m}_k(\mathbf{p}, b) \subseteq \mathbb{R}_-,$$

for all  $k \in \{1, \dots, d\}$ ; thus, in particular,  $\sum_{b \in \mathcal{B}} \Phi_b(\sigma) \check{m}(\mathbf{p}, b) = \check{m}(\mathbf{p}, \sigma) \subseteq \mathbb{R}_-^d$ . ■

We can then extend the result of the previous section as announced; note that no bi-piecewise linearity assumption is needed on the game.

**Theorem 16** *If Condition 1 is satisfied for  $\overline{m}$  and  $\mathbb{R}_-^d$ , then there exists a strategy for  $(r, H)$ -approaching  $\mathbb{R}_-^d$  at a rate of the order of  $T^{-1/5}$ , with a constant per-round complexity.*

**Proof (sketched)** The assumption of the theorem and Property 2. of Lemma 15 imply that Condition 1 holds for  $\mathbb{R}_-^d$  and  $\check{m}$ ; furthermore, the latter corresponds to a bi-piecewise linear game as can be seen by noting, similarly to what was done in the section devoted to regret minimization, that each  $\check{m}_k$ , thus also  $\check{m}$ , is a piecewise linear function. Thus, (the proof of) Theorem 12 guarantees that  $\mathcal{C}$  is  $\check{m}$ -robust approachable. Now, Property 1. of Lemma 15 implies that any  $\check{m}$ -robust approachability strategy of  $\mathcal{C} = \mathbb{R}_-^d$  is also a  $\overline{m}$ -robust approachability strategy. Therefore,  $\mathcal{C}$  is  $\overline{m}$ -robust approachable, hence, following again the methodology used in the proof of Theorem 12, is also  $(r, H)$ -approachable. ■

#### 4.3.2. APPROACHABILITY OF POLYTOPES IN THE CASE OF GENERAL GAMES

If that the target set  $\mathcal{C}$  is a polytope, then  $\mathcal{C}$  can be written as the intersection of a finite number of half-planes, i.e., there exists a finite family  $\{(e_k, f_k) \in \mathbb{R}^d \times \mathbb{R}, k \in \mathcal{K}\}$  such that

$$\mathcal{C} = \{z \in \mathbb{R}^d : \langle z, e_k \rangle \leq f_k, \forall k \in \mathcal{K}\}.$$

Given the original (not necessarily bi-piecewise linear) game  $(r, H)$ , we introduce another game  $(r_{\mathcal{C}}, H)$ , whose payoff function  $r_{\mathcal{C}} : \mathcal{I} \times \mathcal{J} \rightarrow \mathbb{R}^{\mathcal{K}}$  is defined as

$$\forall i \in \mathcal{I}, \quad \forall j \in \mathcal{J}, \quad r_{\mathcal{C}}(i, j) = \left[ \langle r(i, j), e_k \rangle - f_k \right]_{k \in \mathcal{K}}.$$

The following lemma is an exercise of mere rewriting.

**Lemma 17** *Given a polytope  $\mathcal{C}$ , the  $(r, H)$ -approachability of  $\mathcal{C}$  and the  $(r_{\mathcal{C}}, H)$ -approachability of  $\mathbb{R}_{-}^d$  are equivalent in the sense that all strategies for one problem translates to a strategy for the other problem.*

*In addition, Condition 1 holds for  $(r, H)$  and  $\mathcal{C}$  if and only if it holds for  $(r_{\mathcal{C}}, H)$  and  $\mathbb{R}_{-}^d$ .*

Via the lemma above, Theorem 16 indicates that Condition 1 for  $(r, H)$  and  $\mathcal{C}$  is a sufficient condition for the  $(r, H)$ -approachability of  $\mathcal{C}$  and provides a strategy to do so.

#### 4.3.3. APPROACHABILITY OF GENERAL CONVEX SETS IN THE CASE OF GENERAL GAMES

A general closed convex set can always be approximated arbitrarily well by a polytope (where the number of vertices of the latter however increases as the quality of the approximation does). There, via a doubling trick, Condition 1 is also seen to be sufficient to  $(r, H)$ -approach any general closed convex set  $\mathcal{C}$ . However, the computational complexity of the resulting strategy is much larger: the per-round complexity increases over time (as the numbers of vertices of the approximating polytopes do).

## References

- J. Abernethy, P. L. Bartlett, and E. Hazan. Blackwell approachability and low-regret learning are equivalent. In *Proceedings of the Twenty-Fourth Annual Conference on Learning Theory (COLT'11)*. Omnipress, 2011.
- D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956a.
- D. Blackwell. Controlled random walks. In *Proceedings of the International Congress of Mathematicians, 1954, Amsterdam, vol. III*, pages 336–338, 1956b.
- A. Blum and Y. Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324, 2007.
- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31:562–580, 2006.
- D. Foster and R. Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29:7–36, 1999.
- S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150, 2000.
- S. Hart and A. Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98:26–54, 2001.
- E. Lehrer and E. Solan. Learning to play partially-specified equilibrium. Mimeo, 2007.

- G. Lugosi, S. Mannor, and G. Stoltz. Strategies for prediction under imperfect monitoring. *Mathematics of Operations Research*, 33:513–528, 2008. An extended abstract was presented at COLT’07.
- S. Mannor and N. Shimkin. On-line learning with imperfect monitoring. In *Proceedings of the Sixteenth Annual Conference on Learning Theory (COLT’03)*, pages 552–567. Springer, 2003.
- S. Mannor and N. Shimkin. Regret minimization in repeated matrix games with variable stage duration. *Games and Economic Behavior*, 63(1):227–258, 2008.
- S. Mannor, J. Tsitsiklis, and J. Y. Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(Mar):569–590, 2009.
- S. Mannor, V. Perchet, and G. Stoltz. Robust approachability and regret minimization in games with partial monitoring. 2011a. URL <http://hal.archives-ouvertes.fr/hal-00595695>.
- S. Mannor, V. Perchet, and G. Stoltz. Corrigendum to “Robust approachability and regret minimization in games with partial monitoring”. 2011b. URL <http://hal.archives-ouvertes.fr/hal-00617554>.
- J.-F. Mertens, S. Sorin, and S. Zamir. Repeated games. Technical Report no. 9420, 9421, 9422, Université de Louvain-la-Neuve, 1994.
- V. Perchet. Calibration and internal no-regret with random signals. In *Proceedings of the Twentieth International Conference on Algorithmic Learning Theory (ALT’09)*, pages 68–82, 2009.
- V. Perchet. Approachability of convex sets in games with partial monitoring. *Journal of Optimization Theory and Applications*, 149:665–677, 2011a.
- V. Perchet. Internal regret with partial monitoring calibration-based optimal algorithms. *Journal of Machine Learning Research*, 2011b. In press.
- A. Piccolboni and C. Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory (COLT’01)*, pages 208–223, 2001.
- A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Beyond regret. In *Proceedings of the Twenty-Fourth Annual Conference on Learning Theory (COLT’11)*. Omnipress, 2011.
- J. Rambau and G. Ziegler. Projections of polytopes and the generalized Baues conjecture. *Discrete and Computational Geometry*, 16:215–237, 1996.
- A. Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29: 224–243, 1999.

### Acknowledgments

Shie Mannor was partially supported by the ISF under contract 890015. Vianney Perchet benefited from the support of the ANR under grant ANR-10-BLAN 0112. Gilles Stoltz acknowledges support from the French National Research Agency (ANR) under grant EX-PLO/RA (“Exploration–exploitation for efficient resource allocation”) and by the PASCAL2 Network of Excellence under EC grant no. 506778.

# The Rate of Convergence of AdaBoost

**Indraneel Mukherjee**

*Princeton University, Department of Computer Science  
Princeton, NJ 08540 USA*

IMUKHERJ@CS.PRINCETON.EDU

**Cynthia Rudin**

*Massachusetts Institute of Technology  
MIT Sloan School of Management  
Cambridge, MA 02139 USA*

RUDIN@MIT.EDU

**Robert E. Schapire**

*Princeton University, Department of Computer Science  
Princeton, NJ 08540 USA*

SCHAPIRE@CS.PRINCETON.EDU

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

The AdaBoost algorithm of Freund and Schapire (1997) was designed to combine many “weak” hypotheses that perform slightly better than a random guess into a “strong” hypothesis that has very low error. We study the rate at which AdaBoost iteratively converges to the minimum of the “exponential loss” with a fast rate of convergence. Our proofs do not require a weak-learning assumption, nor do they require that minimizers of the exponential loss are finite. Specifically, our first result shows that at iteration  $t$ , the exponential loss of AdaBoost’s computed parameter vector will be at most  $\varepsilon$  more than that of any parameter vector of  $\ell_1$ -norm bounded by  $B$  in a number of rounds that is bounded by a polynomial in  $B$  and  $1/\varepsilon$ . We also provide rate lower bound examples showing a polynomial dependence on these parameters is necessary. Our second result is that within  $C/\varepsilon$  iterations, AdaBoost achieves a value of the exponential loss that is at most  $\varepsilon$  more than the best possible value, where  $C$  depends on the dataset. We show that this dependence of the rate on  $\varepsilon$  is optimal up to constant factors, i.e. at least  $\Omega(1/\varepsilon)$  rounds are necessary to achieve within  $\varepsilon$  of the optimal exponential loss.

**Keywords:** AdaBoost, optimization, coordinate descent, convergence rate.

## 1. Introduction

The AdaBoost algorithm of Freund and Schapire (1997) was designed to combine many “weak” hypotheses that perform slightly better than a random guess into a “strong” hypothesis that has very low error. AdaBoost has been named in 2006 as one of the “top 10” algorithms in data mining (Wu et al., 2008) and has performed favorably with respect to other popular machine learning algorithms in empirical comparisons (Caruana and Niculescu-Mizil, 2006). Despite AdaBoost’s popularity, basic properties of its convergence are not well understood. In this work, we focus on one of those properties, namely to find

convergence rates when there are no simplifying assumptions. For instance, we do not assume that the “weak learning assumption” necessarily holds, that all of the weak hypotheses perform at least slightly better than random guessing. If the weak learning assumption holds, or if other assumptions hold, it is easier to prove a fast convergence rate for AdaBoost. However, in some cases where AdaBoost is commonly applied, such simplifying assumptions do not necessarily hold, and it is not as easy to find a convergence rate.

AdaBoost can be viewed as a coordinate descent (functional gradient descent) algorithm that iteratively minimizes an objective function  $L : \mathbb{R}^n \rightarrow \mathbb{R}$  called the *exponential loss* (Breiman, 1999; Frean and Downs, 1998; Friedman et al., 2000; Friedman, 2001; Mason et al., 1999; Onoda et al., 1998; Rätsch et al., 2001; Schapire and Singer, 1999). The exponential loss is constructed from  $m$  labeled training examples  $(x_1, y_1), \dots, (x_m, y_m)$ , where the  $x_i$ ’s are in some domain  $\mathcal{X}$  and  $y_i \in \{-1, +1\}$ , and a set of hypotheses  $\mathcal{H} = \{\hbar_1, \dots, \hbar_N\}$ , where each  $\hbar_j : \mathcal{X} \rightarrow \{-1, +1\}$ . Specifically, the exponential loss is defined as follows:

$$L(\boldsymbol{\lambda}) := \sum_{i=1}^m \exp \left( - \sum_{j=1}^N \lambda_j y_i \hbar_j(x_i) \right).$$

In each iteration, a coordinate descent algorithm moves some distance along some coordinate direction. For AdaBoost, the coordinate directions are provided by the individual weak hypotheses. Correspondingly, AdaBoost chooses some weak hypotheses and a step length, and then adds that to the current combination. The direction and step length are so chosen that the resulting vector  $\boldsymbol{\lambda}^t$  in iteration  $t$  yields a lower value of the exponential loss than in the previous iteration,  $L(\boldsymbol{\lambda}^t) < L(\boldsymbol{\lambda}^{t-1})$ . This repeats until it reaches a minimizer if one exists. It was shown by Collins et al. (2002), and later by Zhang and Yu (2005), that AdaBoost asymptotically converges to the minimum possible exponential loss. That is,

$$\lim_{t \rightarrow \infty} L(\boldsymbol{\lambda}^t) = \inf_{\boldsymbol{\lambda} \in \mathbb{R}^N} L(\boldsymbol{\lambda}),$$

though that work did not address a convergence rate to the minimizer of the exponential loss.

Our work specifically addresses a recent conjecture of Schapire (2010) stating that there exists a positive constant  $c$  and a polynomial  $\text{poly}()$  such that for all training sets and all finite sets of weak hypotheses, and for all  $B > 0$ ,

$$L(\boldsymbol{\lambda}^t) \leq \min_{\boldsymbol{\lambda}: \|\boldsymbol{\lambda}\|_1 \leq B} L(\boldsymbol{\lambda}) + \frac{\text{poly}(\log N, m, B)}{t^c}. \quad (1)$$

In other words, the exponential loss of AdaBoost will be at most  $\varepsilon$  more than that of any other parameter vector  $\boldsymbol{\lambda}$  of  $\ell_1$ -norm bounded by  $B$  in a number of rounds that is bounded by a polynomial in  $\log N$ ,  $m$ ,  $B$  and  $1/\varepsilon$ . (We require  $\log N$  rather than  $N$  since the number of weak hypotheses will typically be extremely large.) Along with an upper bound that is polynomial in these parameters, we also provide lower bound constructions showing some polynomial dependence on  $B, \varepsilon^{-1}$  is necessary. Without any additional assumptions on the exponential loss  $L$ , and without altering AdaBoost’s minimization algorithm for  $L$ , the best known convergence rate of AdaBoost prior to this work that we are aware of is that of Bickel et al. (2006) who prove a bound on the rate of the form  $O(1/\sqrt{\log t})$ .

We provide also a convergence rate of AdaBoost to the minimum value of the exponential loss. Namely, within  $C/\epsilon$  iterations, AdaBoost achieves a value of the exponential loss that is at most  $\epsilon$  more than the best possible value, where  $C$  depends on the dataset. This convergence rate is different from the one discussed above in that it has better dependence on  $\epsilon$  (in fact the dependence is optimal, as we show), and does not depend on the best solution within a ball of size  $B$ . However, this second convergence rate cannot be used to prove (1) since in certain worst case situations, we show the constant  $C$  may be larger than  $2^m$  (although usually it will be much smaller).

Within the proof of the second convergence rate, we provide a lemma (called the *decomposition lemma*) that shows that the training set can be split into two sets of examples: the *finite margin set*, and the *zero loss set*. Examples in the finite margin set always make a positive contribution to the exponential loss, and they never lie too far from the decision boundary. Examples in the zero loss set do not have these properties. If we consider the exponential loss where the sum is only over the finite margin set (rather than over all training examples), it is minimized by a finite  $\lambda$ . The fact that the training set can be decomposed into these two classes is the key step in proving the second convergence rate.

This problem of determining the rate of convergence is relevant in the proof of the consistency of AdaBoost given by Bartlett and Traskin (2007), where it has a direct impact on the rate at which AdaBoost converges to the Bayes optimal classifier (under suitable assumptions). It may also be relevant to practitioners who wish to have a guarantee on the exponential loss value at iteration  $t$ .

There have been several works that make additional assumptions on the exponential loss in order to attain a better bound on the rate, but those assumptions are not true in general, and cases are known where each of these assumptions are violated. For instance, better bounds are proved by Rätsch et al. (2002) using results from Luo and Tseng (1992), but these appear to require that the exponential loss be minimized by a finite  $\lambda$ , and also depend on quantities that are not easily measured. There are many cases where  $L$  does not have a finite minimizer; in fact, one such case is provided by Schapire (2010). Shalev-Shwartz and Singer (2008) have proven bounds for a variant of AdaBoost. Zhang and Yu (2005) also have given rates of convergence, but their technique requires a bound on the change in the size of  $\lambda^t$  at each iteration that does not necessarily hold for AdaBoost. Many classic results are known on the convergence of iterative algorithms generally (see for instance Luenberger and Ye, 2008; Boyd and Vandenberghe, 2004); however, these typically start by assuming that the minimum is attained at some finite point in the (usually compact) space of interest. When the weak learning assumption holds, there is a parameter  $\gamma > 0$  that governs the improvement of the exponential loss at each iteration. Freund and Schapire (1997) and Schapire and Singer (1999) showed that the exponential loss is at most  $e^{-2t\gamma^2}$  after  $t$  rounds, so AdaBoost rapidly converges to the minimum possible loss under this assumption.

In Section 2 we summarize the coordinate descent view of AdaBoost. Section 3 contains the proof of the conjecture and associated lower bounds. Section 4 provides the  $C/\epsilon$  convergence rate.

---

Given:  $(x_1, y_1), \dots, (x_m, y_m)$  where  $x_i \in \mathcal{X}$ ,  $y_i \in \{-1, +1\}$   
set  $\mathcal{H} = \{\bar{h}_1, \dots, \bar{h}_N\}$  of weak hypotheses  $\bar{h}_j : \mathcal{X} \rightarrow \{-1, +1\}$ .  
Initialize:  $D_1(i) = 1/m$  for  $i = 1, \dots, m$ .  
For  $t = 1, \dots, T$ :

- Train weak learner using distribution  $D_t$ ; that is, find weak hypothesis  $h_t \in \mathcal{H}$  whose correlation  $r_t \triangleq \mathbb{E}_{i \sim D_t} [y_i h_t(x_i)]$  has maximum magnitude  $|r_t|$ .
- Choose  $\alpha_t = \frac{1}{2} \ln \{(1 + r_t) / (1 - r_t)\}$ .
- Update, for  $i = 1, \dots, m$ :  $D_{t+1}(i) = D_t(i) \exp(-\alpha_t y_i h_t(x_i)) / Z_t$   
where  $Z_t$  is a normalization factor (chosen so that  $D_{t+1}$  will be a distribution).

Output the final hypothesis:  $F(x) = \text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right)$ .

Figure 1: The boosting algorithm AdaBoost.

---

## 2. Coordinate Descent View of AdaBoost

From the examples  $(x_1, y_1), \dots, (x_m, y_m)$  and hypotheses  $\mathcal{H} = \{\bar{h}_1, \dots, \bar{h}_N\}$ , AdaBoost iteratively computes the function  $F : \mathcal{X} \rightarrow \mathbb{R}$ , where  $\text{sign}(F(x))$  can be used as a classifier for a new instance  $x$ . The function  $F$  is a linear combination of the hypotheses. At each iteration  $t$ , AdaBoost chooses one of the weak hypotheses,  $h_t$  from the set  $\mathcal{H}$ , and adjusts its coefficient by a specified value  $\alpha_t$ . Then  $F$  is constructed after  $T$  iterations as:  $F(x) = \sum_{t=1}^T \alpha_t h_t(x)$ . Figure 1 shows the AdaBoost algorithm (Freund and Schapire, 1997).

Since each  $h_t$  is equal to  $\bar{h}_{j_t}$  for some  $j_t$ ,  $F$  can also be written  $F(x) = \sum_{j=1}^N \lambda_j \bar{h}_j(x)$  for a vector of values  $\boldsymbol{\lambda} = \langle \lambda_1, \dots, \lambda_N \rangle$  (such vectors will sometimes also be referred to as combinations, since they represent combinations of weak hypotheses). In different notation, we can write AdaBoost as a coordinate descent algorithm on vector  $\boldsymbol{\lambda}$ . We define the “feature matrix”  $M$  elementwise by  $M_{ij} = y_i \bar{h}_j(x_i)$ , so that this matrix contains all of the inputs to AdaBoost (the training examples and hypotheses). Then the exponential loss can be written more compactly as:

$$L(\boldsymbol{\lambda}) = \frac{1}{m} \sum_i e^{-(M\boldsymbol{\lambda})_i}$$

where  $(M\boldsymbol{\lambda})_i$ , the  $i^{\text{th}}$  coordinate of the vector  $M\boldsymbol{\lambda}$ , is the “margin” achieved by vector  $\boldsymbol{\lambda}$  on training example  $i$ .

Coordinate descent algorithms choose a coordinate at each iteration where the directional derivative is the steepest, and choose a step that maximally decreases the objective along that coordinate. To perform coordinate descent on the exponential loss, we determine the coordinate  $j_t$  at iteration  $t$  as follows, where  $\mathbf{e}_j$  is a vector that is 1 in the  $j^{\text{th}}$  position and 0 elsewhere:

$$j_t \in \operatorname{argmax}_j \left| \left( -\frac{dL(\boldsymbol{\lambda}^{t-1} + \alpha \mathbf{e}_j)}{d\alpha} \Big|_{\alpha=0} \right) \right| = \operatorname{argmax}_j \frac{1}{m} \left| \sum_{i=1}^m e^{-(M\boldsymbol{\lambda}^{t-1})_i} M_{ij} \right|. \quad (2)$$

It can be shown (see, for instance Mason et al., 2000) that the distribution  $D_t$  chosen by AdaBoost at each round  $t$  puts weight  $D_t(i)$  proportional to  $e^{(-M\boldsymbol{\lambda}^{t-1})_i}$ . Expression (2) can

now be rewritten as

$$j_t \in \operatorname{argmax}_j \left| \sum_i D_t(i) M_{ij} \right| = \operatorname{argmax}_j \left| \mathbb{E}_{i \sim D_t} [M_{ij}] \right| = \operatorname{argmax}_j \left| \mathbb{E}_{i \sim D_t} [y_i h_j(x_i)] \right|,$$

which is exactly the way AdaBoost chooses a weak hypothesis in each round (see Figure 1). The correlation  $\sum_i D_t(i) M_{ij}$  will be denoted by  $r_t$  and its absolute value  $|r_t|$ , denoted by  $\delta_t$ . The quantity  $\delta_t$  is commonly called the *edge* for round  $t$ . The distance  $\alpha_t$  to travel along direction  $j_t$  is chosen to minimize the  $L(\boldsymbol{\lambda}^{t-1} + \alpha_t \mathbf{e}_{j_t})$ , and can be shown to be equal to  $\alpha_t = \frac{1}{2} \ln \left( \frac{1+r_t}{1-r_t} \right)$  (see, for instance Mason et al., 2000), just as in Figure 1. With this choice of step length, it can be shown (see, for instance Freund and Schapire, 1997) that the exponential loss drops by an amount depending on the edge:  $L(\boldsymbol{\lambda}^t) = L(\boldsymbol{\lambda}^{t-1}) \sqrt{1 - \delta_t^2}$ .

Our rate bounds also hold when the weak-hypotheses are confidence rated, that is, giving real-valued predictions in  $[-1, +1]$ , so that  $h : \mathcal{X} \rightarrow [-1, +1]$ . In that case, the criterion for picking a weak hypothesis in each round remains the same, that is, at round  $t$ , an  $\bar{h}_{j_t}$  maximizing the absolute correlation  $j_t \in \operatorname{argmax}_j \left| \sum_{i=1}^m e^{-(M\boldsymbol{\lambda}^{t-1})_i} M_{ij} \right|$ , is chosen, where  $M_{ij}$  may now be non-integral. An exact analytical line search is no longer possible, but if the step size is chosen in the same way

$$\alpha_t = \frac{1}{2} \ln \left( \frac{1+r_t}{1-r_t} \right), \quad (3)$$

then Freund and Schapire (1997) and Schapire and Singer (1999) show that a similar drop in the loss is still guaranteed

$$L(\boldsymbol{\lambda}^t) \leq L(\boldsymbol{\lambda}^{t-1}) \sqrt{1 - \delta^2}. \quad (4)$$

With confidence rated hypotheses, other implementations may choose the step size in a different way. However, in this paper, by AdaBoost we will always mean the version in (Freund and Schapire, 1997; Schapire and Singer, 1999) which chooses step sizes as in (3), and enjoys the loss guarantee as in (4). That said, all our proofs work more generally, and are robust to numerical inaccuracies in the implementation. In other words, even if the previous conditions are violated by a small amount, similar bounds continue to hold, although we leave out explicit proofs of this fact to simplify the presentation.

### 3. Convergence to any target loss

In this section, we bound the number of rounds of AdaBoost required to get within  $\varepsilon$  of the loss attained by any parameter vector  $\boldsymbol{\lambda}^*$  as a function of  $\varepsilon$  and the  $\ell_1$ -norm  $\|\boldsymbol{\lambda}^*\|_1$ . The vector  $\boldsymbol{\lambda}^*$  serves as a reference based on which we define the target loss  $L(\boldsymbol{\lambda}^*)$ , and its  $\ell_1$ -norm is a measure of difficulty of attaining the target loss. We prove a bound polynomial in  $1/\varepsilon$ ,  $\|\boldsymbol{\lambda}^*\|_1$  and the number of examples  $m$ , showing (1) holds, thereby resolving affirmatively the open problem posed in (Schapire, 2010). Later in the section we provide lower bounds showing how a polynomial dependence on both parameters is necessary.

**Theorem 1** *For any  $\boldsymbol{\lambda}^* \in \mathbb{R}^N$ , AdaBoost achieves loss at most  $L(\boldsymbol{\lambda}^*) + \varepsilon$  in at most  $13\|\boldsymbol{\lambda}^*\|_1^6 \varepsilon^{-5}$  rounds.*

The high level idea behind the proof of the theorem is as follows. To show a fast rate, we require a large edge in each round, as indicated by (4). A large edge is guaranteed if the size of the current solution of AdaBoost is small. Therefore AdaBoost makes good progress if the size of its solution does not grow too fast. On the other hand, the increase in size of its solution is given by the step length, which in turn is proportional to the edge achieved in that round. Therefore, if the solution size grows fast, the loss also drops fast. Either way the algorithm makes good progress. In the rest of the section we make these ideas concrete through a sequence of lemmas. The proof of Theorem 1 is based on these lemmas and appears later. We conclude by indicating possibilities for improvement in our analysis that might help tighten the exponents in the rate bound of Theorem 1.

We provide some more notation. Throughout,  $\boldsymbol{\lambda}^*$  is fixed, and its  $\ell_1$ -norm is denoted by  $B$  (matching the notation in (Schapire, 2010)). One key parameter is the suboptimality  $R_t$  of AdaBoost's solution measured via the logarithm of the exponential loss

$$R_t \triangleq \ln L(\boldsymbol{\lambda}^t) - \ln L(\boldsymbol{\lambda}^*).$$

Another key parameter is the  $\ell_1$ -distance  $S_t$  of AdaBoost's solution from the closest combination that achieves the target loss

$$S_t \triangleq \inf_{\boldsymbol{\lambda}} \{ \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^t\|_1 : L(\boldsymbol{\lambda}) \leq L(\boldsymbol{\lambda}^*) \}.$$

We will also be interested in how they change as captured by

$$\Delta R_t \triangleq R_{t-1} - R_t \geq 0, \quad \Delta S_t \triangleq S_t - S_{t-1}.$$

Notice that  $\Delta R_t$  is always non-negative since AdaBoost decreases the loss, and hence the suboptimality, in each round. Let  $T_0$  be the bound on the number of rounds in Theorem 1. We assume without loss of generality that  $R_0, \dots, R_{T_0}$  and  $S_0, \dots, S_{T_0}$  are all strictly positive, since otherwise the theorem holds trivially. Also, in the rest of the section, we restrict our attention entirely to the first  $T_0$  rounds of boosting. We first show that a  $\text{poly}(B, \varepsilon^{-1})$  rate of convergence follows if the edge is always polynomially large compared to the suboptimality.

**Lemma 2** *If for some constants  $c_1, c_2$ , where  $c_2 > 1/2$ , the edge satisfies  $\delta_t \geq B^{-c_1} R_{t-1}^{c_2}$  in each round  $t$ , then AdaBoost achieves at most  $L(\boldsymbol{\lambda}^*) + \varepsilon$  loss after  $2B^{2c_1}(\varepsilon \ln 2)^{1-2c_2}$  rounds.*

**Proof** From the definition of  $R_t$  and (4) we have

$$\Delta R_t = \ln L(\boldsymbol{\lambda}^{t-1}) - \ln L(\boldsymbol{\lambda}^t) \geq -\frac{1}{2} \ln(1 - \delta_t^2). \quad (5)$$

Combining the above with the inequality  $e^x \geq 1 + x$ , and the assumption on the edge

$$\Delta R_t \geq -\frac{1}{2} \ln(1 - \delta_t^2) \geq \delta_t^2/2 \geq \frac{1}{2} B^{-2c_1} R_{t-1}^{2c_2}.$$

Let  $T = \lceil 2B^{2c_1}(\varepsilon \ln 2)^{1-2c_2} \rceil$  be the bound on the number of rounds in the Lemma. If any of  $R_0, \dots, R_T$  is negative, then by monotonicity  $R_T < 0$  and we are done. Otherwise, they are

all non-negative. Then, applying Lemma 18 from the Appendix to the sequence  $R_0, \dots, R_T$ , and using  $c_2 > 1/2$  we get

$$R_T^{1-2c_2} \geq R_0^{1-2c_2} + c_2 B^{-2c_1} T > (1/2) B^{-2c_1} T \geq (\varepsilon \ln 2)^{1-2c_2} \implies R_T < \varepsilon \ln 2.$$

If either  $\varepsilon$  or  $L(\boldsymbol{\lambda}^*)$  is greater than 1, then the lemma follows since  $L(\boldsymbol{\lambda}^T) \leq L(\boldsymbol{\lambda}^0) = 1 < L(\boldsymbol{\lambda}^*) + \varepsilon$ . Otherwise,

$$L(\boldsymbol{\lambda}^T) < L(\boldsymbol{\lambda}^*) e^{\varepsilon \ln 2} \leq L(\boldsymbol{\lambda}^*)(1 + \varepsilon) \leq L(\boldsymbol{\lambda}^*) + \varepsilon,$$

where the second inequality uses  $e^x \leq 1 + (1/\ln 2)x$  for  $x \in [0, \ln 2]$ .  $\blacksquare$

We next show that large edges are achieved provided  $S_t$  is small compared to  $R_t$ .

**Lemma 3** *In each round  $t$ , the edge satisfies  $\delta_t \geq R_{t-1}/S_{t-1}$ .*

**Proof** For any combination  $\boldsymbol{\lambda}$ , define  $p_{\boldsymbol{\lambda}}$  as the distribution on examples  $\{1, \dots, m\}$  that puts weight proportional to the loss  $p_{\boldsymbol{\lambda}}(i) = e^{-(M\boldsymbol{\lambda})_i}/(mL(\boldsymbol{\lambda}))$ . Choose any  $\boldsymbol{\lambda}$  suffering less than the target loss  $L(\boldsymbol{\lambda}) \leq L(\boldsymbol{\lambda}^*)$ . By non-negativity of relative entropy we get

$$\begin{aligned} 0 &\leq \text{RE}(p_{\boldsymbol{\lambda}^{t-1}} \parallel p_{\boldsymbol{\lambda}}) = \sum_{i=1}^m p_{\boldsymbol{\lambda}^{t-1}} \ln \left\{ \frac{\frac{1}{m} e^{-(M\boldsymbol{\lambda}^{t-1})_i}/L(\boldsymbol{\lambda}^{t-1})}{\frac{1}{m} e^{-(M\boldsymbol{\lambda})_i}/L(\boldsymbol{\lambda})} \right\} \\ &= -R_{t-1} + \sum_{i=1}^m p_{\boldsymbol{\lambda}^{t-1}}(i) (M\boldsymbol{\lambda} - M\boldsymbol{\lambda}^{t-1})_i. \end{aligned} \quad (6)$$

Note that  $p_{\boldsymbol{\lambda}^{t-1}}$  is the distribution  $D_t$  that AdaBoost creates in round  $t$ . The above summation can be rewritten as

$$\begin{aligned} \sum_{i=1}^m p_{\boldsymbol{\lambda}^{t-1}}(i) \sum_{j=1}^N (\lambda_j - \lambda_j^{t-1}) M_{ij} &= \sum_{j=1}^N (\lambda_j - \lambda_j^{t-1}) \sum_{i=1}^m D_t(i) M_{ij} \\ &\leq \left( \sum_{j=1}^N |\lambda_j - \lambda_j^{t-1}| \right) \max_j \left| \sum_{i=1}^m D_t(i) M_{ij} \right| = \delta_t \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{t-1}\|_1. \end{aligned} \quad (7)$$

Since the previous holds for any  $\boldsymbol{\lambda}$  suffering less than the target loss, the last expression is at most  $\delta_t S_{t-1}$ . Combining this with (7) completes the proof.  $\blacksquare$

To complete the proof of Theorem 1, we show  $S_t$  is small compared to  $R_t$  in rounds  $t \leq T_0$  (during which we have assumed  $S_t, R_t$  are all non-negative). In fact we prove:

**Lemma 4** *For any  $t \leq T_0$ ,  $S_t \leq B^3 R_t^{-2}$ .*

This, along with Lemmas 2 and 3, proves Theorem 1. The bound on  $S_t$  in Lemma 4 can be proven if we can first show  $S_t$  grows slowly compared to the rate at which the suboptimality  $R_t$  falls. Intuitively this holds since growth in  $S_t$  is caused by a large step, which in turn will drive down the suboptimality. In fact we can prove the following.

**Lemma 5** In any round  $t \leq T_0$ , we have  $\frac{2\Delta R_t}{R_{t-1}} \geq \frac{\Delta S_t}{S_{t-1}}$ .

**Proof** Firstly, it follows from the definition of  $S_t$  that  $\Delta S_t \leq \|\boldsymbol{\lambda}^t - \boldsymbol{\lambda}^{t-1}\|_1 = |\alpha_t|$ . Next, using (5) and (3) we may write  $\Delta R_t \geq \Upsilon(\delta_t) |\alpha_t|$ , where the function  $\Upsilon$  has been defined in (Rätsch and Warmuth, 2005) as

$$\Upsilon(x) = \frac{-\ln(1-x^2)}{\ln\left(\frac{1+x}{1-x}\right)}.$$

It is known (Rätsch and Warmuth, 2005; Rudin et al., 2007) that  $\Upsilon(x) \geq x/2$  for  $x \in [0, 1]$ . Combining and using Lemma 3

$$\Delta R_t \geq \delta_t \Delta S_t / 2 \geq R_{t-1} (\Delta S_t / 2 S_{t-1}).$$

Rearranging completes the proof. ■

Using this we may prove Lemma 4.

**Proof** We first show  $S_0 \leq B^3 R_0^{-2}$ . Note,  $S_0 \leq \|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^0\|_1 = B$ , and by definition  $R_0 = -\ln\left\{\frac{1}{m} \sum_i e^{-(M\boldsymbol{\lambda}^*)_i}\right\}$ . The quantity  $(M\boldsymbol{\lambda}^*)_i$  is the inner product of row  $i$  of matrix  $M$  with the vector  $\boldsymbol{\lambda}^*$ . Since the entries of  $M$  lie in  $[-1, +1]$ , this is at most  $\|\boldsymbol{\lambda}^*\|_1 = B$ . Therefore  $R_0 \leq -\ln\left\{\frac{1}{m} \sum_i e^{-B}\right\} = B$ , which is what we needed.

To complete the proof, we show that  $R_t^2 S_t$  is non-increasing. It suffices to show for any  $t$  the inequality  $R_t^2 S_t \leq R_{t-1}^2 S_{t-1}$ . This holds by the following chain

$$\begin{aligned} R_t^2 S_t &= (R_{t-1} - \Delta R_t)^2 (S_{t-1} + \Delta S_t) = R_{t-1}^2 S_{t-1} \left(1 - \frac{\Delta R_t}{R_{t-1}}\right)^2 \left(1 + \frac{\Delta S_t}{S_{t-1}}\right) \\ &\leq R_{t-1}^2 S_{t-1} \exp\left(-\frac{2\Delta R_t}{R_{t-1}} + \frac{\Delta S_t}{S_{t-1}}\right) \leq R_{t-1}^2 S_{t-1}, \end{aligned}$$

where the first inequality follows from  $e^x \geq 1 + x$ , and the second one from Lemma 5. ■

### 3.1. Lower-bounds

Here we show that the dependence of the rate in Theorem 1 on the norm  $\|\boldsymbol{\lambda}^*\|_1$  of the solution achieving target accuracy is necessary for a wide class of datasets. The arguments in this section are not tailored to AdaBoost, but hold more generally for any coordinate descent algorithm and a wide variety of loss functions.

**Lemma 6** Suppose the feature matrix  $M$  corresponding to a dataset has two rows with  $\{-1, +1\}$  entries which are complements of each other, i.e. there are two examples on which any hypothesis gets one wrong and one correct prediction. Then the number of rounds required to achieve a target loss  $\phi^*$  is at least  $\inf\{\|\boldsymbol{\lambda}\|_1 : L(\boldsymbol{\lambda}) \leq \phi^*\} / (2 \ln m)$ .

**Proof** We first show that the two examples corresponding to the complementary rows in  $M$  both satisfy a certain margin boundedness property. Since each hypothesis predicts

oppositely on these, in any round  $t$  their margins will be of equal magnitude and opposite sign. Unless both margins lie in  $[-\ln m, \ln m]$ , one of them will be smaller than  $-\ln m$ . But then the exponential loss  $L(\boldsymbol{\lambda}^t) = (1/m) \sum_j e^{-(M\boldsymbol{\lambda}^t)_j}$  in that round will exceed 1, a contradiction since the losses are non-increasing through rounds, and the loss at the start was 1. Thus, assigning one of these examples the index  $i$ , we have the absolute margin  $|(M\boldsymbol{\lambda}^t)_i|$  is bounded by  $\ln m$  in any round  $t$ . Letting  $M(i)$  denote the  $i$ th row of  $M$ , the step length  $\alpha_t$  in round  $t$  therefore satisfies

$$|\alpha_t| = |M_{ij_t} \alpha_t| = |\langle M(i), \alpha_t \mathbf{e}_{j_t} \rangle| = |(M\boldsymbol{\lambda}^t)_i - (M\boldsymbol{\lambda}^{t-1})_i| \leq |(M\boldsymbol{\lambda}^t)_i| + |(M\boldsymbol{\lambda}^{t-1})_i| \leq 2 \ln m,$$

and the statement of the lemma directly follows.  $\blacksquare$

The next lemma constructs a feature matrix satisfying the properties of Lemma 6 and where additionally the smallest size of a solution achieving  $\phi^* + \varepsilon$  loss is at least  $\Omega(2^m \ln(1/\varepsilon))$ , for some fixed  $\phi^*$  and every  $\varepsilon > 0$ . This implies that when  $\varepsilon$  is a small constant (say  $\varepsilon = 0.01$ ) AdaBoost takes at least  $\Omega(2^m / \ln m)$  steps to get within  $\varepsilon/2$  of the loss achieved by some  $\boldsymbol{\lambda}^*$  with loss  $\phi^* + \varepsilon/2$ . Since  $m$  might be arbitrarily larger than  $\varepsilon^{-1}$ , this shows that a polynomial dependence of the convergence rate on the norm of the competing solution is unavoidable. Further this norm might be exponential in the number of training examples and weak hypotheses in the worst case, and hence the bound  $\text{poly}(\|\boldsymbol{\lambda}^*\|_1, 1/\varepsilon)$  in Theorem 1 cannot be replaced by  $\text{poly}(m, N, 1/\varepsilon)$ .

**Lemma 7** <sup>1</sup> Consider the following matrix  $M$  derived out of abstaining weak hypotheses.  $M$  has  $m+1$  rows labeled  $0, \dots, m$  and  $m$  columns labeled  $1, \dots, m$  (assume  $m \geq 2$ ). The square sub-matrix ignoring row zero is an upper triangular matrix, with 1's on the diagonal,  $-1$ 's above the diagonal, and 0 below the diagonal. Therefore row one is  $(+, -, \dots, -)$ , and row zero is defined to be just the complement of row one. Then, for any  $\varepsilon > 0$ , a loss of  $2/(m+1) + \varepsilon$  is achievable on this dataset, but with large norms

$$\inf \{ \|\boldsymbol{\lambda}\|_1 : L(\boldsymbol{\lambda}) \leq 2/(m+1) + \varepsilon \} \geq \Omega(2^m \ln(1/3\varepsilon)).$$

Therefore, by Lemma 6, the minimum number of rounds required for reaching loss at most  $2/(m+1) + \varepsilon$  is at least  $\Omega(2^m / \ln m) \ln(1/3\varepsilon)$ .

**Proof** We first show  $2/(m+1) + \varepsilon$  loss is achievable for any  $\varepsilon$ . Note that if  $\mathbf{x} = (2^m - 1, 2^{m-1}, 2^{m-2}, \dots, 1)$  then  $M\mathbf{x}$  achieves a margin of 1 on examples 2 through  $m$ , and zero margin on the first two examples. Therefore  $\ln(1/\varepsilon)\mathbf{x}$  achieves loss  $(2 + (m-1)\varepsilon)/(m+1) \leq 2/(m+1) + \varepsilon$ , for any  $\varepsilon > 0$ .

Next we lower bound the norm of solutions achieving loss at most  $2/(m+1) + \varepsilon$ . Observe that since rows 0 and 1 are complementary, any solution's loss on just examples 0 and 1 will add up to at least  $2/(m+1)$ . Therefore, to get within  $2/(m+1) + \varepsilon$ , the margins on examples  $2, \dots, m$  should be at least  $\ln \left( \frac{m-1}{(m+1)\varepsilon} \right) \leq \ln(1/3\varepsilon)$  (for  $m \geq 2$ ). Now, a solution  $\boldsymbol{\lambda}$  gets margin at least  $\ln(1/3\varepsilon)$  on example  $m$  implies  $\lambda_m \geq \ln(1/3\varepsilon)$  (since the other columns get zero margin on it). Since column  $m$  gets margin  $-1$  on example  $m-1$ , and column  $m-1$  is the only column with a positive margin on that example, the previous fact forces

---

1. We thank Nikhil Srivastava for informing us of the matrix used in this lemma.

$\lambda_{m-1} \geq \ln(1/3\varepsilon) + \lambda_m \geq 2\ln(1/3\varepsilon)$ . Continuing this way, we get  $\lambda_i \geq (2^{m+1-i} - 1)\ln(1/3\varepsilon)$  for  $i = m, \dots, 2$ . Hence  $\|\boldsymbol{\lambda}\| \geq \ln(1/3\varepsilon)(2 + \dots + 2^{m-1} - (m-2)) = (2^m + 2 - m)\ln(1/3\varepsilon) \geq \Omega(2^m)\ln(1/3\varepsilon)$ .  $\blacksquare$

In the next section we investigate the optimal dependence on the parameter  $\varepsilon$  and show that  $\Omega(1/\varepsilon)$  number of rounds are necessary.

#### 4. Convergence to optimal loss

In the previous section, our rate bound depended on both the approximation parameter  $\varepsilon$ , as well as the size of the smallest solution achieving the target loss. For many datasets, the optimal target loss  $\inf_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda})$  cannot be realized by any finite solution. In such cases, if we want to bound the number of rounds needed to achieve within  $\varepsilon$  of the optimal loss, the only way to use Theorem 1 is to first decompose the accuracy parameter  $\varepsilon$  into two parts  $\varepsilon = \varepsilon_1 + \varepsilon_2$ , find some finite solution  $\boldsymbol{\lambda}^*$  achieving within  $\varepsilon_1$  of the optimal loss, and then use the bound  $\text{poly}(1/\varepsilon_2, \|\boldsymbol{\lambda}^*\|_1)$  to achieve at most  $L(\boldsymbol{\lambda}^*) + \varepsilon_2 = \inf_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}) + \varepsilon$  loss. However, this introduces implicit dependence on  $\varepsilon$  through  $\|\boldsymbol{\lambda}^*\|_1$  which may not be immediately clear. In this section, we show bounds of the form  $C/\varepsilon$ , where the constant  $C$  depends only on the feature matrix  $M$ , and not on  $\varepsilon$ .

**Theorem 8** *AdaBoost reaches within  $\varepsilon$  of the optimal loss in at most  $C/\varepsilon$  rounds, where  $C$  only depends on the feature matrix.*

Additionally, we show that this dependence on  $\varepsilon$  is optimal in Lemma 17 of the Appendix, where  $\Omega(1/\varepsilon)$  rounds are shown to be necessary for converging to within  $\varepsilon$  of the optimal loss on a certain dataset. Finally, we note that the lower bounds in the previous section indicate that  $C$  can be  $\Omega(2^m)$  in the worst case for integer matrices (although it will typically be much smaller), and hence this bound, though stronger than that of Theorem 1 with respect to  $\varepsilon$ , cannot be used to prove the conjecture in (Schapire, 2010), since the constant is not polynomial in the number of examples  $m$ .

Our techniques build upon earlier work on the rate of convergence of AdaBoost, which have mainly considered two particular cases. In the first case, the *weak learning assumption* holds, that is, the edge in each round is at least some fixed constant. In this situation, Freund and Schapire (1997) and Schapire and Singer (1999) show that the optimal loss is zero, no solution with finite size can achieve this loss, but AdaBoost achieves at most  $\varepsilon$  loss within  $O(\ln(1/\varepsilon))$  rounds. In the second case some finite combination of the weak classifiers achieves the optimal loss, and Rätsch et al. (2002), using results from (Luo and Tseng, 1992), show that AdaBoost achieves within  $\varepsilon$  of the optimal loss again within  $O(\ln(1/\varepsilon))$  rounds.

Here we consider the most general situation, where the weak learning assumption may fail to hold, and yet no finite solution may achieve the optimal loss. The dataset used in Lemma 17 and shown in Figure 2 exemplifies this situation. Our main technical contribution shows that the examples in any dataset can be partitioned into a *zero-loss set* and *finite-margin set*, such that a certain form of the weak learning assumption holds within the zero-loss set, while the optimal loss considering only the finite-margin set can be obtained by some finite solution. The two partitions provide different ways of making progress in

every round, and one of the two kinds of progress will always be sufficient for us to prove Theorem 8.

We next state our decomposition result, illustrate it with an example, and then state several lemmas quantifying the nature of the progress we can make in each round. Using these lemmas, we prove Theorem 8.

**Lemma 9** (*Decomposition Lemma*) *For any dataset, there exists a partition of the set of training examples  $X$  into a (possibly empty) zero-loss set  $A$  and a (possibly empty) finite-margin set  $F = A^c \triangleq X \setminus A$  such that the following hold simultaneously :*

1. *For some positive constant  $\gamma > 0$ , there exists some vector  $\boldsymbol{\eta}^\dagger$  with unit  $\ell_1$ -norm  $\|\boldsymbol{\eta}^\dagger\|_1 = 1$  that attains at least  $\gamma$  margin on each example in  $A$ , and exactly zero margin on each example in  $F$*

$$\forall i \in A : (M\boldsymbol{\eta}^\dagger)_i \geq \gamma, \quad \forall i \in F : (M\boldsymbol{\eta}^\dagger)_i = 0.$$

2. *The optimal loss considering only examples within  $F$  is achieved by some finite combination  $\boldsymbol{\eta}^*$ .*
3. *(Corollary to Item 2) There is a constant  $\mu_{\max} < \infty$ , such that for any combination  $\boldsymbol{\eta}$  with bounded loss on the finite-margin set,  $\sum_{i \in F} e^{-(M\boldsymbol{\eta})_i} \leq m$ , the margin  $(M\boldsymbol{\eta})_i$  for any example  $i$  in  $F$  lies in the bounded interval  $[-\ln m, \mu_{\max}]$ .*

A proof is deferred to the next section. The Decomposition Lemma immediately implies that the vector  $\boldsymbol{\eta}^* + \infty \cdot \boldsymbol{\eta}^\dagger$ , defined as the limit of  $\lim_{c \rightarrow \infty} (\boldsymbol{\eta}^* + c\boldsymbol{\eta}^\dagger)$ , is an optimal solution, achieving zero loss on the zero-loss set, but only finite margins (and hence positive losses) on the finite-margin set (thereby justifying the names).

Before proceeding, we give an example dataset and indicate the zero-loss set, finite-margin set,  $\boldsymbol{\eta}^*$  and  $\boldsymbol{\eta}^\dagger$  to illustrate our definitions. Consider a dataset with three examples  $\{a, b, c\}$  and two hypotheses  $\{h_1, h_2\}$  and the following feature matrix  $M$ .

	$h_1$	$h_2$
$a$	+	-
$b$	-	+
$c$	+	+

Figure 2:

Here + means correct ( $M_{ij} = +1$ ) and - means wrong ( $M_{ij} = -1$ ). The optimal solution is  $\infty \cdot (h_1 + h_2)$  with a loss of  $2/3$ . The finite-margin set is  $\{a, b\}$ , the zero-loss set is  $\{c\}$ ,  $\boldsymbol{\eta}^\dagger = (1/2, 1/2)$  and  $\boldsymbol{\eta}^* = (0, 0)$ ; for this dataset these are unique. This dataset also serves as a lower-bound example in Lemma 17, where we show that  $0.22/\varepsilon$  rounds are necessary for AdaBoost to achieve less than  $(2/3) + \varepsilon$  loss on it.

Before providing proofs, we introduce some notation. By  $\|\cdot\|$  we will mean  $\ell_2$ -norm; every other norm will have an appropriate subscript, such as  $\|\cdot\|_1, \|\cdot\|_\infty$ , etc. The set of all training examples will be denoted by  $X$ . By  $\ell^\lambda(i)$  we mean the exp-loss  $e^{-(M\lambda)_i}$  on example  $i$ . For any subset  $S \subset X$  of examples,  $\ell^\lambda(S) = \sum_{i \in S} \ell^\lambda(i)$  denotes the total exp-loss on

the set  $S$ . Notice  $L(\boldsymbol{\lambda}) = (1/m)\ell^{\boldsymbol{\lambda}}(X)$ , and that  $D_t(i) = \ell^{\boldsymbol{\lambda}^t}(i)/\ell^{\boldsymbol{\lambda}^t}(X)$ , where  $\boldsymbol{\lambda}^t$  is the combination found by AdaBoost in round  $t$ . By  $\delta_S(\boldsymbol{\eta}; \boldsymbol{\lambda})$  we mean the edge obtained on the set  $S$  by the vector  $\boldsymbol{\eta}$ , when the weights over the examples are given by  $\ell^{\boldsymbol{\lambda}}(\cdot)/\ell^{\boldsymbol{\lambda}}(S)$ :

$$\delta_S(\boldsymbol{\eta}; \boldsymbol{\lambda}) = \left| \frac{1}{\ell^{\boldsymbol{\lambda}}(S)} \sum_{i \in S} \ell^{\boldsymbol{\lambda}}(i)(M\boldsymbol{\eta})_i \right|.$$

In the rest of the section, by loss we mean the unnormalized loss  $\ell^{\boldsymbol{\lambda}}(X) = mL(\boldsymbol{\lambda})$  and study convergence to within  $\varepsilon$  of the optimal unnormalized loss  $\inf_{\boldsymbol{\lambda}} \ell^{\boldsymbol{\lambda}}(X)$ , henceforth denoted by  $K$ . Note that this is the same as converging to within  $\varepsilon/m$  of the optimal normalized loss, that is to within  $\inf_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}) + \varepsilon/m$ . Hence a  $C'/\varepsilon$  bound for the unnormalized loss translates to a  $C'm^{-1}/\varepsilon$  bound for the normalized loss and vice versa, and does not affect the result in Theorem 8. The progress due to the zero-loss set is now immediate from Item 1 of the Decomposition Lemma:

**Lemma 10** *In any round  $t$ , the maximum edge  $\delta_t$  is at least  $\gamma \left\{ \frac{\ell^{\boldsymbol{\lambda}^{t-1}}(A)}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)} \right\}$ , where  $\gamma$  is as in Item 1 of the Decomposition Lemma.*

**Proof** Recall the distribution  $D_t$  created by AdaBoost in round  $t$  puts weight  $D_t(i) = \ell^{\boldsymbol{\lambda}^{t-1}}(i)/\ell^{\boldsymbol{\lambda}^{t-1}}(X)$  on each example  $i$ . From Item 1 we get

$$\delta_X(\boldsymbol{\eta}^\dagger; \boldsymbol{\lambda}^{t-1}) = \left| \frac{1}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)} \sum_{i \in X} \ell^{\boldsymbol{\lambda}^{t-1}}(i)(M\boldsymbol{\eta}^\dagger)_i \right| = \frac{1}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)} \sum_{i \in A} \gamma \ell^{\boldsymbol{\lambda}^{t-1}}(i) = \gamma \left\{ \frac{\ell^{\boldsymbol{\lambda}^{t-1}}(A)}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)} \right\}.$$

Next define  $p$  to be a distribution on the columns  $\{1, \dots, N\}$  of  $M$  which puts probability  $p(j)$  proportional to  $|\boldsymbol{\eta}_j^\dagger|$  on column  $j$ . Since  $(M\boldsymbol{\eta}^\dagger)_i = \sum_j \eta_j^\dagger (M\mathbf{e}_j)_i$ , we may rewrite the edge  $\delta_X(\boldsymbol{\eta}^\dagger; \boldsymbol{\lambda}^{t-1})$  as follows

$$\begin{aligned} \delta_X(\boldsymbol{\eta}^\dagger; \boldsymbol{\lambda}^{t-1}) &= \left| \frac{1}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)} \sum_{i \in X} \ell^{\boldsymbol{\lambda}^{t-1}}(i) \sum_j \eta_j^\dagger (M\mathbf{e}_j)_i \right| \\ &= \left| \sum_j \eta_j^\dagger \frac{1}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)} \sum_{i \in X} \ell^{\boldsymbol{\lambda}^{t-1}}(i)(M\mathbf{e}_j)_i \right| = \left| \sum_j \eta_j^\dagger \delta_X(\mathbf{e}_j; \boldsymbol{\lambda}^{t-1}) \right| \leq \sum_j |\eta_j^\dagger| \delta_X(\mathbf{e}_j; \boldsymbol{\lambda}^{t-1}). \end{aligned}$$

Since the  $\ell_1$ -norm of  $\boldsymbol{\eta}^\dagger$  is 1, the weights  $|\eta_j^\dagger|$  form some distribution  $p$  over the columns  $1, \dots, N$ . We may therefore conclude

$$\gamma \left\{ \frac{\ell^{\boldsymbol{\lambda}^{t-1}}(A)}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)} \right\} \leq \delta_X(\boldsymbol{\eta}^\dagger; \boldsymbol{\lambda}^{t-1}) \leq \mathbb{E}_{j \sim p} [\delta_X(\mathbf{e}_j; \boldsymbol{\lambda}^{t-1})] \leq \max_j \delta_X(\mathbf{e}_j; \boldsymbol{\lambda}^{t-1}) \leq \delta_t.$$

■

If the set  $F$  were empty, then Lemma 10 implies an edge of  $\gamma$  is available in each round. This in fact means that the weak learning assumption holds, and using (4), we can

show a  $O(\ln(1/\varepsilon)\gamma^{-2})$  bound matching the rate bounds in (Freund and Schapire, 1997) and (Schapire and Singer, 1999). So henceforth, we assume that  $F$  is non-empty. Note that this implies that the optimal loss  $K$  is at least 1 (since any solution will get non-positive margin on some example in  $F$ ), a fact we will use later in the proofs.

Lemma 10 says that the edge is large if the loss on the zero-loss set is large. On the other hand, when it is small, Lemmas 11 and 12 together show how AdaBoost can make good progress using the finite margin set. Lemma 11 uses second order methods to show how progress is made in the case where there is a finite solution; such arguments may have appeared in earlier work.

**Lemma 11** *Suppose  $\lambda$  is a combination such that  $m \geq \ell^\lambda(F) \geq K$ . Then in some coordinate direction the edge is at least  $\sqrt{C_0(\ell^\lambda(F) - K)/\ell^\lambda(F)}$ , where  $C_0$  is a constant depending only on the feature matrix  $M$ .*

**Proof** Let  $M_F \in \mathbb{R}^{|F| \times N}$  be the matrix  $M$  restricted to only the rows corresponding to the examples in  $F$ . Choose  $\eta$  such that  $\lambda + \eta = \eta^*$  is an optimal solution over  $F$ . Without loss of generality assume that  $\eta$  lies in the orthogonal subspace of the null-space  $\{\mathbf{u} : M_F \mathbf{u} = \mathbf{0}\}$  of  $M_F$  (since we can translate  $\eta^*$  along the null space if necessary for this to hold). If  $\eta = \mathbf{0}$ , then  $\ell^\lambda(F) = K$  and we are done. Otherwise  $\|M_F \eta\| \geq \lambda_{\min} \|\eta\|$ , where  $\lambda_{\min}^2$  is the smallest positive eigenvalue of the symmetric matrix  $M_F^T M_F$  (exists since  $M_F \eta \neq \mathbf{0}$ ). Now define  $f : [0, 1] \rightarrow \mathbb{R}$  as the loss along the (rescaled) segment  $[\eta^*, \lambda]$

$$f(x) \triangleq \ell^{(\eta^* - x\eta)}(F) = \sum_{i \in F} \ell^{\eta^*}(i) e^{x(M\eta)_i}.$$

This implies that  $f(0) = K$  and  $f(1) = \ell^\lambda(F)$ . Notice that the first and second derivatives of  $f(x)$  are given by

$$f'(x) = \sum_{i \in F} (M_F \eta)_i \ell^{(\eta^* - x\eta)}(i), \quad f''(x) = \sum_{i \in F} (M_F \eta)_i^2 \ell^{(\eta^* - x\eta)}(i).$$

We next lower bound possible values of the second derivative. We define a distribution  $q$  on examples  $\{1, \dots, m\}$  which puts probability proportional to  $(M_F \eta)_i^2$  on example  $i$ . Then we may rewrite the second derivative as

$$f''(x) = \|M_F \eta\|^2 \mathbb{E}_{i \sim q} [\ell^{(\eta^* - x\eta)}(i)] \geq \|M_F \eta\|^2 \min_i \ell^{(\eta^* - x\eta)}(i).$$

Since both  $\lambda = \eta^* - \eta$ , and  $\eta^*$  suffer total loss at most  $m$ , therefore, by convexity, so does  $\eta^* - x\eta$  for any  $x \in [0, 1]$ . Hence we may apply Item 3 of the Decomposition Lemma to the vector  $\eta^* - x\eta$ , for any  $x \in [0, 1]$ , to conclude that  $\ell^{(\eta^* - x\eta)}(i) = \exp\{-(M_F(\eta^* - x\eta))_i\} \geq e^{-\mu_{\max}}$  on every example  $i$ . Therefore we have,

$$f''(x) \geq \|M_F \eta\|^2 e^{-\mu_{\max}} \geq \lambda_{\min}^2 e^{-\mu_{\max}} \|\eta\|^2 \text{ (by choice of } \eta\text{)}.$$

A standard second-order result is (see e.g. Boyd and Vandenberghe, 2004, eqn. (9.9))

$$|f'(1)|^2 \geq 2 \left( \inf_{x \in [0, 1]} f''(x) \right) \{f(1) - f(0)\}.$$

Collecting our results so far, we get

$$\sum_{i \in F} \ell^\lambda(i)(M\eta)_i = |f'(1)| \geq \|\eta\| \sqrt{2\lambda_{\min}^2 e^{-\mu_{\max}} \{\ell^\lambda(F) - K\}}.$$

Next let  $\tilde{\eta} = \eta / \|\eta\|_1$  be  $\eta$  rescaled to have unit  $\ell_1$  norm. Then we have

$$\sum_{i \in F} \ell^\lambda(i)(M\tilde{\eta})_i = \frac{1}{\|\eta\|_1} \sum_i \ell^\lambda(i)(M\eta)_i \geq \frac{\|\eta\|}{\|\eta\|_1} \sqrt{2\lambda_{\min}^2 e^{-\mu_{\max}} \{\ell^\lambda(F) - K\}}.$$

Applying the Cauchy-Schwartz inequality, we may lower bound  $\frac{\|\eta\|}{\|\eta\|_1}$  by  $1/\sqrt{N}$  (since  $\eta \in \mathbb{R}^N$ ). Along with the fact  $\ell^\lambda(F) \leq m$ , we may write

$$\frac{1}{\ell^\lambda(F)} \sum_{i \in F} \ell^\lambda(i)(M\tilde{\eta})_i \geq \sqrt{2\lambda_{\min}^2 N^{-1} m^{-1} e^{-\mu_{\max}}} \sqrt{\{\ell^\lambda(F) - K\} / \ell^\lambda(F)}$$

If we define  $p$  to be a distribution on the columns  $\{1, \dots, N\}$  of  $M_F$  which puts probability  $p(j)$  proportional to  $|\tilde{\eta}_j|$  on column  $j$ , then we have

$$\frac{1}{\ell^\lambda(F)} \sum_{i \in F} \ell^\lambda(i)(M\tilde{\eta})_i \leq \mathbb{E}_{j \sim p} \left| \frac{1}{\ell^\lambda(F)} \sum_{i \in F} \ell^\lambda(i)(M\mathbf{e}_j)_i \right| \leq \max_j \left| \frac{1}{\ell^\lambda(F)} \sum_{i \in F} \ell^\lambda(i)(M\mathbf{e}_j)_i \right|.$$

Notice the quantity inside max is precisely the edge  $\delta_F(\mathbf{e}_j; \lambda)$  in direction  $j$ . Combining everything, the maximum possible edge is

$$\max_j \delta_F(\mathbf{e}_j; \lambda) \geq \sqrt{C_0 \{\ell^\lambda(F) - K\} / \ell^\lambda(F)},$$

where we define  $C_0 = 2m^{-1}N^{-1}\lambda_{\min}^2 e^{-\mu_{\max}}$ . ■

**Lemma 12** Suppose, at some stage of boosting, the combination found by AdaBoost is  $\lambda$ , and the loss is  $K + \theta$ . Let  $\Delta\theta$  denote the drop in the suboptimality  $\theta$  after one more round; i.e. the loss after one more round is  $K + \theta - \Delta\theta$ . Then, there are constants  $C_1, C_2$  depending only on the feature matrix (and not on  $\theta$ ), such that if  $\ell^\lambda(A) < C_1\theta$ , then  $\Delta\theta \geq C_2\theta$ .

**Proof** Let  $\lambda$  be the current solution found by boosting. Using Lemma 11 pick a direction  $j$  in which the edge  $\delta_F(\mathbf{e}_j; \lambda)$  restricted to the finite loss set is at least  $\sqrt{2C_0(\ell^\lambda(F) - K)/\ell^\lambda(F)}$ . We can bound the edge  $\delta_X(\mathbf{e}_j; \lambda)$  on the entire set of examples as follows

$$\begin{aligned} \delta_X(\mathbf{e}_j; \lambda) &= \frac{1}{\ell^\lambda(X)} \left| \sum_{i \in F} \ell^\lambda(i)(M\mathbf{e}_j)_i + \sum_{i \in A} \ell^\lambda(i)(M\mathbf{e}_j)_i \right| \\ &\geq \frac{1}{\ell^\lambda(X)} \left( |\ell^\lambda(F)\delta_F(\mathbf{e}_j; \lambda)| - \sum_{i \in A} \ell^\lambda(i) \right) \\ &\geq \frac{1}{\ell^\lambda(X)} \left( \sqrt{2C_0(\ell^\lambda(F) - K)\ell^\lambda(F)} - \ell^\lambda(A) \right). \end{aligned}$$

Now,  $\ell^\lambda(A) < C_1\theta$ , and  $\ell^\lambda(F) - K = \theta - \ell^\lambda(A) \geq (1 - C_1)\theta$ . Further, we will choose  $C_1 < 1$ , so that  $\ell^\lambda(F) \geq K \geq 1$ . Hence, the previous inequality implies

$$\delta_X(\mathbf{e}_j; \boldsymbol{\lambda}) \geq \frac{1}{K + \theta} \left( \sqrt{2C_0(1 - C_1)\theta} - C_1\theta \right).$$

Set  $C_1 = \min \left\{ 1/2, (1/4)\sqrt{C_0/m} \right\}$ . Using  $\theta \leq K + \theta = \ell^\lambda(X) \leq m$ , we can bound the square of the term in brackets on the previous line as

$$\begin{aligned} \left( \sqrt{2C_0(1 - C_1)\theta} - C_1\theta \right)^2 &\geq 2C_0(1 - C_1)\theta - 2C_1\theta\sqrt{2C_0(1 - C_1)\theta} \\ &\geq 2C_0(1 - 1/2)\theta - 2 \left\{ (1/4)\sqrt{C_0/m} \right\} \theta\sqrt{2C_0(1 - 1/2)m} \geq C_0\theta/2. \end{aligned}$$

So, if  $\delta$  is the maximum edge in any direction, then  $\delta \geq \delta_X(\mathbf{e}_j; \boldsymbol{\lambda}) \geq \sqrt{C_0\theta/(2m(K + \theta))}$  (again using  $1/(K + \theta) \leq \sqrt{(K + \theta)m}$ ). Therefore the loss after one more step is at most  $(K + \theta)\sqrt{1 - \delta^2} \leq (K + \theta)(1 - \delta^2/2) \leq K + \theta - \frac{C_0}{2m}\theta$ . Setting  $C_2 = C_0/(2m)$  completes the proof. ■

**Proof of Theorem 8.** At any stage of boosting, let  $\boldsymbol{\lambda}$  be the current combination, and  $K + \theta$  be the current loss. We show that the new loss is at most  $K + \theta - \Delta\theta$  for  $\Delta\theta \geq C_3\theta^2$  for some constant  $C_3$  depending only on the dataset (and not  $\theta$ ). Lemma 18 and some algebra will then complete the proof.

To show this, either  $\ell^\lambda(A) < C_1\theta$ , in which case Lemma 12 applies, and  $\Delta\theta \geq C_2\theta \geq (C_2/m)\theta^2$  (since  $\theta = \ell^\lambda(X) - K \leq m$ ). Or  $\ell^\lambda(A) \geq C_1\theta$ , in which case applying Lemma 10 yields  $\delta \geq \gamma C_1\theta/\ell^\lambda(X) \geq (\gamma C_1/m)\theta$ . By (4),  $\Delta\theta \geq \ell^\lambda(X)(1 - \sqrt{1 - \delta^2}) \geq \ell^\lambda(X)\delta^2/2 \geq (K/2)(\gamma C_1/m)^2\theta^2$ . Using  $K \geq 1$  and choosing  $C$  appropriately gives the required condition. ■

#### 4.1. Proof of the Decomposition Lemma

Throughout this section we only consider (unless otherwise stated) *admissible* combinations  $\boldsymbol{\lambda}$  of weak classifiers, which have loss  $\ell^\lambda(X)$  bounded by  $m$  (since such are the ones found by boosting). We prove Lemma 9 in three steps. We begin with a simple lemma that rigorously defines the zero-loss and finite-margin sets.

**Lemma 13** *For any sequence  $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots$ , of admissible combinations of weak classifiers, we can find a subsequence  $\boldsymbol{\eta}_{(1)} = \boldsymbol{\eta}_{t_1}, \boldsymbol{\eta}_{(2)} = \boldsymbol{\eta}_{t_2}, \dots$ , whose losses converge to zero on all examples in some fixed (possibly empty) subset  $S$  (the zero-loss set), and losses bounded away from zero in its complement  $X \setminus S$  (the finite-margin set)*

$$\forall x \in S : \lim_{t \rightarrow \infty} \ell^{\boldsymbol{\eta}(t)}(x) = 0, \quad \forall x \in X \setminus S : \inf_i \ell^{\boldsymbol{\eta}(t)}(x) > 0. \quad (8)$$

**Proof** We will build a zero-loss set and the final subsequence incrementally. Initially the set is empty. Pick the first example. If the infimal loss ever attained on the example in the sequence is bounded away from zero, then we do not add it to the set. Otherwise we add it, and consider only the subsequence whose  $t^{\text{th}}$  element attains loss less than  $1/t$  on the

example. Beginning with this subsequence, we now repeat with other examples. The final sequence is the required subsequence, and the examples we have added form the zero-loss set. ■

We apply Lemma 13 to some admissible sequence converging to the optimal loss (e.g. the one found by AdaBoost). Let us call the resulting subsequence  $\boldsymbol{\eta}_{(t)}^*$ , the obtained zero-loss set  $A$ , and the finite-margin set  $F = X \setminus A$ . The next lemma shows how to extract a single combination out of the sequence  $\boldsymbol{\eta}_{(t)}^*$  that satisfies the properties in Item 1 of the Decomposition Lemma.

**Lemma 14** *Suppose  $M$  is the feature matrix,  $A$  is a subset of the examples, and  $\boldsymbol{\eta}_{(1)}, \boldsymbol{\eta}_{(2)}, \dots$ , is a sequence of combinations of weak classifiers such that  $A$  is its zero loss set, and  $X \setminus A$  its finite loss set, that is, (8) holds. Then there is a combination  $\boldsymbol{\eta}^\dagger$  of weak classifiers that achieves positive margin on every example in  $A$ , and zero margin on every example in its complement  $X \setminus A$*

$$(M\boldsymbol{\eta}^\dagger)_i \begin{cases} > 0 & \text{if } i \in A, \\ = 0 & \text{if } i \in X \setminus A. \end{cases}$$

**Proof** Since the  $\boldsymbol{\eta}_{(i)}$  achieve arbitrarily large positive margins on  $A$ ,  $\|\boldsymbol{\eta}_{(i)}\|$  will be unbounded, and it will be hard to extract a useful single solution out of them. On the other hand, the rescaled combinations  $\boldsymbol{\eta}_{(i)}/\|\boldsymbol{\eta}_{(i)}\|$  lie on a compact set, and therefore have a limit point, which might have useful properties. We formalize this next.

We use induction on the total number of training examples  $|X|$ . If  $X$  is zero, then the lemma holds vacuously for any  $\boldsymbol{\eta}^\dagger$ . Assume inductively for all  $X$  of size less than  $m > 0$ , and consider  $X$  of size  $m$ . Since translating a vector along the null space of  $M$ ,  $\ker M = \{\mathbf{x} : M\mathbf{x} = \mathbf{0}\}$  has no effect on the margins produced by the vector, assume without loss of generality that the  $\boldsymbol{\eta}_{(t)}$ 's are orthogonal to  $\ker M$ . Also, since the margins produced on the zero loss set are unbounded, so are the norms of  $\boldsymbol{\eta}_{(t)}$ . Therefore assume (by picking a subsequence and relabeling if necessary) that  $\|\boldsymbol{\eta}_{(t)}\| > t$ . Let  $\boldsymbol{\eta}'$  be a limit point of the sequence  $\boldsymbol{\eta}_{(t)}/\|\boldsymbol{\eta}_{(t)}\|$ , a unit vector that is also orthogonal to the null-space. Then firstly  $\boldsymbol{\eta}'$  achieves non-negative margin on every example; otherwise by continuity for some extremely large  $t$ , the margin of  $\boldsymbol{\eta}_{(t)}/\|\boldsymbol{\eta}_{(t)}\|$  on that example is also negative and bounded away from zero, and therefore  $\boldsymbol{\eta}_{(t)}$ 's loss is more than  $m$ , a contradiction to admissibility. Secondly, the margin of  $\boldsymbol{\eta}'$  on each example in  $X \setminus A$  is zero; otherwise, by continuity, for arbitrarily large  $t$  the margin of  $\boldsymbol{\eta}_{(t)}/\|\boldsymbol{\eta}_{(t)}\|$  on an example in  $X \setminus A$  is positive and bounded away from zero, and hence that example attains arbitrarily small loss in the sequence, a contradiction to (8). Finally, if  $\boldsymbol{\eta}'$  achieves zero margin everywhere in  $A$ , then  $\boldsymbol{\eta}'$ , being orthogonal to the null-space, must be  $\mathbf{0}$ , a contradiction since  $\boldsymbol{\eta}'$  is a unit vector. Therefore  $\boldsymbol{\eta}'$  must achieve positive margin on some non-empty subset  $Z$  of  $A$ , and zero margins on every other example.

Next we use induction on the reduced set of examples  $X' = X \setminus Z$ . Since  $Z$  is non-empty,  $|X'| < m$ . Further, using the same sequence  $\boldsymbol{\eta}_{(t)}$ , the zero-loss and finite-loss sets, restricted to  $X'$ , are  $A' = A \setminus Z$  and  $(X \setminus A) \setminus Z = X' \setminus A'$  (since  $Z \subseteq A$ ) =  $X' \setminus A'$ . By the inductive hypothesis, there exists some  $\boldsymbol{\eta}''$  which achieves positive margins on  $A'$ , and zero margins on  $X' \setminus A' = X \setminus A$ . Therefore, by setting  $\boldsymbol{\eta}^\dagger = \boldsymbol{\eta}' + c\boldsymbol{\eta}''$  for a large enough  $c$ , we can achieve

the desired properties. ■

Applying Lemma 14 to the sequence  $\boldsymbol{\eta}_{(t)}^*$  yields some convex combination  $\boldsymbol{\eta}^\dagger$  having margin at least  $\gamma > 0$  (for some  $\gamma$ ) on  $A$  and zero margin on its complement, proving Item 1 of the Decomposition Lemma. The next lemma proves Item 2.

**Lemma 15** *There is a (finite) combination  $\boldsymbol{\eta}^*$  which achieves the same margins on  $F$  as the optimal solution.*

**Proof** The existence of  $\boldsymbol{\eta}^\dagger$  with properties as in Lemma 14 implies that the optimal loss is the same whether considering all the examples, or just examples in  $F$ . Therefore it suffices to show the existence of finite  $\boldsymbol{\eta}^*$  that achieves loss  $K$  on  $F$ , that is,  $\ell^{\boldsymbol{\eta}^*}(F) = K$ .

Recall  $M_F$  denotes the matrix  $M$  restricted to the rows corresponding to examples in  $F$ . Let  $\ker M_F = \{\mathbf{x} : M_F \mathbf{x} = 0\}$  be the null-space of  $M_F$ . Let  $\boldsymbol{\eta}^{(t)}$  be the projection of  $\boldsymbol{\eta}_{(t)}^*$  onto the orthogonal subspace of  $\ker M_F$ . Then the losses  $\ell^{\boldsymbol{\eta}^{(t)}}(F) = \ell^{\boldsymbol{\eta}_{(t)}^*}(F)$  converge to the optimal loss  $K$ . If  $M_F$  is identically zero, then each  $\boldsymbol{\eta}^{(t)} = \mathbf{0}$ , and then  $\boldsymbol{\eta}^* = \mathbf{0}$  has loss  $K$  on  $F$ . Otherwise, let  $\lambda^2$  be the smallest positive eigenvalue of  $M_F^T M_F$ . Then  $\|M \boldsymbol{\eta}^{(t)}\| \geq \lambda \|\boldsymbol{\eta}^{(t)}\|$ . By the definition of finite margin set,  $\inf_{t \rightarrow \infty} \ell^{\boldsymbol{\eta}^{(t)}}(F) = \inf_{t \rightarrow \infty} \ell^{\boldsymbol{\eta}_{(t)}^*}(F) > 0$ . Therefore, the margins  $\|M \boldsymbol{\eta}^{(t)}\|$  are bounded, and hence the  $\boldsymbol{\eta}^{(t)}$  are also bounded in norm. Therefore they have a (finite) limit point  $\boldsymbol{\eta}^*$  which must have loss  $K$  over  $F$ . ■

As a corollary, we prove Item 3.

**Lemma 16** *There is a constant  $\mu_{\max} < \infty$ , such that for any combination  $\boldsymbol{\eta}$  that achieves bounded loss on the finite-margin set,  $\ell^{\boldsymbol{\eta}}(F) \leq m$ , the margin  $(M\boldsymbol{\eta})_i$  for any example  $i$  in  $F$  lies in the bounded interval  $[-\ln m, \mu_{\max}]$ .*

**Proof** The loss  $\ell^{\boldsymbol{\eta}}(F)$  at most  $m$  implies no margin may be less than  $-\ln m$ . If Item 3 of the Decomposition Lemma were false, then for some example  $x \in F$  there exists a sequence of combinations of weak classifiers, whose  $t^{\text{th}}$  element achieves more than margin  $t$  on  $x$  but has loss at most  $m$  on  $F$ . Applying Lemma 13 we can find a subsequence  $\boldsymbol{\lambda}^{(t)}$  whose tail achieves zero-loss on some non-empty subset  $S$  of  $F$  containing  $x$ , and bounded margins in  $F \setminus S$ . Applying Lemma 14 to  $\boldsymbol{\lambda}^{(t)}$  we get some convex combination  $\boldsymbol{\lambda}^\dagger$  which has positive margins on  $S$  and zero margin on  $F \setminus S$ . Let  $\boldsymbol{\eta}^*$  be as in Lemma 15, a finite combination achieving the optimal loss on  $F$ . Then  $\boldsymbol{\eta}^* + \infty \cdot \boldsymbol{\lambda}^\dagger$  achieves the same loss on every example in  $F \setminus S$  as the optimal solution  $\boldsymbol{\eta}^*$ , but zero loss for examples in  $S$ . This solution is strictly better than  $\boldsymbol{\eta}^*$  on  $F$ , a contradiction to the optimality of  $\boldsymbol{\eta}^*$ . ■

## References

- Peter L. Bartlett and Mikhail Traskin. AdaBoost is consistent. *Journal of Machine Learning Research*, 8:2347–2368, 2007.
- Peter J. Bickel, Ya’acov Ritov, and Alon Zakai. Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7:705–732, 2006.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Leo Breiman. Prediction games and arcing classifiers. *Neural Computation*, 11(7):1493–1517, 1999.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1/2/3), 2002.

Marcus Frean and Tom Downs. A simple cost function for boosting. Technical report, Department of Computer Science and Electrical Engineering, University of Queensland, 1998.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28(2):337–374, April 2000.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), October 2001.

David G. Luenberger and Yinyu Ye. *Linear and nonlinear programming*. Springer, third edition, 2008.

Z. Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, January 1992.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Functional gradient techniques for combining hypotheses. In *Advances in Large Margin Classifiers*. MIT Press, 1999.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems 12*, 2000.

T. Onoda, G. Rätsch, and K.-R. Müller. An asymptotic analysis of AdaBoost in the binary classification case. In *Proceedings of the 8th International Conference on Artificial Neural Networks*, pages 195–200, 1998.

G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for AdaBoost. *Machine Learning*, 42(3):287–320, 2001.

Gunnar Rätsch and Manfred K. Warmuth. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6:2131–2152, 2005.

THE RATE OF CONVERGENCE OF ADABoost

- Gunnar Rätsch, Sebastian Mika, and Manfred K. Warmuth. On the convergence of leveraging. In *Advances in Neural Information Processing Systems 14*, 2002.
- Cynthia Rudin, Robert E. Schapire, and Ingrid Daubechies. Analysis of boosting algorithms using the smooth margin function. *Annals of Statistics*, 35(6):2723–2768, 2007.
- Robert E. Schapire. The convergence rate of AdaBoost. In *The 23rd Conference on Learning Theory*, 2010. open problem.
- Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, December 1999.
- Shai Shalev-Shwartz and Yoram Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. In *21st Annual Conference on Learning Theory*, 2008.
- Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.
- Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 33(4):1538–1579, 2005.

**Lemma 17** *To get within  $\varepsilon < 0.1$  of the optimum loss on the dataset in Table 2, AdaBoost takes at least  $0.22/\varepsilon$  steps.*

**Proof** Note that optimum loss is  $2/3$ , and we are bounding the number of rounds necessary to get within  $(2/3) + \varepsilon$  loss for  $\varepsilon < 0.1$ . We begin by showing that for rounds  $t \geq 3$ , the edge achieved is  $1/t$ . First observe that the edges in rounds 1 and 2 are  $1/3$  and  $1/2$ . Our claim will follow from the following stronger claim. Let  $w_a^t, w_b^t, w_c^t$  denote the normalized-losses (adding up to 1) or weights on examples  $a, b, c$  at the beginning of round  $t$ , and  $\delta_t$  the edge in round  $t$ . Then for  $t \geq 2$ ,

1. Either  $1/2 = w_a^t$  or  $1/2 = w_b^t$ .
2.  $\delta_{t+1} = \delta_t/(1 + \delta_t)$ .

Proof by induction. Base case may be checked. Suppose the inductive assumption holds for  $t$ . Assume without loss of generality that  $1/2 = w_a^t > w_b^t > w_c^t$ . Then in round  $t$ ,  $h_a$  gets picked, the edge  $\delta_t = 2w_c^t$ , and  $w_b^{t+1} = 1/2, w_c^{t+1} = (w_c^t/2)/(1/2 + w_c^t) = w_c^t/(1 + 2w_c^t)$ . Hence, in round  $t + 1$   $h_b$  gets picked and we get edge  $\delta_{t+1} = 2w_c^t/(1 + 2w_c^t) = \delta_t/(1 + \delta_t)$ . Proof follows by induction. Note the recurrence on  $\delta_t$  yields  $\delta_t = 1/t$  for  $t \geq 3$ .

Next we find the loss after each iteration. The loss after  $T$  rounds is

$$\sqrt{1 - (1/3)^2} \prod_{i=2}^T \sqrt{1 - 1/t^2}$$

and can be computed as follows. Notice that in the following list

$$\begin{aligned} 1 - (1/2)^2 &= (1 \cdot 3)/(2 \cdot 2), \\ 1 - (1/3)^2 &= (2 \cdot 4)/(3 \cdot 3), \\ 1 - (1/4)^2 &= (3 \cdot 5)/(4 \cdot 4), \\ \dots &= \dots, \end{aligned}$$

the middle denominator  $(3 \cdot 3)$  gets canceled by the right term of the first numerator and the left term of the third denominator. Continuing this way, the product till term  $1 - (1/T)^2$  is  $(1/2) \{(T + 1)/T\}$ . Therefore the loss after round  $T$  is  $(2/3)\sqrt{1 + 1/T} \geq (2/3) + (2/9)T$ , for  $T \geq 3$ . Since the error after 3 rounds is still at least  $(2/3) + 0.1$  the Lemma holds for  $\varepsilon < 0.1$ .  $\blacksquare$

**Lemma 18** *Suppose  $u_0, u_1, \dots$ , are non-negative numbers satisfying*

$$u_t - u_{t+1} \geq c_0 u_t^{1+c_1},$$

*for some non-negative constants  $c_0, c_1$ . Then, for any  $t$ ,*

$$\frac{1}{u_t^{c_1}} - \frac{1}{u_0^{c_1}} \geq c_1 c_0 t.$$

THE RATE OF CONVERGENCE OF ADABoost

**Proof** By induction on  $t$ . The base case is an identity. Assume Lemma holds for  $t$ . Then,

$$\frac{1}{u_{t+1}^{c_1}} - \frac{1}{u_0^{c_1}} \geq \left( \frac{1}{u_{t+1}^{c_1}} - \frac{1}{u_t^{c_1}} \right) + \left( \frac{1}{u_t^{c_1}} - \frac{1}{u_0^{c_1}} \right) \geq \frac{1}{u_{t+1}^{c_1}} - \frac{1}{u_t^{c_1}} + c_0 t, \text{ (by induction).}$$

Thus it suffices to show

$$\frac{1}{u_{t+1}^{c_1}} - \frac{1}{u_t^{c_1}} \geq c_1 c_0 \iff \left( \frac{u_t}{u_{t+1}} \right)^{c_1} \geq 1 + c_1 c_0 u_t^{c_1} \iff \frac{1}{(1 - c_0 u_t^{c_1})^{c_1}} \geq 1 + c_1 c_0 u_t^{c_1}.$$

Since  $1 + c_1 c_0 u_t^{c_1} \leq (1 + c_0 u_t^{c_1})^{c_1}$ , and  $(1 + c_0 u_t^{c_1})(1 - c_0 u_t^{c_1}) < 1$ , the inequality holds. ■

MUKHERJEE RUDIN SCHAPIRE

# Online Learning: Beyond Regret

**Alexander Rakhlin**

*Department of Statistics  
University of Pennsylvania*

**Karthik Sridharan**

*TTI-Chicago*

**Ambuj Tewari**

*Department of Computer Science  
University of Texas at Austin*

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We study online learnability of a wide class of problems, extending the results of Rakhlin et al. (2010a) to general notions of performance measure well beyond external regret. Our framework simultaneously captures such well-known notions as internal and general  $\Phi$ -regret, learning with non-additive global cost functions, Blackwell’s approachability, calibration of forecasters, and more. We show that learnability in all these situations is due to control of the same three quantities: a martingale convergence term, a term describing the ability to perform well if future is known, and a generalization of sequential Rademacher complexity, studied in Rakhlin et al. (2010a). Since we directly study complexity of the problem instead of focusing on efficient algorithms, we are able to improve and extend many known results which have been previously derived via an algorithmic construction.

## 1. Introduction

In the companion paper Rakhlin et al. (2010a) (hereafter referred to as *RST*), we analyzed learnability in the **Online Learning Model** when the value of the game is defined through minimax *regret*. However, regret (also known as *external regret*) is not the only way to measure performance of an online learning procedure. In the present paper, we extend the results of *RST* to other performance measures, encompassing a wide spectrum of notions which appear in the literature. Our framework gives the same footing to external regret, internal and general  $\Phi$ -regret, learning with non-additive global cost functions, Blackwell’s approachability, calibration of forecasters, and more. We recover, extend, and improve some existing results, and (what is more important) show that they all follow from control of the same quantities. In particular, sequential Rademacher complexity, introduced in *RST*, plays a key role in our derivations.

A reflection on the past two decades of research in learning theory reveals (in our somewhat biased view) an interesting difference between Statistical Learning Theory and Online Learning. In the former, the focus has been primarily on understanding *complexity measures* rather than *algorithms*. There are good reasons for this: if a supervised problem with i.i.d. data is learnable, Empirical Risk Minimization is the algorithm that will perform

well if one disregards computational aspects. In contrast, Online Learning has been mainly centered around algorithms. Given an algorithm, a non-trivial bound serves as a certificate that the problem is learnable. This algorithm-focused approach has dominated research in Online Learning for several decades. Many important tools (such as optimization-based algorithms for online convex optimization) have emerged, yet the results lacked a unified approach for determining learnability.

With the tools developed in *RST*, the question of learnability can now be addressed in a variety of situations in a unified manner. In fact, *RST* presents a number of examples of provably learnable problems for which computationally feasible online learning methods have not yet been developed. In the present paper, we show that the scope of problems whose learnability and precise rates can be characterized is much larger than those defined in *RST* through external regret. Within this circle of problems are such well-known results as Blackwell’s approachability and calibration of forecasters. For instance, our complexity-based (rather than algorithm-based) approach yields a proof of Blackwell’s approachability in Banach spaces without ever mentioning an algorithm. Let us remark that Blackwell’s approachability has been a key tool for showing learnability (Cesa-Bianchi and Lugosi, 2006); as our results imply approachability, they can be utilized whenever Blackwell’s approachability has been successful. The results can also be used in situations where phrasing a problem as an approachability question is not necessarily natural. In Section 4.2, we discuss the relation of our results to approachability in greater detail. Our contributions can be broken down into three parts:

1. We formulate the online learning problem, with a performance measure (a form of *regret*), defined in terms of certain payoff transformations. While this formulation might appear unusual, we show that it is general enough to encompass many seemingly different frameworks, yet specific enough that we can provide generic upper bounds.
2. We develop upper and lower bounds on the value of the game under various natural assumptions. These tools allow us to deal with performance measures well beyond the standard additive notion of external regret. Such performance measures include smooth non-additive functions of payoffs, generalizing the “cumulative payoff” notion often considered in the literature. The abstract definition in terms of payoff transformations lets us consider rich classes of mappings whose complexity can be studied through random averages, covering numbers, and combinatorial parameters.
3. We apply our machinery to a number of well-known problems. Unfortunately, in this extended abstract we are not able to fit all the details. We refer the reader to Rakhlin et al. (2010b).
  - (a) For the usual notion of external regret, the results boil down to those of Rakhlin et al. (2010a).
  - (b) For the more general  $\Phi$ -regret (see e.g. Stoltz and Lugosi (2007); Gordon et al. (2008); Hazan and Kale (2007)), we recover and improve several known results. In particular, for convergence to  $\Phi$ -correlated equilibria, we improve upon the results of Stoltz and Lugosi (Stoltz and Lugosi, 2007).

- (c) We study the game of Blackwell's approachability (Blackwell, 1956) in (possibly infinite-dimensional) separable Banach spaces. Specifically, we show that variation of the worst-case martingale upper and lower-bounds (to within a constant) the rate of convergence to the set.
- (d) We also consider the game of calibrated forecasting. We improve upon the results of Mannor and Stoltz (Mannor and Stoltz, 2010) and prove (to the best of our knowledge) the first known  $O(\sqrt{T})$  rates for calibration with more than 2 outcomes. Our approach is markedly different from those found in the literature.
- (e) We use our framework to study games with global cost functions and as an example we extend the bounds recently obtained by Even-Dar et al. (2009).
- (f) We provide techniques for bounding notions of regret where algorithm's performance is measured against a time-varying comparator (see e.g. Herbster and Warmuth (1998); Bousquet and Warmuth (2002); Zinkevich (2003)).

The intent of this paper is to provide a framework and tools for studying problems that can be phrased as repeated games. However, unlike much of existing research in online learning, we are not solving the general problem by exhibiting an algorithm and studying its performance. Rather, we proceed by directly attacking the value of the game. Alas, the value is a complicated object, and the non-invitingly long sequence of infima and suprema can single-handedly extinguish any desire to study it. Our results attest to the power of *symmetrization*, which emerges as a key tool for studying the value of the game. In the literature, symmetrization has been used for i.i.d. data (Giné and Zinn, 1984). In *RST* (see also Abernethy et al. (2009)), it was shown that symmetrization can also be used in situations beyond the traditional setting. What is even more surprising, we are able to employ symmetrization ideas even when the objective function is not a summation of terms but rather a global function of many variables. We hope that these tools can have an impact not only on online learning but also on game theory.

We believe that there are many more examples falling under the present framework. We only chose a few to demonstrate how upper and lower bounds arise from the complexity of the problem. Along with an upper bound, a (computationally inefficient) algorithm can always be recovered from the minimax analysis. Finding efficient algorithms is often a difficult enterprise, and it is important to be able to understand the inherent complexity even before focusing on computation.

## 2. The Setting

At a very abstract level, the problem of online learning can be phrased as that of optimization of a given function  $\mathbf{R}_T(f_1, x_1, \dots, f_T, x_T)$  with coordinates being chosen *sequentially* by the player and the adversary. Of course, at this level of generality not much can be said. Hence, we make some minimal assumptions on the function  $\mathbf{R}_T$  which lead to meaningful guarantees on the online optimization process.<sup>1</sup> These assumptions are satisfied by a number of natural performance measures, as illustrated by the examples below.

---

1. The question of general conditions on the function under which such sequential minimization is possible was put forth by Peter Bartlett a few years ago in a coffee conversation. This paper paves way towards addressing this question.

Let  $\mathcal{F}$  and  $\mathcal{X}$  be the sets of moves of the learner (player) and the adversary, respectively. Generalizing the **Online Learning Model** considered in *RST*, we study the following  $T$ -round interaction between the learner and the adversary: On round  $t = 1, \dots, T$ , the learner chooses a mixed strategy  $q_t$  (distribution on  $\mathcal{F}$ ), the adversary picks  $x_t \in \mathcal{X}$ , the learner draws  $f_t \in \mathcal{F}$  from  $q_t$  and receives payoff (loss) signal  $\ell(f_t, x_t) \in \mathcal{H}$ .

We would like to specify that we are in the full information setting and that at the end of each round both the player and the adversary observe each other's moves  $f_t, x_t$ . The payoff space  $\mathcal{H}$  is a (not necessarily convex) subset of a separable Banach space  $\mathcal{B}$ . Both the player and the adversary can be randomized and adaptive. The goal of the learner is to minimize the following general form

$$\mathbf{R}_T = \mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) - \inf_{\phi \in \Phi_T} \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T)) \quad (1)$$

of performance measure, where

- (1) The function  $\ell : \mathcal{F} \times \mathcal{X} \mapsto \mathcal{H}$  is an  $\mathcal{H}$ -valued payoff (or loss) function.
- (2) The function  $\mathbf{B} : \mathcal{H}^T \mapsto \mathbb{R}$  is a (not necessarily additive or convex) form of cumulative payoff.
- (3) The set  $\Phi_T$  consists of sequences  $\phi = (\phi_1, \dots, \phi_T)$  of measurable payoff transformation mappings  $\phi_t : \mathcal{H}^{\mathcal{F} \times \mathcal{X}} \mapsto \mathcal{H}^{\mathcal{F} \times \mathcal{X}}$  that transform the payoff function  $\ell$  into a payoff function  $\ell_{\phi_t}$ .

The goal of the adversary is to maximize the same quantity (1), making it a zero-sum game.

This paper is concerned with learnability and with identifying *complexity measures* that govern learnability. But complexity of what should we focus on? After all, the general online learning problem is defined by the choice of five components:  $\mathbf{B}, \ell, \mathcal{F}, \mathcal{X}$ , and  $\Phi_T$ . In *RST*, the choice was easy: it should be the complexity of the function class  $\mathcal{F}$  that plays the key role. That was natural because the payoff was written as  $\ell(f, x) = f(x)$ , which suggested that the function class  $\mathcal{F}$  is the object of study. The present formulation, however, is much more general. When this work commenced, it seemed likely that complexity of the problem will be some interaction between the complexity of  $\Phi_T$  and complexity of  $\mathcal{F}$ . As we show below, one may just focus on the complexity of  $\Phi_T$ , while  $\mathcal{F}$  and  $\mathcal{X}$  are now on the same footing. For instance, even if it might seem unusual at first, we will introduce a notion of a cover of the set of sequences of payoff transformations  $\Phi_T$ . In summary, while all five components  $\mathbf{B}, \ell, \mathcal{F}, \mathcal{X}$ , and  $\Phi_T$  play a role in determining learnability, we will mainly refer to the complexity of the payoff mapping  $\ell$  and the payoff transformation  $\Phi_T$  without an explicit reference to  $\mathcal{F}, \mathcal{X}$ , and  $\mathbf{B}$ . We emphasize that most flexibility comes from the payoff mapping  $\ell$  and from the transformations  $\Phi_T$  of the payoffs.

Important classes of payoff transformation mappings are those that transform the payoff function  $\ell$  by acting only on the first argument of  $\ell$ , i.e. only modifying the player's action. Formally, a class of sequences of payoff transformations  $\Phi_T$  is said to be a *departure mapping class* if there exists a class  $\Phi'_T$  of sequences  $\phi' = (\phi'_1, \dots, \phi'_T)$  with  $\phi'_i : \mathcal{F} \mapsto \mathcal{F}$  such that for each  $\phi \in \Phi_T$  there exists a  $\phi' \in \Phi'_T$  with  $\ell_{\phi_t}(f, x) := \ell(\phi'_t(f), x)$  that for all  $t \in [T]$ ,  $f \in \mathcal{F}$  and  $x \in \mathcal{X}$ . We shall slightly abuse notation and use  $\Phi_T$  to represent both the class of payoff transformation and the class of departure mappings from  $\mathcal{F}$  to itself. Another class of interest are payoff transformations that do not vary with time. We say

that  $\Phi_T$  is *time-invariant* if all sequences of payoff transformation are constant in time:  $\Phi_T = \{(\phi, \dots, \phi) : \phi \in \Phi\}$ , where  $\Phi$  is a “basis” class of mappings  $\mathcal{H}^{\mathcal{F} \times \mathcal{X}} \mapsto \mathcal{H}^{\mathcal{F} \times \mathcal{X}}$ .

In the following, we assume that  $\mathcal{F}$  and  $\mathcal{X}$  are subsets of a separable metric space. Let  $\mathcal{Q}$  and  $\mathcal{P}$  be the sets of probability distributions on  $\mathcal{F}$  and  $\mathcal{X}$ , respectively. Assume that  $\mathcal{Q}$  and  $\mathcal{P}$  are weakly compact. From the outset, we assume that the adversary is non-oblivious (that is, adaptive). Formally, define a learner’s strategy  $\pi$  as a sequence of mappings  $\pi_t : (\mathcal{P} \times \mathcal{F} \times \mathcal{X})^{t-1} \mapsto \mathcal{Q}$  for each  $t \in [T]$ . The form (1) of the performance measure gives rise to the value of the game:

$$\begin{aligned} \mathcal{V}_T(\ell, \Phi_T) = & \inf_{q_1} \sup_{x_1} \mathbb{E}_{f_1 \sim q_1} \dots \inf_{q_T} \sup_{x_T} \mathbb{E}_{f_T \sim q_T} \sup_{\phi \in \Phi_T} \{ \mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) \\ & - \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T)) \} \end{aligned} \quad (2)$$

where  $q_t$  and  $x_t$  range over  $\mathcal{Q}$  and  $\mathcal{X}$ , respectively. With this definition of a value, the (deterministic) strategy of the adversary is a sequence of mappings  $(\mathcal{Q} \times \mathcal{F} \times \mathcal{X})^{t-1} \times \mathcal{Q} \mapsto \mathcal{X}$  for each  $t \in [T]$ . The problem is said to be *online learnable* if  $\limsup_{T \rightarrow \infty} \mathcal{V}_T(\ell, \Phi_T) = 0$ .

The value of the game is defined as an *expected* performance measure. As such, it yields “in probability” statements. While beyond the scope of this paper, we can also define the value of the game using a *high probability* performance measure, leading to “almost sure” convergence (Rakhlin et al., 2010b).

## 2.1. Examples

A reader might wonder why we have defined the game in terms of abstract payoff transformation mappings. It turns out that with this definition, various seemingly different frameworks become nothing but special cases, as illustrated by the following examples.

**Example 1 (External Regret Game, Section 4.1.1)** Let  $\mathcal{H} = \mathbb{R}$ , let  $\mathbf{B}(z_1, \dots, z_T) = \frac{1}{T} \sum_{t=1}^T z_t$ , and

$$\Phi_T = \{(\phi_f, \dots, \phi_f) : f \in \mathcal{F} \text{ and } \phi_f : \mathcal{F} \mapsto \mathcal{F} \text{ is a constant mapping } \phi_f(g) = f \forall g \in \mathcal{F}\}.$$

It is easy to see that (1) becomes external regret:

$$\mathbf{R}_T = \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) - \inf_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^T \ell(f, x_t).$$

**Example 2 ( $\Phi$ -Regret, Section 4.1)** Let  $\mathcal{H} = \mathbb{R}$ , let  $\mathbf{B}(z_1, \dots, z_T) = \frac{1}{T} \sum_{t=1}^T z_t$ , and  $\Phi_T = \{(\phi, \dots, \phi) : \phi \in \Phi\}$  for a fixed family  $\Phi$  of  $\mathcal{F} \mapsto \mathcal{F}$  mappings. Performance measure in (1) becomes

$$\mathbf{R}_T = \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) - \inf_{\phi \in \Phi} \frac{1}{T} \sum_{t=1}^T \ell(\phi(f_t), x_t). \quad (3)$$

This example covers a variety of notions such as external, internal, and swap regrets.

**Example 3 (Blackwell Approachability, Section 4.2)** Let  $\mathcal{H}$  a subset of a Banach space  $\mathcal{B}$ ,  $S \subset \mathcal{B}$  be a closed convex set, and  $\mathbf{B}(z_1, \dots, z_T) = \inf_{c \in S} \left\| \frac{1}{T} \sum_{t=1}^T z_t - c \right\|$ . The set  $\Phi_T$  contains sequences  $(\phi_1, \dots, \phi_T)$  such that  $\ell_{\phi_t}(f, x) = c_t \in S$  for all  $f \in \mathcal{F}$ ,  $x \in \mathcal{X}$ , and  $1 \leq t \leq T$ . Eq. (1) becomes the distance to the set  $S$ :

$$\mathbf{R}_T = \inf_{c \in S} \left\| \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) - c \right\| \quad (4)$$

**Example 4 (Calibration of Forecasters, Section 4.3)** Let  $\mathcal{H} = \mathbb{R}^k$ ,  $\mathcal{F}$  the probability simplex in  $\mathbb{R}^k$ , and  $\mathcal{X}$  the vertices of  $\mathcal{F}$ . Define  $\ell(f, x) = 0$ . Further,  $\mathbf{B}(z_1, \dots, z_T) = - \left\| \frac{1}{T} \sum_{t=1}^T z_t \right\|$  for some norm  $\|\cdot\|$  on  $\mathbb{R}^k$ , and  $\Phi_T = \{(\phi_{p,\lambda}, \dots, \phi_{p,\lambda}) : p \in \Delta(k), \lambda > 0\}$  contains time-invariant mappings defined by  $\ell_{\phi_{p,\lambda}}(f, x) = \mathbf{1}\{\|f - p\| \leq \lambda\} \cdot (f - x)$ . Performance measure in (1) then becomes

$$\mathbf{R}_T = \sup_{\lambda > 0} \sup_{p \in \Delta(k)} \left\| \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{\|f_t - p\| \leq \lambda\} \cdot (f_t - x_t) \right\|. \quad (5)$$

**Example 5 (Global Cost Online Learning Game, Section 4.4)** (see also the original paper Even-Dar et al. (2009)) Let  $\mathcal{H} = \mathbb{R}^k$ ,  $\mathcal{X} = [0, 1]^k$ ,  $\mathcal{F} = \Delta(k)$ ,  $\ell(f, x) = f \odot x = (f^1 \cdot x^1, \dots, f^k \cdot x^k)$ . Let  $\mathbf{B}(z_1, \dots, z_T) = \left\| \frac{1}{T} \sum_{t=1}^T z_t \right\|$  and

$$\Phi_T = \{(\phi_f, \dots, \phi_f) : f \in \mathcal{F} \text{ and } \phi_f : \mathcal{F} \mapsto \mathcal{F} \text{ is a constant mapping } \phi_f(g) = f \forall g \in \mathcal{F}\}.$$

Then

$$\mathbf{R}_T = \left\| \frac{1}{T} \sum_{t=1}^T f_t \odot x_t \right\| - \inf_{f \in \mathcal{F}} \left\| \frac{1}{T} \sum_{t=1}^T f \odot x_t \right\|. \quad (6)$$

## 2.2. Notation

We let  $\mathbb{E}_{x \sim p}$  denote expectation w.r.t. a random variable  $x$  with a distribution  $p$ . For random variables  $x_1, \dots, x_T$  with distributions  $p_1, \dots, p_T$ , we will use the shorthand  $\mathbb{E}_{x_{1:T} \sim p_{1:T}}$  to denote expectation w.r.t. all these variables. Let  $q$  and  $p$  be distributions on  $\mathcal{F}$  and  $\mathcal{X}$ , respectively. We define a shorthand  $\ell(q, p) = \mathbb{E}_{f \sim q, x \sim p} \ell(f, x)$  and  $\ell_\phi(q, p) = \mathbb{E}_{f \sim q, x \sim p} \ell_\phi(f, x)$ . The Dirac delta distribution is denoted by  $\delta_x$ . A Rademacher random variable is symmetric  $\{\pm 1\}$ . The notation  $x_{a:b}$  denotes the sequence  $x_a, \dots, x_b$ . The indicator of an event  $A$  is denoted by  $\mathbf{1}\{A\}$ . The set  $\{1, \dots, T\}$  is denoted by  $[T]$ , while the  $k$ -dimensional probability simplex is denoted by  $\Delta(k)$ . The set of all functions from  $\mathcal{X}$  to  $\mathcal{Y}$  is denoted by  $\mathcal{Y}^\mathcal{X}$ , and the  $t$ -fold product is denoted by  $\mathcal{X}^t$ . Whenever a supremum (infimum) is written as  $\sup_a$  without  $a$  being quantified, it is assumed that  $a$  ranges over the set of all possible values which will be understood from the context. For a separable Banach space  $\mathcal{B}$  equipped with a norm  $\|\cdot\|$ , let  $B_{\|\cdot\|}$  be the unit ball. Let  $\mathcal{B}^*$  denote the dual space and  $B_{\|\cdot\|_*}$  the corresponding dual ball. For  $a \in \mathcal{B}^*$ ,  $\|a\|_* = \sup_{b \in B_{\|\cdot\|}} |\langle a, b \rangle|$ . For  $b \in \mathcal{B}$ , we write  $\langle a, b \rangle = a(b)$  for the continuous linear functional  $a \in \mathcal{B}^*$  on  $\mathcal{B}$ . Let  $\phi_{\text{id}}$  be identity payoff transformation  $\ell_{\phi_{\text{id}}}(f, x) = \ell(f, x)$  for all  $f \in \mathcal{F}$ ,  $x \in \mathcal{X}$ . The singleton set containing the time-invariant

sequence of identity transformations is denoted by  $\mathcal{I} = \{(\phi_{\text{id}}, \dots, \phi_{\text{id}})\}$ . Following *RST*, we define binary trees as follows. Given some set  $\mathcal{Z}$ , a  $\mathcal{Z}$ -valued tree of depth  $T$  is a sequence  $(\mathbf{z}_1, \dots, \mathbf{z}_T)$  of  $T$  mappings  $\mathbf{z}_i : \{\pm 1\}^{i-1} \mapsto \mathcal{Z}$ . The root of the tree  $\mathbf{z}$  is the constant function  $\mathbf{z}_1 \in \mathcal{Z}$ . Unless specified otherwise,  $\epsilon = (\epsilon_1, \dots, \epsilon_T) \in \{\pm 1\}^T$  will define a path. Slightly abusing the notation, we will write  $\mathbf{z}_t(\epsilon)$  instead of  $\mathbf{z}_t(\epsilon_{1:t-1})$ .

### 3. General Upper Bounds

This section is devoted to upper bounds on the value of the game. We start by introducing the Triplex Inequality, which requires no assumptions beyond those described in Section 2. Under the additional weak assumption of subadditivity of  $\mathbf{B}$ , we can perform symmetrization and further upper bound two of the three terms in Triplex Inequality by a non-additive version of sequential Rademacher complexity. As we progress through the section, we make additional assumptions and specialize and refine the upper bounds. The following definition generalizes the notion of sequential Rademacher complexity, introduced in *RST*, to “global” functions  $\mathbf{B}$  of the payoff sequence.

**Definition 1** *The sequential complexity with respect to the payoff function  $\ell$  and payoff transformation mappings  $\Phi_T$  is defined as*

$$\mathfrak{R}_T(\ell, \Phi_T, \mathbf{B}) = \sup_{\mathbf{f}, \mathbf{x}} \mathbb{E}_{\epsilon_{1:T}} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\epsilon_1 \ell_{\phi_1}(\mathbf{f}_1(\epsilon), \mathbf{x}_1(\epsilon)), \dots, \epsilon_T \ell_{\phi_T}(\mathbf{f}_T(\epsilon), \mathbf{x}_T(\epsilon))\right)$$

where the outer supremum is taken over all  $(\mathcal{F} \times \mathcal{X})$ -valued trees of depth  $T$  and  $\epsilon = (\epsilon_1, \dots, \epsilon_T)$  is a sequence of i.i.d. Rademacher random variables.

Whenever  $\mathbf{B}$  is clear from the context, we will omit it from the notation:  $\mathfrak{R}_T(\ell, \Phi_T)$ . If  $\Phi_T$  is a set of sequences of time-invariant transformations obtained from the base class  $\Phi$ , we will simply write  $\mathfrak{R}_T(\ell, \Phi)$ . Let us remark that the moves of the player and the adversary appear “on the same footing” in  $\mathbf{R}_T$  and in the above definition of sequential complexity. The “asymmetry” of sequential Rademacher complexity as studied in *RST* (where the supremum is taken over the *player’s* best choice) arises precisely from the asymmetry of the notion of external regret, which, in turn, is due to  $\Phi_T$  acting on the player choice only. In Section 4.1.1, we show that the notion studied in *RST* is indeed recovered for the case of external regret. An equivalent way to write sequential complexity is

$$\mathfrak{R}_T(\ell, \Phi_T, \mathbf{B}) = \sup_{f_1, x_1} \mathbb{E}_{\epsilon_1} \sup_{f_2, x_2} \mathbb{E}_{\epsilon_2} \dots \sup_{f_T, x_T} \mathbb{E}_{\epsilon_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\epsilon_1 \ell_{\phi_1}(f_1, x_1), \dots, \epsilon_T \ell_{\phi_T}(f_T, x_T)\right) \quad (7)$$

where the supremum on  $t$ -th step is over  $f_t \in \mathcal{F}$ ,  $x_t \in \mathcal{X}$ .

#### 3.1. Triplex Inequality

The following theorem is the starting point for all further analysis. Because of its importance, we shall call it the *Triplex Inequality*. The three terms in the upper bound of the theorem are the three key players in the process of online learning: martingale convergence, the ability to perform well if the future is known, and complexity of the class in terms of sequential complexity.

**Theorem 2 (Triplex Inequality)** *The following 3-term upper bound on the value of the game holds:*

$$\begin{aligned}
 & \mathcal{V}_T(\ell, \Phi_T) \\
 & \leq \sup_{p_1, q_1} \mathbb{E}_{x_1, f_1} \dots \sup_{p_T, q_T} \mathbb{E}_{x_T, f_T} \left\{ \mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) - \mathbb{E}_{x'_{1:T}, f'_{1:T}} \mathbf{B}(\ell(f'_1, x'_1), \dots, \ell(f'_T, x'_T)) \right\} \\
 & + \sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \sup_{\phi \in \Phi_T} \mathbb{E}_{x_{1:T}, f_{1:T}} \left\{ \mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) - \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T)) \right\} \\
 & + \sup_{p_1, q_1} \mathbb{E}_{x_1, f_1} \dots \sup_{p_T, q_T} \mathbb{E}_{x_T, f_T} \\
 & \quad \sup_{\phi \in \Phi_T} \left\{ \mathbb{E}_{x'_{1:T}, f'_{1:T}} \mathbf{B}(\ell_{\phi_1}(f'_1, x'_1), \dots, \ell_{\phi_T}(f'_T, x'_T)) - \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T)) \right\}
 \end{aligned} \tag{8}$$

In the statement of the theorem, the random variables  $f_t, f'_t$  have distribution  $q_t$  while  $x_t, x'_t$  have distribution  $p_t$ . We remark that convexity of  $\mathbf{B}$  is *not required* for the Triplex Inequality to hold. Under a subadditivity condition, the following result gives upper bounds on the first and third terms.

**Theorem 3** *If  $\mathbf{B}$  is subadditive, then the last term in the Triplex Inequality is upper bounded by twice the sequential complexity,  $2\mathfrak{R}_T(\ell, \Phi_T, \mathbf{B})$ , and the first term is bounded by  $2\mathfrak{R}_T(\ell, \mathcal{I}, \mathbf{B})$  where  $\mathcal{I}$  is the singleton set consisting of the identity mapping. Similarly, if  $-\mathbf{B}$  is subadditive, then the last term is upper bounded by  $2\mathfrak{R}_T(\ell, \Phi_T, -\mathbf{B})$  and the first term is bounded by  $2\mathfrak{R}_T(\ell, \mathcal{I}, -\mathbf{B})$ .*

**Discussion of Theorem 2 and Theorem 3** We note that the first and the third terms are similar in their form. In fact, the first term can be equivalently written in a form similar to the third term, with only one difference that  $\phi$  belongs to a singleton set  $\mathcal{I}$  containing the identity mapping. If  $\mathcal{I} \subseteq \Phi_T$ , then, trivially,  $\mathfrak{R}_T(\ell, \mathcal{I}, \mathbf{B}) \leq \mathfrak{R}_T(\ell, \Phi_T, \mathbf{B})$  and, therefore, an upper bound on the third term yields an upper bound on the first. However, in some situations  $\Phi_T$  is “simpler” or incomparable to  $\mathcal{I}$  and, hence, the first and the third term in the Triplex Inequality are distinct.

What exactly is achieved by Theorem 3? Let us compare the third term in the Triplex Inequality to its sequential complexity upper bound given by Eq. (7). Both quantities involve interleaved suprema and expected values. However, in the former, the suprema are over the choice of distributions  $p_t, q_t$  and the expected values are draws of  $x_t, f_t$  from these mixed strategies. In contrast, sequential complexity, as written in Eq. (7), contains suprema over the choices  $x_t, f_t$  followed by a random draw of the next sign  $\epsilon_t$ . Crucially, it is easier to work with the sequential complexity as opposed to the third term in the Triplex Inequality since in the former the only randomness comes from the random signs. In mathematical terms, the  $\sigma$ -algebra is generated by  $\{\epsilon_t\}$  rather than a complicated stochastic process arising from the Triplex Inequality. This is one of the key observations of the paper.

Depending on a particular problem, some of the terms in the Triplex Inequality might be easier to control than others. However, it is often the case that the first term is the easiest, as it naturally leads to the question of martingale convergence. The second term is typically bounded by providing a specific response strategy for the player if the mixed strategy of the adversary is known. This response strategy is similar to the so-called Blackwell’s condition for approachability (see Section 4.2 for further comparison). The third term is arguably the

most difficult as it captures complexity of the set of payoff transformations  $\Phi_T$ . Under the subadditivity assumption on  $\mathbf{B}$ , Theorem 3 upper bounds the first and third terms by the sequential complexity.

The following observation gives us a simple condition under which we can replace  $\mathbf{B}$  with some other  $\mathbf{B}'$ , and we shall find it useful in scenarios when it is difficult to directly deal with  $\mathbf{B}$ . If  $\mathbf{B} : \mathcal{H}^T \mapsto \mathbb{R}$  and  $\mathbf{B}' : \mathcal{H}^T \mapsto \mathbb{R}$  are such that  $\forall z_1, \dots, z_T \in \mathcal{H}$ ,  $\mathbf{B}(z_1, \dots, z_T) \leq \mathbf{B}'(z_1, \dots, z_T)$  then we have that for any class of transformations  $\Phi_T$ ,  $\mathfrak{R}_T(\ell, \Phi_T, \mathbf{B}) \leq \mathfrak{R}_T(\ell, \Phi_T, \mathbf{B}')$ .

This completes our discussion of the main theorems. We now turn to the question of upper bounding the terms in the Triplex Inequality. To this end, we need to define the notion of a smooth function. A function  $g : \mathcal{H} \mapsto \mathbb{R}$  is said to be  $(\sigma, p)$ -uniformly smooth for some  $p \in (1, 2]$  and  $\sigma \geq 0$  if for all  $z, z' \in \mathcal{H}$  we have,

$$g(z) \leq g(z') + \langle \nabla g(z'), z - z' \rangle + \frac{\sigma}{p} \|z - z'\|^p.$$

We say that  $g$  is *uniformly smooth* if there exist finite  $\sigma$  and  $p$  such that  $g$  is  $(\sigma, p)$ -uniformly smooth. We say that a norm  $\|\cdot\|$  is  $(\sigma, p)$ -smooth if  $\|\cdot\|^p/p$  is a  $(\sigma, p)$ -smooth function.

A function  $\mathbf{B}$  which is smooth in its arguments can be “sequentially linearized”, with additional second-order terms appearing as norms of the increments. Informally, the smoothness assumption provides a link from a “global” function  $\mathbf{B}$  of coordinates to a sum of its parts. From the point of view of online learning, this is very promising, as it appears to be difficult to sequentially optimize a “global” function of many decisions. Due to limited space in this extended abstract, we will not present the most general bounds based solely on smoothness of  $\mathbf{B}$  (we refer the reader to Rakhlin et al. (2010b) for these results). However, we will state bounds for a smooth function of the average of coordinates.

### 3.2. When $\mathbf{B}$ is a Function of the Average

For the rest of this sub-section we assume that  $\mathbf{B} = G\left(\frac{1}{T} \sum_{t=1}^T z_t\right)$ , where (some power of)  $G$  is an appropriately smooth function on the convex set  $\text{conv}(\mathcal{H})$ . This form of  $\mathbf{B}$  occurs naturally in many games including Blackwell’s approachability and calibration. Among the most basic smooth functions are powers of norms. For the  $\ell_q$  norms, the following smoothness results are known. For any  $q \in (1, 2]$ ,  $G(z) = \|z\|_q^q$  is  $(q, q)$ -uniformly smooth and for any  $q \in [2, \infty)$  the function  $G(z) = \|z\|_q^2$  is  $(2(q-1), 2)$ -uniformly smooth. The  $\ell_\infty$  cannot be made smooth by raising it to any finite power  $s$ . However, for any  $z \in \mathcal{H}$  and any  $q' \in (1, \infty)$ ,  $\|z\|_\infty \leq \|z\|_{q'}$ . Hence as discussed above,  $\mathfrak{R}_T(\ell, \Phi_T, \mathbf{B}) \leq \mathfrak{R}_T(\ell, \Phi_T, \mathbf{B}')$  where  $\mathbf{B}'(z_1, \dots, z_T) = \left\| \frac{1}{T} \sum_{t=1}^T z_t \right\|_{q'}$ . By choosing  $q'$  appropriately and using the smoothness of the  $\ell_{q'}$  norm we can provide upper bounds for the value of the game. Similarly to  $\ell_\infty$ , no finite power of the  $\ell_1$  norm is smooth. However if  $\mathcal{H} \subseteq \mathbb{R}^d$ , we can upper bound the  $\ell_1$  norm by, say,  $\ell_2$  norm multiplied by a factor  $\sqrt{d}$ . Smoothness of this latter norm can then be used. This is indeed the approach that is employed for proving rates for calibration.

The following result shows that if  $G$  is 1-Lipschitz and  $G^2$  is 2-smooth, we obtain a  $O(1/\sqrt{T})$  convergence rate whenever  $\Phi_T$  is a finite set. We refer to Rakhlin et al. (2010b) for the case when  $G$  is not Lipschitz, as well as for the case of a  $(\gamma, p)$ -smooth function  $G$  for  $1 < p \leq 2$ .

**Lemma 4** Let  $\Phi_T$  be a finite set of payoff transformations. Let

$$\mathbf{B}(z_1, \dots, z_T) = G\left(\frac{1}{T} \sum_{t=1}^T z_t\right)$$

where  $G$  is 1-Lipschitz with respect to a norm  $\|\cdot\|$  and  $G(0) = 0$ . Suppose that  $G^2$  is  $(\gamma, 2)$ -smooth function on the convex set  $\text{conv}(\mathcal{H})$ . Further, suppose that for any  $x \in \mathcal{X}$ ,  $f \in \mathcal{F}$ ,  $\phi \in \Phi_T$  and  $t \in [T]$ , it is true that  $\|\ell_{\phi_t}(f, x)\| \leq \eta$ . Then it holds that

$$\mathfrak{R}_T(\ell, \Phi_T) \leq \sqrt{\frac{\gamma \eta^2 \log(2|\Phi_T|)}{T}}.$$

Having a bound on the complexity of a finite set of payoff transformations, we seek to extend the results to infinite sets. A natural approach is to pass to a finite cover of the set at an expense of losing an amount proportional to the resolution of the cover. The following definition can be seen as a generalization of the corresponding notion introduced in *RST*. We remark that the object, for which we would like to provide a cover, is the set  $\Phi_T$ . Whenever payoff transformations are simply constant time-invariant departure mappings, complexity of  $\Phi_T$  identical to that of  $\mathcal{F}$ , yielding the online cover of class  $\mathcal{F}$ . In general, however, the set of payoff transformations can be much more complex than (or not even comparable to)  $\mathcal{F}$ .

**Definition 5** A set  $V$  of  $\mathcal{H}$ -valued trees of depth  $T$  is an  $\alpha$ -cover (with respect to  $\ell_p$ -norm) of  $\Phi_T$  on an  $(\mathcal{F} \times \mathcal{X})$ -valued tree  $(\mathbf{f}, \mathbf{x})$  of depth  $T$  if

$$\forall \phi \in \Phi_T, \forall \epsilon \in \{\pm 1\}^T \exists \mathbf{v} \in V \text{ s.t. } \frac{1}{T} \sum_{t=1}^T \|\mathbf{v}_t(\epsilon) - \ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon))\|^p \leq \alpha^p. \quad (9)$$

The covering number of the set of payoff transformations  $\Phi_T$  on a given tree  $(\mathbf{f}, \mathbf{x})$  is defined as

$$\mathcal{N}_p(\alpha, \Phi_T, (\mathbf{f}, \mathbf{x})) = \min\{|V| : V \text{ is an } \alpha\text{-cover w.r.t. } \ell_p\text{-norm of } \Phi_T \text{ on } (\mathbf{f}, \mathbf{x}) \text{ tree}\}.$$

Further define  $\mathcal{N}_p(\alpha, \Phi_T, T) = \sup_{(\mathbf{f}, \mathbf{x})} \mathcal{N}_p(\alpha, \Phi_T, (\mathbf{f}, \mathbf{x}))$ , the maximal  $\ell_p$  covering number of  $\Phi_T$ .

In sections that follow, we specialize this definition to fit particular assumptions on  $\Phi_T$ . The next theorem shows that sequential complexity can be bounded above in terms of the covering number, integrated over all the scales. This is a generalization of the analogous result in *RST*.

**Theorem 6** Assume that  $\mathbf{B}(z_1, \dots, z_T) = G\left(\frac{1}{T} \sum_{t=1}^T z_t\right)$  where  $G$  is 1-Lipschitz with respect to a norm  $\|\cdot\|$  and  $G(0) = 0$ . Suppose that  $G^2$  is  $(\gamma, 2)$ -smooth function on the convex set  $\text{conv}(\mathcal{H})$ . Further, suppose that for any  $x \in \mathcal{X}$ ,  $f \in \mathcal{F}$ ,  $\phi \in \Phi_T$  and  $t \in [T]$ , it is true that  $\|\ell_{\phi_t}(f, x)\| \leq \eta$ . Then it holds that

$$\mathfrak{R}_T(\ell, \Phi_T) \leq 4 \inf_{\alpha > 0} \left\{ \alpha + 3 \sqrt{\frac{\gamma}{T} \int_{\alpha}^{\eta} \sqrt{\log \mathcal{N}_{\infty}(\beta, \Phi_T, T)} d\beta} \right\}.$$

### 3.3. General Bounds Under Linearity Assumptions on $\mathbf{B}$

The general results of the previous section can be restated in simpler terms if more assumptions are made. In particular, some of the terms in the Triplex Inequality can be dropped as soon as  $\mathbf{B}$  is linear. While some of the results below can be repeated for a more general form of  $\mathbf{B}$ , for simplicity we assume that  $\mathbf{B}$  is an average of its arguments and that  $\mathcal{H} \subseteq \mathbb{R}$ :  $\mathbf{B}(z_1, \dots, z_T) = \frac{1}{T} \sum_{t=1}^T z_t$ .

**Corollary 7** *Under the assumption on  $\mathbf{B}$  made above, the following statements hold:*

(a) *The first term in the Triplex Inequality is zero.*

(b) *If  $\Phi_T$  is a class of departure mappings, then the second term in the Triplex Inequality is non-positive. Hence,*

$$\mathcal{V}_T(\ell, \Phi_T) \leq 2\mathfrak{R}_T(\ell, \Phi_T).$$

(c) *For  $\mathcal{H} \subseteq [-1, 1]$ , and assuming  $|\ell(f, x)| \leq \eta$  for any  $x \in \mathcal{X}$  and  $f \in \mathcal{F}$ ,*

$$\mathfrak{R}_T(\ell, \Phi_T) \leq 4 \inf_{\alpha \geq 0} \left\{ \alpha + 3\sqrt{2} \int_{\alpha}^{\eta} \sqrt{\frac{\log \mathcal{N}_{\infty}(\delta, \Phi_T, T)}{T}} d\delta \right\}.$$

Experts in the area will notice the use of  $\ell_{\infty}$  (as opposed to  $\ell_2$  in the classical Dudley integral bound) covering numbers in the two results above. This can certainly be done (Rakhlin et al., 2010b) for Corollary 7 and most probably even for the more general Theorem 6. However, in applications, one seldom loses more than a mild logarithmic factor (in  $T$ ) due to the use of  $\ell_{\infty}$  covering numbers.

When  $\mathbf{B}$  is the average of its coordinates, the sequential complexity takes on a familiar form:

$$\mathfrak{R}_T(\ell, \Phi_T) = \sup_{\mathbf{f}, \mathbf{x}} \mathbb{E}_{\epsilon_{1:T}} \left\{ \sup_{\phi \in \Phi_T} \frac{1}{T} \sum_{t=1}^T \epsilon_t \ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) \right\}.$$

Further, for  $\mathcal{H} \subseteq \mathbb{R}$ , Eq. (9) in definition of the cover becomes

$$\forall \phi \in \Phi_T, \forall \epsilon \in \{\pm 1\}^T \exists \mathbf{v} \in V \text{ s.t. } \frac{1}{T} \sum_{t=1}^T |\mathbf{v}_t(\epsilon) - \ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon))|^p \leq \alpha^p$$

where  $V$  is now a set of  $\mathbb{R}$ -valued trees. A further simplification of various notions is obtained for time-invariant payoff transformations. Moreover, for time-invariant payoff transformations we can define combinatorial parameters, generalizing the Littlestone (Littlestone, 1988; Ben-David et al., 2009) and (online) fat-shattering dimensions (Rakhlin et al., 2010a). This is the subject of the next section.

#### 3.3.1. COMBINATORIAL PARAMETERS FOR TIME-INVARIANT PAYOFF TRANSFORMATIONS

Assume  $\mathcal{H} \subseteq \mathbb{R}$ . Consider time-invariant payoff transformations generated from some base class of payoff transformations  $\Phi$ . That is,  $\Phi_T = \{(\phi, \dots, \phi) : \phi \in \Phi\}$ . We have the following definition of a generalized shattering dimension.

**Definition 8** Let  $\mathcal{H} = \{\pm 1\}$ . An  $(\mathcal{F} \times \mathcal{X})$ -valued tree  $(\mathbf{f}, \mathbf{x})$  of depth  $d$  is shattered<sup>2</sup> by a payoff transformation class  $\Phi$  if for all  $\epsilon \in \{\pm 1\}^d$ , there exists  $\phi \in \Phi$  such that  $\ell_\phi(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) = \epsilon_t$  for all  $t \in [d]$ . The shattering dimension  $\text{Sdim}(\Phi)$  is the largest  $d$  such that  $\Phi$  shatters an  $(\mathcal{F} \times \mathcal{X})$ -valued tree of depth  $d$ .

We can also define the scale-sensitive version of the shattering dimension, generalizing the fat-shattering dimension of *RST*.

**Definition 9** An  $(\mathcal{F} \times \mathcal{X})$ -valued tree  $(\mathbf{f}, \mathbf{x})$  of depth  $d$  is  $\alpha$ -shattered by a payoff transformation class  $\Phi$ , if there exists an  $\mathbb{R}$ -valued tree  $\mathbf{s}$  of depth  $d$  such that

$$\forall \epsilon \in \{\pm 1\}^d, \exists \phi \in \Phi \quad s.t. \quad \forall t \in [d], \epsilon_t (\ell_\phi(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) \geq \alpha/2$$

The tree  $\mathbf{s}$  is called the witness to shattering. The fat-shattering dimension  $\text{fat}_\alpha(\Phi)$  at scale  $\alpha$  is the largest  $d$  such that  $\Phi$   $\alpha$ -shatters an  $(\mathcal{F} \times \mathcal{X})$ -valued tree of depth  $d$ .

Slightly abusing notation, we write  $\mathcal{N}_p(\alpha, \Phi, (\mathbf{f}, \mathbf{x}))$  instead of  $\mathcal{N}_p(\alpha, \Phi_T, (\mathbf{f}, \mathbf{x}))$  whenever  $\Phi_T$  consists of sequences of time-invariant payoff transformations with a base class  $\Phi$ .

The combinatorial parameters are useful if they can be shown to control problem complexity through, for instance, covering numbers. We state the following two results without proofs, as the arguments are identical to the ones given in the companion paper *RST*. To be precise, the  $(\mathbf{f}, \mathbf{x})$  tree here plays the role of the  $\mathbf{x}$  tree in *RST*,  $\ell_\phi$  for  $\phi \in \Phi$  plays the role of  $f \in \mathcal{F}$  in *RST*.

**Theorem 10** Let  $\mathcal{H} \subseteq \{0, \dots, k\}$  and  $\text{fat}_2(\Phi) = d_2$ ,  $\text{fat}_1(\Phi) = d_1$ . Then

$$\mathcal{N}_\infty(1/2, \Phi, T) \leq \sum_{i=0}^{d_2} \binom{T}{i} k^i \leq (ekT)^{d_2}, \quad \mathcal{N}(0, \Phi, T) \leq \sum_{i=0}^{d_1} \binom{T}{i} k^i \leq (ekT)^{d_1}.$$

In particular, the result holds for binary-valued payoffs ( $k = 1$ ), in which case  $\text{fat}_1(\Phi) = \text{Sdim}(\Phi)$ . We now show that the covering numbers are bounded in terms of the fat-shattering dimension.

**Corollary 11** Suppose  $\mathcal{H} \subseteq [-1, 1]$ . Then for any  $\alpha > 0$ , any  $T > 0$ , and any  $(\mathcal{F} \times \mathcal{X})$ -valued tree  $(\mathbf{f}, \mathbf{x})$  of depth  $T$ ,

$$\mathcal{N}_1(\alpha, \Phi, (\mathbf{f}, \mathbf{x})) \leq \mathcal{N}_2(\alpha, \Phi, (\mathbf{f}, \mathbf{x})) \leq \mathcal{N}_\infty(\alpha, \Phi, (\mathbf{f}, \mathbf{x})) \leq \left(\frac{2eT}{\alpha}\right)^{\text{fat}_\alpha(\Phi)}$$

The generality of these results is evident, as both the combinatorial parameters and covering numbers are defined for any performance measure (1) with time-invariant payoff transformations. In particular, this includes  $\Phi$ -regret (see Section 4.1).

---

2. As an aside, the term “shattered set” was introduced by J. Michael Steele in his Ph.D. thesis in 1975.

### 3.4. Covering Number Bounds for Slowly-Varying Payoff Transformations

In this subsection, we lift the assumption of time-invariance and observe that size of  $\Phi_T$  or an appropriately behaving covering number  $\mathcal{N}(\alpha, \Phi_T, T)$  is key for bounding the sequential complexity. If payoff transformations change wildly in time, there is little hope of getting non-trivial bounds. Under some assumptions on the variability of the sequences in  $\Phi_T$ , we can get a bound on the covering number of  $\Phi_T$ . It has been shown in Herbster and Warmuth (1998); Bousquet and Warmuth (2002) that it is possible to have small external regret against comparators that change a limited number of times. In Zinkevich (2003), *dynamic regret* is defined with respect to a comparator whose path length is bounded. In general, one can consider situations where we would like to compete with a budgeted comparator. We now show that the assumptions of slowly-varying or budgeted comparators are naturally captured by our framework through the notion of slowly-changing payoff transformations  $\Phi_T$ . Furthermore, the control of covering numbers of  $\Phi_T$  becomes transparent under such assumptions. Our goal here is not to provide a comprehensive list of possible results, but rather to show versatility of our framework.

Suppose  $\Phi_T$  consists of payoff transformations  $(\phi_1, \dots, \phi_T)$  which are “almost” time-invariant within each of  $k + 1$  intervals. Consider the following definition:

$$\begin{aligned} \Phi_T^{k,\alpha} = \left\{ (\phi_1, \dots, \phi_T) : 1 = i_0 \leq \dots \leq i_k \leq T, \right. \\ \left. \sup_{f,x} \|\ell_{\phi_t}(f, x) - \ell_{\phi_{t'}}(f, x)\| \leq \alpha \text{ if } i_s \leq t \leq t' < i_{s+1} \text{ for } s \geq 0 \right\}. \end{aligned}$$

One can think of the time-invariant segments as “accumulation points” where the payoff transformations do not vary much. This, of course, includes the case when  $\Phi_T$  is constant over the  $k + 1$  intervals by setting  $\alpha = 0$ . The following result controls the covering number of  $\Phi_T^{k,\alpha}$ .

**Lemma 12** *If  $\mathcal{N}_\infty(\alpha, \Phi, T)$  is finite,  $\mathcal{N}_\infty(2\alpha, \Phi_T^{k,\alpha}, T) \leq \binom{T}{k} \cdot \mathcal{N}_\infty(\alpha, \Phi, T)^{k+1}$ .*

Further extending the above results, we will now study the size of an online cover if  $\Phi_T$  consists of payoff transformations of bounded length. In general, “length” can be defined as some budget given by the setting at hand. Here, we present a straightforward approach without an attempt to give very general and tight bounds. The length of a sequence  $(\phi_1, \dots, \phi_T)$  of payoff transformations (with respect to  $L_\infty$  distance) is defined as  $\text{len}(\phi_1, \dots, \phi_T) := \sum_{t=1}^{T-1} \sup_{f,x} \|\ell_{\phi_t}(f, x) - \ell_{\phi_{t+1}}(f, x)\|$ .

**Lemma 13** *Assume that for all  $(\phi_1, \dots, \phi_T) \in \Phi_T$ , we have  $\text{len}(\phi_1, \dots, \phi_T) \leq L$ . Then, we have,*

$$\mathcal{N}_\infty(2\alpha, \Phi_T, T) \leq \binom{T}{L/\alpha} \cdot \mathcal{N}_\infty(\alpha, \Phi, T)^{L/\alpha+1}.$$

## 4. Examples and Comparison to Known Results

We now turn to several specific settings studied in the literature and look at them through the prism of our general results. While we believe that online learnability in many different

scenarios can be established through our framework, we decided to focus on several major problems. On the surface, these problems are quite different; yet, through our unified approach we show that learnability can be seamlessly established for all of them. The unification not only leads to simpler proofs and sharper results, but also yields insight into the inherent complexity of problems.

#### 4.1. $\Phi$ -Regret

In this section, we consider a particular notion of performance measure, known as  $\Phi$ -regret (Stoltz and Lugosi, 2007; Gordon et al., 2008; Hazan and Kale, 2007). In our framework, this means that we restrict ourselves to only *time-invariant departure mapping classes*  $\Phi_T$  specified by a base class  $\Phi$  of mappings from  $\mathcal{F}$  to itself. The particular choices of  $\Phi$  lead to various notions, such as external, internal, swap regret, and more. To define  $\Phi$ -regret, we fix a set  $\Phi$  of departure mappings which map  $\mathcal{F}$  to  $\mathcal{F}$  and define the set of time-invariant departure mappings  $\Phi_T := \{(\phi, \dots, \phi) : \phi \in \Phi\}$ . Then the measure of performance becomes  $\Phi$ -regret (Eq. (3)). Since  $\mathbf{B}$  is the average of its arguments, Corollary 7 implies that in the setting of  $\Phi$ -regret,  $\mathcal{V}_T(\ell, \Phi) \leq 2\mathfrak{R}(\ell, \Phi)$ . Specializing the definition of sequential complexity to  $\Phi$ -regret, we obtain:

**Definition 14** *The sequential complexity for  $\Phi$ -regret is defined as*

$$\mathfrak{R}_T(\ell, \Phi) = \sup_{(\mathbf{f}, \mathbf{x})} \mathbb{E}_{\epsilon_{1:T}} \sup_{\phi \in \Phi} \frac{1}{T} \sum_{t=1}^T \epsilon_t \ell(\phi \circ \mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) . \quad (10)$$

Sequential complexity for  $\Phi$ -regret enjoys some of the nice properties of the sequential Rademacher complexity for external regret. Suppose  $\ell$  is convex in the first argument and  $\text{conv}(\Phi)$  maps  $\mathcal{F}$  into  $\mathcal{F}$ . Then  $\mathfrak{R}_T(\ell, \text{conv}(\Phi)) = \mathfrak{R}_T(\ell, \Phi)$ . This allows us to obtain bounds for convex hulls of finite sets  $\Phi$ .

To capture complexity via covering numbers, Definition 5 can be specialized to the case of  $\Phi$ -regret:

**Definition 15** *A set  $V$  of  $\mathbb{R}$ -valued trees of depth  $T$  is an  $\alpha$ -cover (with respect to  $\ell_p$ -norm) of  $\Phi_T$  on the  $\mathcal{F} \times \mathcal{X}$ -valued tree  $(\mathbf{f}, \mathbf{x})$  of depth  $T$  if*

$$\forall \phi \in \Phi, \forall \epsilon \in \{\pm 1\}^T \exists \mathbf{v} \in V \text{ s.t. } \frac{1}{T} \sum_{t=1}^T |\mathbf{v}_t(\epsilon) - \ell(\phi \circ \mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon))|^p \leq \alpha^p$$

##### 4.1.1. EXTERNAL REGRET

External regret is the simplest example of  $\Phi$ -regret. We separate it from the general discussion in order to show that for external regret the various notions introduced in this paper reduce to the ones proposed in *RST*. Considering the definitions in Example 1, notice that the time-invariant departure mappings class  $\Phi_T$  is chosen to be the class of sequences of *constant* mappings  $\{(\phi_f, \dots, \phi_f) : f \in \mathcal{F} \text{ and } \phi_f(g) = f \forall g \in \mathcal{F}\}$ . It is precisely because of this constancy of  $\phi$  that the dependence on the  $\mathcal{F}$ -valued tree  $\mathbf{f}$  disappears from all the definitions and results. Further, because of the obvious bijection between elements of  $\Phi_T$  and  $\mathcal{F}$ , minimization (maximization) over  $\Phi_T$  can be written as minimization (maximization) over  $\mathcal{F}$ . Notice that the action of  $\phi_f$  on the payoff is  $\ell_{\phi_f}(f_t, x_t) = \ell(f, x_t)$ . Let us

turn to Definition 14 of the sequential complexity for  $\Phi$ -regret. Because each  $\phi_f \in \Phi$  is a constant mapping, we have

$$\mathfrak{R}_T(\ell, \Phi) = \sup_{\mathbf{f}, \mathbf{x}} \mathbb{E}_{\epsilon_{1:T}} \sup_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^T \epsilon_t \ell(f, \mathbf{x}_t(\epsilon)) = \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{1:T}} \sup_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^T \epsilon_t \ell(f, \mathbf{x}_t(\epsilon)). \quad (11)$$

If payoff is written as  $\ell(f, x) = f(x)$ , this is precisely the sequential Rademacher complexity defined in *RST*. Next, we show that Definition 15 reduces to the definition of online covering given in *RST*. Indeed,  $\ell_{\phi_f}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) = \ell(f, \mathbf{x}_t(\epsilon))$  for the constant mappings  $\phi = (\phi_f, \dots, \phi_f)$ . Further, the payoff space  $\mathcal{H} \subseteq \mathbb{R}$ . With these simplifications, the closeness to a covering element in Definition 15 becomes

$$\forall f \in \mathcal{F}, \forall \epsilon \in \{\pm 1\}^T \exists \mathbf{v} \in V \text{ s.t. } \frac{1}{T} \sum_{t=1}^T |\mathbf{v}_t(\epsilon) - \ell(f, \mathbf{x}_t(\epsilon))|^p \leq \alpha^p$$

where  $V$  is a set of  $\mathbb{R}$ -valued trees. It is then immediate that Corollary 7 recovers the corresponding result of *RST*. For a detailed study of external regret, we refer to the companion paper *RST*.

#### 4.1.2. INTERNAL AND SWAP REGRET

Assume the cardinality  $N = |\mathcal{F}|$  is finite. For internal regret,  $\Phi$  is the set of mappings  $\{\phi_{f \rightarrow g} : \phi_{f \rightarrow g}(f) = g \text{ and } \phi_{f \rightarrow g}(h) = h \forall h \neq f, h \in \mathcal{F}\}$ . For swap regret (Blum and Mansour, 2005; Cesa-Bianchi and Lugosi, 2006),  $\Phi$  contains all  $N^N$  functions from  $\mathcal{F}$  to itself. It is easy to see that applying Corollary 7 in the finite class case ( $|\Phi_T| < \infty$ ) immediately recovers the  $O(\sqrt{T \log N})$  bound for internal and external regret and the  $O(\sqrt{TN \log N})$  bound for the swap regret (Cesa-Bianchi and Lugosi, 2006). Our general tools, however, allow us to go well beyond finite sets of departure mappings. In the following sections, we consider several examples of infinite classes of departure mappings which have been considered in the literature. In some of these cases, an explicit strategy requires computation of a fixed-point (Foster and Vohra, 1997; Hazan and Kale, 2007; Gordon et al., 2008). Since we are not providing efficient algorithms in order to obtain bounds, we are able to get sharp results by directly focusing on the complexity of these infinite classes of departure mappings.

#### 4.1.3. CONVERGENCE TO $\Phi$ -CORRELATED EQUILIBRIA

A beautiful result of Foster and Vohra (1997) shows that convergence to the set of correlated equilibria can be achieved if players follow *internal* regret minimization strategies. What is surprising, no coordination is required to achieve this goal. Stoltz and Lugosi (2007) extended this result to compact and convex sets of strategies in normed spaces. In this section we show that their results can be improved in certain situations. Let us consider their setting in a bit more detail. Suppose there are  $N$  players each playing in a strategy set  $\mathcal{F}$ . We could make the strategy set player dependent but it only complicates notation. There are  $N$  loss functions mapping a strategy profile  $(f_1, \dots, f_N)$  to  $\{\ell_k(f_1, \dots, f_N)\}_{k=1}^N$ , the losses for each of the  $N$  players. Consider a set of departure mappings  $\Phi \subseteq \{\phi : \mathcal{F} \rightarrow \mathcal{F}\}$ . A  $\Phi$ -correlated equilibrium is a distribution  $\pi$  over strategy profiles such that if the player

jointly play according to it, no player has an incentive to unilaterally transform its action using a mapping from  $\Phi$ . That is,  $\mathbb{E}_{(f_1, \dots, f_N) \sim \pi} [\ell_k(f_k, f_{-k})] \leq \mathbb{E}_{(f_1, \dots, f_N) \sim \pi} [\ell_k(\phi(f_k), f_{-k})]$  for all  $k \in [N], \phi \in \Phi$ . Theorem 18 in Stoltz and Lugosi (2007) shows the following. If  $\mathcal{F}$  is convex compact subset of a normed vector space,  $\ell_k$ 's are continuous and  $\Phi$  is a separable subset of  $\mathcal{C}(\mathcal{F})^3$ , then there exist regret minimizing algorithms such that, if every player follows the algorithm then the sequence of empirical plays jointly converges to the set of  $\Phi$ -correlated equilibria.

Consider the case where  $\mathcal{F}$  is some compact subset of the unit ball in some normed space with a norm  $\|\cdot\|$ , the loss function  $\ell_k$  is a 1-Lipschitz convex function, and the class  $\Phi$  of departure functions has finite metric entropy  $\mathcal{N}_{\text{metric}}(\Phi, \alpha)$  for all  $\alpha > 0$ . Metric entropy is simply the log covering number where covers of  $\Phi$  are built for the supremum norm  $\|\phi\|_\infty = \sup_{f \in \mathcal{F}} \|\phi(f)\|$ . Let us consider a typical situation where  $\mathcal{N}_{\text{metric}}(\Phi, \alpha) = \Theta(1/\alpha^p)$ . The adversary's set  $\mathcal{X}$  here is simply  $\{f \mapsto \ell_k(f, g) : g \in \mathcal{F}^{k-1}\}$ , where  $g$  is a strategy profile over the remaining  $k - 1$  players. To upper bound the  $\Phi$ -regret we can always make the set of adversary's moves larger. In fact, we may set  $\mathcal{X} = \mathcal{C}_\mathcal{F}$ , where  $\mathcal{C}_\mathcal{F} = \{x : \mathcal{F} \rightarrow \mathbb{R} : x \text{ convex and 1-Lipschitz}\}$ . Moreover, the value of the convex-Lipschitz game is equal to the value of the linear game (see Rakhlin et al. (2010b)):  $\mathcal{V}_T(\mathcal{C}_\mathcal{F}, \mathcal{F}, \Phi) = \mathcal{V}_T(\mathcal{L}_\mathcal{F}, \mathcal{F}, \Phi)$  where  $\mathcal{L}_\mathcal{F} = \{x : \mathcal{F} \rightarrow \mathbb{R} \text{ is linear and 1-Lipschitz}\}$ . Then the sequential complexity bound is

$$\sup_{(\mathbf{f}, \mathbf{x})} \mathbb{E}_{\epsilon_{1:T}} \sup_{\phi \in \Phi} \frac{1}{T} \sum_{t=1}^T \epsilon_t \langle \phi(\mathbf{f}_t(\epsilon)), \mathbf{x}_t(\epsilon) \rangle . \quad (12)$$

Note that the set  $\mathcal{X}$  is now just the set of 1-Lipschitz linear functions. Since  $\|\phi_1 - \phi_2\|_\infty \leq \alpha$  implies  $|\langle \phi_1(f), x \rangle - \langle \phi_2(f), x \rangle| \leq \alpha$  for any  $x \in \mathcal{X}$ , we can use metric entropy inside Dudley's integral to upper bound the sequential complexity by  $c \inf_\alpha \{\alpha T + \sqrt{T} \int_{\alpha'=\alpha}^1 \sqrt{1/\alpha'^p} d\alpha'\}$ . This behaves as  $O(\sqrt{T})$ , if  $p < 2$ , as  $O(\sqrt{T \log(T)})$  if  $p = 2$ , and as  $O(T^{(p-1)/p})$  if  $p > 2$ . These are better than the  $O(T^{(p+1)/(p+2)})$  bound (derived using an explicit learning algorithm) given in Example 23 of Stoltz and Lugosi (2007).

## 4.2. Blackwell's Approachability

Blackwell's Approachability Theorem (Blackwell, 1956; Mertens et al., 1994; Lehrer, 2003; Cesa-Bianchi and Lugosi, 2006) is a fundamental result for repeated two-player zero-sum games. By means of this theorem, learnability (Hannan consistency) can be established for a wide array of problems, as illustrated in Cesa-Bianchi and Lugosi (2006). For instance, existence of calibrated forecasters can be deduced from Blackwell's Approachability Theorem (Mannor and Stoltz, 2010; Foster and Vohra, 1997). Let us first discuss the relation of our results to Blackwell's Theorem. A proof of Blackwell's Theorem (e.g. Cesa-Bianchi and Lugosi (2006)) reveals that (a) martingale convergence has to take place in the payoff space, and (b) the so-called Blackwell's one-shot approachability condition has to be satisfied. The former is closely related to the first term in our Triplex Inequality, while the latter is related to the second term (ability to play well if the next move is known). What is interesting, in the literature, Blackwell's Theorem is applied by embedding the problem at hand into an often high-dimensional space. The dimensionality represents the complexity of the problem, but this embedding is often artificial. In contrast, the problem complexity is captured by

---

3. The set of continuous functions on  $\mathcal{F}$  equipped with the supremum norm.

the third term of our decomposition, the *sequential complexity*, and it is explicitly written as a complexity measure rather than an embedding into some other space. The ability to upper bound problem complexity with tools similar to those developed in *RST* (e.g. covering numbers) means that learnability can be established for a wide class of problems. In this section, we show that Blackwell's approachability can be viewed as an online game with a particular performance measure (distance to the set). Using the techniques developed in this paper, we prove Blackwell's approachability in Banach spaces for which martingale convergence holds (Theorem 16). We also show that martingale convergence is necessary for the result to hold (Theorem 17). To the best of our knowledge, both of these results are novel. To define the problem precisely, suppose  $\mathcal{H}$  a subset of a Banach space  $\mathcal{B}$  and  $S \subset \mathcal{B}$  is a closed convex set. For the moves  $f \in \mathcal{F}$  of the player and  $x \in \mathcal{X}$  of the adversary,  $\ell(f, x) \in \mathcal{H}$  is a Banach space valued signal. The goal of the player is to keep the average of the signals  $\frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t)$  close to the set  $S$ . By defining  $\mathbf{B}$  as in Example 3,  $\mathbf{R}_T$  becomes distance of the average payoff to the set (see Eq. (4)). The comparator term is zero by our assumption that  $\Phi_T$  contains sequences  $(\phi_1, \dots, \phi_T)$  of constant mappings which transform our actions to a point inside  $S$ :  $\ell_{\phi_t}(f, x) = c_t \in S$  for all  $f \in \mathcal{F}$ ,  $x \in \mathcal{X}$ , and  $1 \leq t \leq T$ .

The Blackwell's approachability game is said to be *one shot approachable* if for every mixed strategy  $p$  of the adversary, there exists a mixed strategy  $q$  for a player such that  $\ell(q, p) \in S$ . This condition says that the player should be able to choose a “good” mixed strategy  $q$  in response to a given adversarial strategy  $p$ . Recall that  $\ell(q, p)$  is simply a shorthand for the expected payoff  $\mathbb{E}_{f \sim q, x \sim p} \ell(f, x)$  (we make no assumptions about linearity of  $\ell$ ). Blackwell's one-shot approachability condition is akin the second term in the Triplex Inequality, where the order of who plays first is switched. If the one-shot condition is satisfied, it remains to check martingale convergence. We now show that, under the one-shot approachability condition, a variation of the worst-case martingale in the subset of the Banach space provides an upper bound on the distance to the set.

**Theorem 16** *For any game that is one shot approachable, we have that*

$$\mathcal{V}_T(\ell, \Phi_T) \leq 4 \sup_{\mathbf{M}} \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=1}^T d_t \right\| \right]$$

where  $\sup$  is over distributions  $\mathbf{M}$  of  $\text{conv}(\mathcal{H} \cup -\mathcal{H})$ -valued martingale difference sequences  $\{d_t\}_{t \in \mathbb{N}}$ .

The notion of approachability considered in this paper corresponds to weak approachability. Extending the techniques of this work to a slightly different notion of a value (see Rakhlin et al. (2010b)), we can guarantee almost sure convergence and, hence, strong approachability.

It is straightforward that for any Blackwell's approachability game to have vanishing regret, one shot approachability for the game is a necessary condition. We now show that martingale convergence in the space of payoffs is necessary for Blackwell's approachability. To the best of our knowledge, this result has not appeared in the literature.

**Theorem 17** *For every symmetric convex set  $\mathcal{H}$  there exists a one shot approachable game with payoff's mapping to  $\mathcal{H}$  such that*

$$\mathcal{V}_T(\ell, \Phi_T) \geq \frac{1}{2} \sup_{\mathbf{M}} \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=1}^T d_t \right\| \right]$$

where  $\sup$  is over distributions  $\mathbf{M}$  of  $\mathcal{H}$ -valued martingale difference sequences  $\{d_t\}_{t \in \mathbb{N}}$ .

### 4.3. Calibration

Calibration is an important notion for forecasting binary sequences (Dawid, 1982). Example 4 corresponds to the notion of  $\lambda$ -calibration for  $\{1, \dots, k\}$ -valued sequences (Cesa-Bianchi and Lugosi, 2006) and defines the measure of performance  $\mathbf{R}$ . We are interested in sharp rates on the value of the calibration game and compare our results with the recent work of Mannor and Stoltz (2010). Note that the definition of value allows the worst scale  $\lambda$  to be chosen at the end of the game, making it a stronger requirement than what is required for building a well calibrated forecaster. Using our techniques, for the  $\ell_1$ -calibration game with  $k$  outcomes, for  $T \geq 3$  and some absolute constant  $c$ , we show that

$\mathcal{V}_T(\ell, \Phi_T) \leq ck^2 ((\log T)/T)^{1/2}$ . That is, the rate of calibration is  $\tilde{O}(T^{-1/2})$ . For  $k > 2$ , the best rates known to us (due to Mannor and Stoltz (2010)) deteriorate with  $k$  because the authors in fact calibrate with respect to all Borel sets.

### 4.4. External Regret with Global Costs

Let us first state a more general setting where the (vector) loss is  $\ell(f, x)$  rather than the specific choice  $f \odot x$  in Example 5. To state the result we need the following Assumption.

**Assumption 1** For any  $p_1, p_2$ ,  $\inf_f \|\ell(f, p_1) + \ell(f, p_2)\| \geq \inf_f \|\ell(f, p_1)\| + \inf_f \|\ell(f, p_2)\|$ .

**Theorem 18** For the setting of Example 5 with vector valued loss  $\ell(f, x)$ , under Assumption 1 :

$$\mathcal{V}_T \leq 4 \sup_{\mathbf{M}} \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=1}^T d_t \right\| \right] + 2 \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{1:T}} \sup_{f \in \mathcal{F}} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \ell(f, \mathbf{x}_t(\epsilon)) \right\| , \quad (13)$$

where  $\sup$  is over distributions  $\mathbf{M}$  of  $\text{conv}(\mathcal{H} \cup -\mathcal{H})$ -valued martingale difference sequences  $\{d_t\}_{t \in \mathbb{N}}$ .

Let us see what this implies in a specific case of Example 5, the setting studied in Even-Dar et al. (2009), i.e.  $\ell(f, x) = f \odot x$ . Let us first verify if Assumption 1 holds here. By linearity of the vector loss, we just have to verify whether, for arbitrary  $p_1, p_2$ , we have

$$\inf_{q \in \Delta(k)} \|q \odot \underline{p_1} + q \odot \underline{p_2}\| \geq \inf_{q \in \Delta(k)} \|q \odot \underline{p_1}\| + \inf_{q \in \Delta(k)} \|q \odot \underline{p_2}\| .$$

where the notation  $\underline{p_i}$  stands for the mean of the distribution  $p_i$ . This is equivalent to asking whether the function  $x \mapsto \inf_{f \in \mathcal{F}} \|f \odot x\|$  is *concave*. Lemma 22 in the appendix proves that it is. Note that in Even-Dar et al. (2009), it is shown that the above function is concave for the  $\ell_p$  norms (including  $p = \infty$ ). It turns out that it remains concave no matter what norm is chosen. Thus, the general upper bound (13) holds. In the case we are considering, we can further massage the second term in that upper bound. Note that for any  $f$  and  $y$ ,  $\|f \odot y\| \leq \|f\|_\infty \|y\| \leq \|y\|$ . Using this in (13) we see that

$$\mathcal{V}_T \leq 4 \sup_{\mathbf{M}} \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=1}^T d_t \right\| \right] + 2 \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{1:T}} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \mathbf{x}_t(\epsilon) \right\| \leq 6 \sup_{\mathbf{M}} \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=1}^T d_t \right\| \right]$$

where the last inequality is because  $(\epsilon_t \mathbf{x}_t(\epsilon))_{t=1}^T$  is a martingale difference sequence. In the last inequality the supremum is over distributions  $\mathbf{M}$  of  $[-1, 1]^k$ -valued martingale difference sequences  $\{d_t\}_{t \in \mathbb{N}}$ . For  $\ell_p$  norms we recover the rates in Even-Dar et al. (2009), specifically for  $\ell_\infty$  norm the bound is  $6\sqrt{\log(k)/T}$

## Acknowledgments

We thank Dean Foster for many insightful discussions about calibration and Blackwell's approachability. A. Rakhlin gratefully acknowledges the support of NSF under grant CAREER DMS-0954737 and Dean's Research Fund.

## References

- J. Abernethy, A. Agarwal, P. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.
- S. Ben-David, D. Pal, and S. Shalev-Shwartz. Agnostic online learning. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.
- D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.
- A. Blum and Y. Mansour. From external to internal regret. In *Proceedings of the 18th Annual Conference on Learning Theory*, pages 621–636. Springer, 2005.
- O. Bousquet and M. K. Warmuth. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3:363–396, 2002.
- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- A.P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.
- E. Even-Dar, R. Kleinberg, S. Mannor, and Y. Mansour. Online learning for global cost functions. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.
- D. P. Foster and R. V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40–55, October 1997.
- E. Giné and J. Zinn. Some limit theorems for empirical processes. *Ann. Prob.*, 12(4): 929–989, 1984.
- P.W. Goldberg and M.R. Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2):131–148, 1995.
- G.J. Gordon, A. Greenwald, and C. Marks. No-regret learning in convex games. In *Proceedings of the 25th International Conference on Machine learning*, pages 360–367. ACM, 2008.

- E. Hazan and S. Kale. Computational equivalence of fixed points and no regret algorithms, and convergence to equilibria. In *NIPS*, 2007.
- M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.
- E. Lehrer. Approachability in infinite dimensional spaces. *International Journal of Game Theory*, 31(2):253–268, 2003.
- N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 04 1988.
- S. Mannor and G. Stoltz. A geometric proof of calibration. *Mathematics of Operations Research*, 35(4):721–727, 2010.
- J.-F. Mertens, S. Sorin, and S. Zamir. Repeated games: Part A Background material. Technical Report 9420, CORE, Universite Catholique de Louvain, 1994.
- I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Prob.*, 22(4):1679–1706, 1994.
- G. Pisier. Martingales with values in uniformly convex spaces. *Israel Journal of Mathematics*, 20(3):326–350, 1975.
- A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In *NIPS*, 2010a. Full version available at arXiv:1006.1138.
- A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Beyond regret. *ArXiv preprint arXiv:1011.3168v2*, 2010b.
- G. Stoltz and G. Lugosi. Learning correlated equilibria in games with compact sets of strategies. *Games and Economic Behavior*, 59(1):187–208, 2007.
- M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, 2003.

## Appendix

### Appendix A. Proofs for General Upper Bounds (Section 3)

**Proof [of Theorem 2]** The value of the game, defined in (2), is

$$\begin{aligned}\mathcal{V}_T(\ell, \Phi_T) &= \inf_{q_1} \sup_{p_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \inf_{q_T} \sup_{p_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \sup_{\phi \in \Phi_T} \{\mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) - \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T))\} \\ &= \sup_{p_1} \inf_{q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T} \inf_{q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \sup_{\phi \in \Phi_T} \{\mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) - \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T))\}\end{aligned}$$

via an application of the minimax theorem. Adding and subtracting terms to the expression above leads to

$$\begin{aligned}\mathcal{V}_T(\ell, \Phi_T) &= \sup_{p_1} \inf_{q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T} \inf_{q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \left[ \mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) - \mathbb{E}_{\substack{f'_1 \sim q_{1:T} \\ x'_1 \sim p_{1:T}}} \mathbf{B}(\ell(f'_1, x'_1), \dots, \ell(f'_T, x'_T)) \right. \\ &\quad \left. + \sup_{\phi \in \Phi_T} \left\{ \mathbb{E}_{\substack{f'_{1:T} \sim q_{1:T} \\ x'_{1:T} \sim p_{1:T}}} \mathbf{B}(\ell(f'_1, x'_1), \dots, \ell(f'_T, x'_T)) - \mathbf{B}(\ell_{\phi_1}(f'_1, x'_1), \dots, \ell_{\phi_T}(f'_T, x'_T)) \right\} \right] \\ &\leq \sup_{p_1} \inf_{q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T} \inf_{q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \left[ \mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) - \mathbb{E}_{\substack{f'_{1:T} \sim q_{1:T} \\ x'_{1:T} \sim p_{1:T}}} \mathbf{B}(\ell(f'_1, x'_1), \dots, \ell(f'_T, x'_T)) \right. \\ &\quad \left. + \sup_{\phi \in \Phi_T} \mathbb{E}_{\substack{f'_{1:T} \sim q_{1:T} \\ x'_{1:T} \sim p_{1:T}}} \left\{ \mathbf{B}(\ell(f'_1, x'_1), \dots, \ell(f'_T, x'_T)) - \mathbf{B}(\ell_{\phi_1}(f'_1, x'_1), \dots, \ell_{\phi_T}(f'_T, x'_T)) \right\} \right. \\ &\quad \left. + \sup_{\phi \in \Phi_T} \left\{ \mathbb{E}_{\substack{f'_{1:T} \sim q_{1:T} \\ x'_{1:T} \sim p_{1:T}}} \mathbf{B}(\ell_{\phi_1}(f'_1, x'_1), \dots, \ell_{\phi_T}(f'_T, x'_T)) - \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T)) \right\} \right]\end{aligned}$$

At this point, we would like to break up the expression into three terms. To do so, notice that expectation is linear and sup is a convex function, while for the infimum,

$$\inf_a [C_1(a) + C_2(a) + C_3(a)] \leq \left[ \sup_a C_1(a) \right] + \left[ \inf_a C_2(a) \right] + \left[ \sup_a C_3(a) \right]$$

for functions  $C_1, C_2, C_3$ . We use these properties of inf, sup, and expectation, starting from the inside of the nested expression and splitting the expression in three parts. We arrive at

$$\begin{aligned}\mathcal{V}_T(\ell, \Phi_T) &\leq \sup_{p_1} \sup_{q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T} \sup_{q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \left[ \mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) - \mathbb{E}_{\substack{f'_{1:T} \sim q_{1:T} \\ x'_{1:T} \sim p_{1:T}}} \mathbf{B}(\ell(f'_1, x'_1), \dots, \ell(f'_T, x'_T)) \right] \\ &\quad + \sup_{p_1} \inf_{q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T} \inf_{q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \left[ \sup_{\phi \in \Phi_T} \mathbb{E}_{\substack{f'_{1:T} \sim q_{1:T} \\ x'_{1:T} \sim p_{1:T}}} \left\{ \mathbf{B}(\ell(f'_1, x'_1), \dots, \ell(f'_T, x'_T)) - \mathbf{B}(\ell_{\phi_1}(f'_1, x'_1), \dots, \ell_{\phi_T}(f'_T, x'_T)) \right\} \right] \\ &\quad + \sup_{p_1} \sup_{q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T} \sup_{q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \left[ \sup_{\phi \in \Phi_T} \left\{ \mathbb{E}_{\substack{f'_{1:T} \sim q_{1:T} \\ x'_{1:T} \sim p_{1:T}}} \mathbf{B}(\ell_{\phi_1}(f'_1, x'_1), \dots, \ell_{\phi_T}(f'_T, x'_T)) - \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T)) \right\} \right]\end{aligned}$$

The replacement of infima by suprema in the first and third terms appears to be a loose step and, indeed, one can pick a particular response strategy  $\{q_t^*\}$  instead of passing to the supremum. For instance, this can be the best-response strategy for the second term. However, in the examples we have considered so far, passing to the supremum still yields the results we need. This is due to the fact that the online learning setting is worst-case.

Consider the second term in the above decomposition. We claim that

$$\begin{aligned} & \sup_{p_1} \inf_{q_1} \mathbb{E}_{f_1 \sim q_1} \dots \sup_{p_T} \inf_{q_T} \mathbb{E}_{f_T \sim q_T} \left[ \sup_{\phi \in \Phi_T} \mathbb{E}_{f'_1, \dots, f'_{T-1} \sim q_{1:T}} [\mathbf{B}(\ell(f'_1, x'_1), \dots, \ell(f'_T, x'_T)) - \mathbf{B}(\ell_{\phi_1}(f'_1, x'_1), \dots, \ell_{\phi_T}(f'_T, x'_T))] \right] \\ &= \sup_{p_1} \dots \sup_{q_1} \sup_{p_T} \mathbb{E}_{f_1, \dots, f_T \sim p_{1:T}} [\mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) - \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T))] \end{aligned}$$

because the objective

$$\mathbb{E}_{f'_1, \dots, f'_{T-1} \sim q_{1:T}} [\mathbf{B}(\ell(f'_1, x'_1), \dots, \ell(f'_T, x'_T)) - \mathbf{B}(\ell_{\phi_1}(f'_1, x'_1), \dots, \ell_{\phi_T}(f'_T, x'_T))]$$

does not depend on the random draws  $f_1, x_1, \dots, f_T, x_T$ . We then rename  $f'_t, x'_t$  into  $f_t, x_t$ . This concludes the proof of the Triplex Inequality.  $\blacksquare$

**Proof [of Theorem 3]** We turn to the third term in the Triplex Inequality. If  $\mathbf{B}$  is subadditive,

$$\begin{aligned} & \mathbb{E}_{f'_1, \dots, f'_{T-1} \sim q_{1:T}} [\mathbf{B}(\ell_{\phi_1}(f'_1, x'_1), \dots, \ell_{\phi_T}(f'_T, x'_T)) - \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T))] \\ & \leq \mathbb{E}_{f'_1, \dots, f'_{T-1} \sim q_{1:T}} [\mathbf{B}(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T))]. \end{aligned}$$

If, on the other hand,  $-\mathbf{B}$  is subadditive,

$$\begin{aligned} & \mathbb{E}_{f'_1, \dots, f'_{T-1} \sim q_{1:T}} [\mathbf{B}(\ell_{\phi_1}(f'_1, x'_1), \dots, \ell_{\phi_T}(f'_T, x'_T)) - \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T))] \\ & \leq - \mathbb{E}_{f'_1, \dots, f'_{T-1} \sim q_{1:T}} [\mathbf{B}(\ell_{\phi_1}(f_1, x_1) - \ell_{\phi_1}(f'_1, x'_1), \dots, \ell_{\phi_T}(f_T, x_T) - \ell_{\phi_T}(f'_T, x'_T))]. \end{aligned} \quad (14)$$

Below assume that  $\mathbf{B}$  is subadditive, and the proof of the other case is identical.

To prove the bound on the third term in terms of twice sequential complexity, we proceed as in Rakhlin et al. (2010a), applying the symmetrization technique from inside out. To this end, first note that,

$$\begin{aligned} & \sup_{p_1, q_1} \mathbb{E}_{f_1 \sim q_1} \dots \sup_{p_T, q_T} \mathbb{E}_{f_T \sim q_T} \sup_{\phi \in \Phi_T} \mathbb{E}_{f'_1 \sim q_1, \dots, f'_{T-1} \sim q_{T-1}, x'_1 \sim p_1, \dots, x'_{T-1} \sim p_{T-1}} \mathbf{B}(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T)) \\ & \leq \sup_{p_1, q_1} \mathbb{E}_{f_1, f'_1 \sim q_1} \dots \sup_{p_T, q_T} \mathbb{E}_{f_T, f'_T \sim q_T} \sup_{\phi \in \Phi_T} \mathbf{B}(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T)). \end{aligned}$$

The above is true because the expectations are pulled outside the suprema, thus resulting in an upper bound. Now notice that conditioned on history  $f_T, f'_T$  are distributed identically and independently drawn from  $q_T$ . Similarly  $x_T, x'_T$  are also identically distributed conditioned on history. Hence renaming them we see that

$$\begin{aligned} & \mathbb{E}_{f_T, f'_T \sim q_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T)\right) \\ &= \mathbb{E}_{f'_T, f_T \sim q_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T) - \ell_{\phi_T}(f'_T, x'_T)\right) \\ &= \mathbb{E}_{f_T, f'_T \sim q_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, -(\ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T))\right) \end{aligned}$$

where only the last argument of  $\mathbf{B}$  is changing sign. Thus,

$$\begin{aligned} & \mathbb{E}_{f_T, f'_T \sim q_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T)\right) \\ &= \mathbb{E}_{\epsilon_T} \mathbb{E}_{f_T, f'_T \sim q_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \epsilon_T(\ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T))\right) \end{aligned}$$

where  $\epsilon_T$  is a Rademacher random variable. Furthermore,

$$\begin{aligned} & \sup_{p_T, q_T} \mathbb{E}_{f_T, f'_T \sim q_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T)\right) \\ &= \sup_{p_T, q_T} \mathbb{E}_{f'_T, f_T \sim q_T} \mathbb{E}_{\epsilon_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \epsilon_T(\ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T))\right) \\ &\leq \sup_{x_T, x'_T \in \mathcal{X}} \mathbb{E}_{\epsilon_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \epsilon_T(\ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T))\right) \end{aligned}$$

Proceeding similarly notice that since given history  $x_{T-1}, x'_{T-1}$  and  $f_{T-1}, f'_{T-1}$  are distributed independently and identically we have,

$$\begin{aligned} & \sup_{p_{T-1}, q_{T-1}} \mathbb{E}_{f_{T-1}, f'_{T-1} \sim q_{T-1}} \sup_{x_T, x'_T \in \mathcal{X}} \mathbb{E}_{\epsilon_T} \sup_{\phi \in \Phi_T} \\ & \quad \mathbf{B}\left(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_{T-1}}(f'_{T-1}, x'_{T-1}) - \ell_{\phi_{T-1}}(f_{T-1}, x_{T-1}), \epsilon_T(\ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T))\right) \\ &= \sup_{p_{T-1}, q_{T-1}} \mathbb{E}_{f_{T-1}, f'_{T-1} \sim q_{T-1}} \mathbb{E}_{\epsilon_{T-1}} \sup_{x_T, x'_T \in \mathcal{X}} \mathbb{E}_{\epsilon_T} \sup_{\phi \in \Phi_T} \\ & \quad \mathbf{B}\left(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \epsilon_{T-1}(\ell_{\phi_T}(f'_{T-1}, x'_{T-1}) - \ell_{\phi_{T-1}}(f_{T-1}, x_{T-1})), \epsilon_T(\ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T))\right) \\ &\leq \sup_{x_{T-1}, x'_{T-1} \in \mathcal{X}} \mathbb{E}_{\epsilon_{T-1}} \sup_{f_{T-1}, f'_{T-1} \in \mathcal{F}} \mathbb{E}_{\epsilon_T} \sup_{\phi \in \Phi_T} \\ & \quad \mathbf{B}\left(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \epsilon_{T-1}(\ell_{\phi_{T-1}}(f'_{T-1}, x'_{T-1}) - \ell_{\phi_{T-1}}(f_{T-1}, x_{T-1})), \epsilon_T(\ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T))\right) \end{aligned}$$

Proceeding in similar fashion introducing Rademacher random variables all the way to  $\epsilon_1$  we arrive at

$$\begin{aligned} & \sup_{p_1, q_1} \mathbb{E}_{f_1, f'_1 \sim q_1} \dots \sup_{p_T, q_T} \mathbb{E}_{f_T, f'_T \sim q_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T)\right) \\ & \leq \sup_{x_1, x'_1 \in \mathcal{X}} \mathbb{E}_{\epsilon_1} \dots \sup_{x_T, x'_T \in \mathcal{X}} \mathbb{E}_{\epsilon_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\epsilon_1(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1)), \dots, \epsilon_T(\ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T))\right) \end{aligned}$$

Subadditivity of  $\mathbf{B}$  implies  $\mathbf{B}(a - b) \leq \mathbf{B}(a) + \mathbf{B}(-b)$ , and thus

$$\begin{aligned} & \mathbf{B}\left(\epsilon_1(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1)), \dots, \epsilon_T(\ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T))\right) \\ & \leq \mathbf{B}\left(\epsilon_1 \ell_{\phi_1}(f'_1, x'_1), \dots, \epsilon_T \ell_{\phi_T}(f'_T, x'_T)\right) + \mathbf{B}\left(-\epsilon_1 \ell_{\phi_1}(f_1, x_1), \dots, -\epsilon_T \ell_{\phi_T}(f_T, x_T)\right) \end{aligned}$$

We, therefore, arrive at

$$\begin{aligned} & \sup_{x_1, x'_1 \in \mathcal{X}} \mathbb{E}_{\epsilon_1} \dots \sup_{x_T, x'_T \in \mathcal{X}} \mathbb{E}_{\epsilon_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\epsilon_1(\ell_{\phi_1}(f'_1, x'_1) - \ell_{\phi_1}(f_1, x_1)), \dots, \epsilon_T(\ell_{\phi_T}(f'_T, x'_T) - \ell_{\phi_T}(f_T, x_T))\right) \\ & \leq 2 \sup_{f_1 \in \mathcal{F}, x_1 \in \mathcal{X}} \mathbb{E}_{\epsilon_1} \dots \sup_{f_T \in \mathcal{F}, x_T \in \mathcal{X}} \mathbb{E}_{\epsilon_T} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\epsilon_1 \ell_{\phi_1}(f_1, x_1), \dots, \epsilon_T \ell_{\phi_T}(f_T, x_T)\right) \\ & = 2 \sup_{(\mathbf{f}, \mathbf{x})} \mathbb{E}_{\epsilon_{1:T}} \sup_{\phi \in \Phi_T} \mathbf{B}\left(\epsilon_1 \ell_{\phi_1}(\mathbf{f}_1(\epsilon), \mathbf{x}_1(\epsilon)), \dots, \epsilon_T \ell_{\phi_T}(\mathbf{f}_T(\epsilon), \mathbf{x}_T(\epsilon))\right) \end{aligned}$$

where in the last step we passed to the supremum over  $(\mathcal{F} \times \mathcal{X})$ -valued trees. This concludes the proof for the case of  $\mathbf{B}$  being subadditive. Starting from Eq. (14), the proof for the case of  $-\mathbf{B}$  being subadditive and convex in each of its coordinates leads to the bound of

$$2 \sup_{(\mathbf{f}, \mathbf{x})} \mathbb{E}_{\epsilon_{1:T}} \sup_{\phi \in \Phi_T} -\mathbf{B}\left(\epsilon_1 \ell_{\phi_1}(\mathbf{f}_1(\epsilon), \mathbf{x}_1(\epsilon)), \dots, \epsilon_T \ell_{\phi_T}(\mathbf{f}_T(\epsilon), \mathbf{x}_T(\epsilon))\right).$$

The complete proof can be repeated for the first term in the Triplex Inequality in order to bound it by  $2\mathfrak{R}_T(\ell, \mathcal{I}, \mathbf{B})$  (or respectively  $2\mathfrak{R}_T(\ell, \mathcal{I}, -\mathbf{B})$ ).  $\blacksquare$

**Proof [of Lemma 4]** The lemma follows directly from a result on concentration of 2-smooth functions of martingales, due to Pinelis (1994). A detailed proof appears in Rakhlin et al. (2010b).  $\blacksquare$

**Proof [of Theorem 6]** Define  $\beta_0 = \eta$  and  $\beta_j = 2^{-j}$ . For a fixed tree  $(\mathbf{f}, \mathbf{x})$  of depth  $T$ , let  $V_j$  be an  $\ell_\infty$ -cover at scale  $\beta_j$ . For any path  $\epsilon \in \{\pm 1\}^T$  and any  $\phi \in \Phi_T$ , let  $\mathbf{v}[\phi, \epsilon]^j \in V_j$  a

$\beta_j$ -close element of the cover in the  $\ell_\infty$  sense. Now, for any  $\phi \in \Phi_T$ ,

$$\begin{aligned}
& G \left( \frac{1}{T} \sum_{t=1}^T \epsilon_t \ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) \right) \\
& \leq G \left( \frac{1}{T} \sum_{t=1}^T \epsilon_t (\ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) - \mathbf{v}[\phi, \epsilon]_t^N) \right) + \sum_{j=1}^N G \left( \frac{1}{T} \sum_{t=1}^T \epsilon_t (\mathbf{v}[\phi, \epsilon]_t^j - \mathbf{v}[\phi, \epsilon]_t^{j-1}) \right) \\
& \leq \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t (\ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) - \mathbf{v}[\phi, \epsilon]_t^N) \right\| + \sum_{j=1}^N G \left( \frac{1}{T} \sum_{t=1}^T \epsilon_t (\mathbf{v}[\phi, \epsilon]_t^j - \mathbf{v}[\phi, \epsilon]_t^{j-1}) \right) \\
& \leq \max_{t=1}^T \| \ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) - \mathbf{v}[\phi, \epsilon]_t^N \| + \sum_{j=1}^N G \left( \frac{1}{T} \sum_{t=1}^T \epsilon_t (\mathbf{v}[\phi, \epsilon]_t^j - \mathbf{v}[\phi, \epsilon]_t^{j-1}) \right)
\end{aligned}$$

Thus,

$$\sup_{\phi \in \Phi_T} G \left( \frac{1}{T} \sum_{t=1}^T \epsilon_t \ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) \right) \leq \beta_N + \sup_{\phi \in \Phi_T} \left\{ \sum_{j=1}^N G \left( \frac{1}{T} \sum_{t=1}^T \epsilon_t (\mathbf{v}[\phi, \epsilon]_t^j - \mathbf{v}[\phi, \epsilon]_t^{j-1}) \right) \right\}$$

We now proceed to upper bound the second term. Consider all possible pairs of  $\mathbf{v}^s \in V_j$  and  $\mathbf{v}^r \in V_{j-1}$ , for  $1 \leq s \leq |V_j|$ ,  $1 \leq r \leq |V_{j-1}|$ , where we assumed an arbitrary enumeration of elements. For each pair  $(\mathbf{v}^s, \mathbf{v}^r)$ , define a real-valued tree  $\mathbf{w}^{(s,r)}$  by

$$\mathbf{w}_t^{(s,r)}(\epsilon) = \begin{cases} \mathbf{v}_t^s(\epsilon) - \mathbf{v}_t^r(\epsilon) & \text{if there exists } \phi \in \Phi_T \text{ s.t. } \mathbf{v}^s = \mathbf{v}[\phi, \epsilon]^j, \mathbf{v}^r = \mathbf{v}[\phi, \epsilon]^{j-1} \\ 0 & \text{otherwise.} \end{cases}$$

for all  $t \in [T]$  and  $\epsilon \in \{\pm 1\}^T$ . It is crucial that  $\mathbf{w}^{(s,r)}$  can be non-zero only on those paths  $\epsilon$  for which  $\mathbf{v}^s$  and  $\mathbf{v}^r$  are indeed the members of the covers (at successive resolutions) close in the  $\ell_2$  sense to some  $\phi \in \Phi_T$ . It is easy to see that  $\mathbf{w}^{(s,r)}$  is well-defined. Let the set of trees  $W_j$  be defined as

$$W_j = \left\{ \mathbf{w}^{(s,r)} : 1 \leq s \leq |V_j|, 1 \leq r \leq |V_{j-1}| \right\}$$

Using the above notations we see that

$$\begin{aligned}
& \mathbb{E}_\epsilon \left[ \sup_{\phi \in \Phi_T} G \left( \frac{1}{T} \sum_{t=1}^T \epsilon_t \ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) \right) \right] \\
& \leq \beta_N + \mathbb{E}_\epsilon \left[ \sup_{\phi \in \Phi_T} \left\{ \sum_{j=1}^N G \left( \frac{1}{T} \sum_{t=1}^T \epsilon_t (\mathbf{v}[\phi, \epsilon]_t^j - \mathbf{v}[\phi, \epsilon]_t^{j-1}) \right) \right\} \right] \\
& \leq \beta_N + \mathbb{E}_\epsilon \left[ \sum_{j=1}^N \sup_{\mathbf{w}^j \in W_j} G \left( \frac{1}{T} \sum_{t=1}^T \epsilon_t \mathbf{w}_t^j(\epsilon) \right) \right]
\end{aligned} \tag{15}$$

Similarly to the corresponding proof in Rakhlin et al. (2010a), we can show that  $\max_{t=1}^T \|\mathbf{w}_t^j(\epsilon)\| \leq 3\beta_j$  for any  $\mathbf{w}^j \in \mathcal{W}^j$  and any path  $\epsilon$ . Using concentration inequalities for 2-smooth functions in Banach spaces (see Pinelis (1994) or the full version Rakhlin et al. (2010b) of this extended abstract), we get

$$\begin{aligned} \mathbb{E}_\epsilon \left[ \sup_{\phi \in \Phi_T} G \left( \frac{1}{T} \sum_{t=1}^T \epsilon_t \ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) \right) \right] &\leq \beta_N + \sum_{j=1}^N \beta_j \sqrt{\frac{\gamma \log(2|W_j|)}{T}} \\ &\leq \beta_N + \sum_{j=1}^N \beta_j \sqrt{\frac{\gamma \log(2|V_j| \cdot |V_{j-1}|)}{T}} \\ &\leq \beta_N + \frac{6\sqrt{\gamma}}{\sqrt{T}} \sum_{j=1}^N \beta_j \sqrt{\log(|V_j|)} \\ &\leq \beta_N + \frac{12\sqrt{\gamma}}{\sqrt{T}} \sum_{j=1}^N (\beta_j - \beta_{j+1}) \sqrt{\log \mathcal{N}_\infty(\beta_j, \Phi_T, T)}. \end{aligned}$$

Using standard arguments, this gives the bound,

$$\inf_\alpha 4\alpha + \frac{12\sqrt{\gamma}}{\sqrt{T}} \int_\alpha^\eta \sqrt{\log \mathcal{N}_\infty(\beta, \Phi_T, T)} d\beta.$$

■

**Proof [of Corollary 7]** The first statement is trivially verified. In fact, for this to hold we only require that  $\mathbf{B}$  is subadditive, affine in its arguments, and  $\mathbf{B}(0, \dots, 0) = 0$ . Indeed, the expectations can be sequentially moved inside of  $\mathbf{B}$ , making the coordinates of  $\mathbf{B}$  zero, and making the suprema over the distributions irrelevant.

For the second claim, consider the second term in (8), specialized to the case of departure mappings:

$$\sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \sup_{\phi \in \Phi_T} \mathbb{E}_{\substack{f_{1:T} \sim q_{1:T} \\ x_{1:T} \sim p_{1:T}}} \left\{ \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) - \ell(\phi_t(f_t), x_t) \right\} \quad (16)$$

Pick a particular (sub)optimal response  $q_t$  which puts all mass on  $f_t^* = \arg \min_{f \in \mathcal{F}} \mathbb{E}_{x \sim p_t} \ell(f, x)$ . It follows that  $\ell(f_t, x_t) - \ell(\phi_t(f_t), x_t) \leq 0$ , ensuring that the quantity in (16) is non-positive.

The third claim is a straightforward consequence of Theorem 6. Indeed,  $\mathcal{H} \subset [-\eta, \eta]$  and  $G$  is the identity mapping, hence  $G^2$  is  $(2, 2)$ -smooth. ■

**Proof [of Lemma 12]** Fix an  $(\mathcal{F} \times \mathcal{X})$ -valued tree  $(\mathbf{f}, \mathbf{x})$  of depth  $T$ . Let  $(i_0, \dots, i_k)$  be the sequence which defines intervals of time-invariant mappings for the sequence  $(\phi_1, \dots, \phi_T)$ . Fix  $\epsilon \in \{\pm 1\}^T$ . Let  $\mathbf{v}^{i_0}, \dots, \mathbf{v}^{i_k} \in V$  be the elements of the  $L_\infty$  cover closest to  $\phi_{i_0}, \dots, \phi_{i_k}$ , respectively, on the path  $\epsilon$ . That is, for any  $a \in \{i_0, \dots, i_k\}$ ,

$$\max_t \|\ell_{\phi_a}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) - \mathbf{v}_t^a(\epsilon)\| \leq \alpha.$$

By our assumption, on any interval  $I$ , defined by the endpoints  $a = i_j$  and  $b = i_{j+1}$ ,

$$\max_{t \in \{a, \dots, b-1\}} \|\ell_{\phi_a}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) - \ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon))\| \leq \alpha,$$

Hence,

$$\max_{t \in \{a, \dots, b-1\}} \|\ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) - \mathbf{v}_t^a(\epsilon)\| \leq 2\alpha$$

Denoting by  $a(t) \in \{i_0, \dots, i_k\}$  the left endpoint of an interval to which  $t$  belongs,

$$\max_{t \in \{1, \dots, T\}} \|\ell_{\phi_t}(\mathbf{f}_t(\epsilon), \mathbf{x}_t(\epsilon)) - \mathbf{v}_t^{a(t)}(\epsilon)\| \leq 2\alpha$$

It is then clear that to construct a  $2\alpha$ -cover for  $\Phi_T^{k,\alpha}$  in  $L_\infty$  norm, it is enough to concatenate trees in  $V$ . More precisely, this is done as follows. Construct a set  $V^k$  of  $\mathcal{H}$ -valued trees as

$$V^k = \{\mathbf{v}' = \mathbf{v}'(\mathbf{v}^0, \dots, \mathbf{v}^k, i_0, \dots, i_k) : 1 = i_0 \leq i_1 \leq \dots \leq i_k \leq T, \mathbf{v}^0, \dots, \mathbf{v}^k \in V\}$$

and  $\mathbf{v}' = \mathbf{v}'(\mathbf{v}^0, \dots, \mathbf{v}^k, i_0, \dots, i_k)$  is defined as a sequence of  $T$  mappings

$$\mathbf{v}'_t(\epsilon) = \mathbf{v}_t^{a(t)}(\epsilon) \quad t \in I_{a(t)}$$

for any  $\epsilon \in \{\pm 1\}^T$ . Here  $I_a = \{i_j, \dots, i_{j+1} - 1\}$  and  $a(t)$  is the index of the interval to which  $t$  belongs. In plain words, we consider all ways of partitioning  $\{1, \dots, T\}$  into  $k + 1$  intervals and defining a new set of trees out of  $V$  in such a way that within the interval, the values are given by a fixed tree from  $V$ . As before, it is clear that

$$\mathcal{N}_\infty(2\alpha, \Phi_T^{k,\alpha}, T) = |V^k| \leq \binom{T}{k} \cdot \mathcal{N}_\infty(\alpha, \Phi, T)^{k+1},$$

providing a control on the complexity of  $\Phi_T^{k,\alpha}$ . ■

**Proof [of Lemma 13]** We claim that by choosing  $k$  large enough, the set of covering trees  $V^k$  defined in the proof of Lemma 12 provides a cover for  $\Phi_T$  at a given scale  $\alpha > 0$ . Consider any  $(\phi_1, \dots, \phi_T) \in \Phi_T$ . We construct the nondecreasing sequence  $i_1, \dots, i_j, \dots \in \{1, \dots, T\}$  of “change-points” as follows: increase  $t$  until the next payoff transformation is farther than  $\alpha$  from the payoff transformation at  $i_j$ :

$$i_{j+1} = \inf_{t > i_j} \left\{ \sup_{f,x} \|\ell_{\phi_{i_j}}(f, x) - \ell_{\phi_t}(f, x)\| \geq \alpha \right\}$$

Let  $k$  be the length of the largest such sequence for all elements of  $\Phi_T$ . We have simply reduced the problem to the one studied in Lemma 12: within each block, all the payoff transformations are close. Clearly,  $k = k(\alpha) \leq L/\alpha$ , but can potentially be smaller under additional assumptions on  $\Phi_T$ . We then have a bound on the size of a  $2\alpha$ -cover of  $\Phi_T$ :

$$\mathcal{N}_\infty(2\alpha, \Phi_T, T) \leq \binom{T}{k(\alpha)} \cdot \mathcal{N}_\infty(\alpha, \Phi, T)^{k(\alpha)+1} \leq \binom{T}{L/\alpha} \cdot \mathcal{N}_\infty(\alpha, \Phi, T)^{L/\alpha+1}.$$
■

## Appendix B. Techniques for Lower Bounds

It is well-known that an *equalizing strategy* (i.e. a strategy that makes the move of the other player “irrelevant”) can often be shown to be minimax optimal. In this section, we define a notion of an equalizer for our repeated game and show that it can be used to prove *lower bounds* on the value of the game. While existence of an equalizer has to be established for particular problems at hand, the lower bounds below hold whenever such an equalizer exists.

**Definition 19** A strategy  $\{p_t^*\}$  for the adversary is said to be an *equalizer strategy* if

$$\mathbb{E}_{\substack{x_1 \sim p_1^* \\ f_1 \sim q_1^*}} \dots \mathbb{E}_{\substack{x_T \sim p_T^* \\ f_T \sim q_T^*}} \mathbf{R}_T((f_1, x_1), \dots, (f_T, x_T)) = \mathbb{E}_{\substack{x_1 \sim \overline{p}_1^* \\ f_1 \sim \overline{q}_1^*}} \dots \mathbb{E}_{\substack{x_T \sim \overline{p}_T^* \\ f_T \sim \overline{q}_T^*}} \mathbf{R}_T((f_1, x_1), \dots, (f_T, x_T))$$

for all strategies  $\{q_t^*\}$  and  $\{\overline{q}_t^*\}$  of the player. Here  $\mathbf{R}_T$  is defined as in (1).

Using the above definition of an equalizer we have the following proposition as an immediate consequence.

**Proposition 20** For any Equalizer strategy  $\{p_t^*\}$  we have that for any  $f \in \mathcal{F}$ ,

$$\mathcal{V}_T(\ell, \Phi_T) \geq \mathbb{E}_{x_1 \sim p_1} \dots \mathbb{E}_{x_T \sim p_T} \left[ \mathbf{B}(\ell(f, x_1), \dots, \ell(f, x_T)) - \inf_{\phi \in \Phi_T} \mathbf{B}(\ell_{\phi_1}(f, x_1), \dots, \ell_{\phi_T}(f, x_T)) \right]$$

$$\text{where } p_t = p_t^* \left( \{f_s = f, x_s\}_{s=1}^{t-1} \right)$$

**Remark 21** For many interesting games we consider it is often the case that for any  $x_1, \dots, x_T$  and any  $f_1, \dots, f_T, f'_1, \dots, f'_T$ ,

$$\inf_{\phi \in \Phi_T} \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T)) = \inf_{\phi \in \Phi_T} \mathbf{B}(\ell_{\phi_1}(f'_1, x_1), \dots, \ell_{\phi_T}(f'_T, x_T))$$

In these cases since the player’s actions do not even affect the second term of the regret, to check if a strategy  $\{p_t^*\}$  is an equalizer or not we only need to check if

$$\mathbb{E}_{\substack{x_1 \sim p_1^* \\ f_1 \sim q_1^*}} \dots \mathbb{E}_{\substack{x_T \sim p_T^* \\ f_T \sim q_T^*}} \mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) = \mathbb{E}_{\substack{x_1 \sim \overline{p}_1^* \\ f_1 \sim \overline{q}_1^*}} \dots \mathbb{E}_{\substack{x_T \sim \overline{p}_T^* \\ f_T \sim \overline{q}_T^*}} \mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T))$$

for all strategies  $\{q_t^*\}$  and  $\{\overline{q}_t^*\}$  of the player.

Interestingly enough, many of the existing lower bounds in online learning literature are, in fact, equalizers (see e.g. (Cesa-Bianchi and Lugosi, 2006, p. 252)). In particular, in Abernethy et al. (2009), a lower bound on the value of the game was derived by looking at a certain *face* of a convex hull of loss vectors. The face, supported by a probability distribution  $p$ , corresponds to the set of functions with the same expected loss under the distribution  $p$ . Hence,  $p$  is an equalizing strategy for those functions. Since these functions are the “best” with respect to this distribution, a lower bound in terms of complexity of this set was derived in Abernethy et al. (2009). Furthermore, (Lee, Bartlett and Williamson, 1998) shows that a lower bound on the rate of convergence in the i.i.d. setting is achieved when there are two distinct minimizers of expected error for a given distribution. Again, this distribution can be viewed as an equalizer for the non-singleton set of minimizers of expected error.

### Appendix C. Proofs for Blackwell Approachability (Section 4.2)

**Proof [of Theorem 16]** Now we apply Theorem 2 to the Blackwell Approachability game. Note that for any sequence  $(\phi_1, \dots, \phi_T)$ ,  $\phi_t$  maps the payoff to some element of  $S$ . Hence,

$$\mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T)) = 0$$

for any  $f_1, \dots, f_T \in \mathcal{F}$ ,  $x_1, \dots, x_T \in \mathcal{X}$ . We then conclude that

$$\begin{aligned} \mathcal{V}_T(\ell, \Phi_T) &\leq \sup_{p_1, q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T, q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \left\{ \mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) - \mathbb{E}_{\substack{f'_1 \sim q_{1:T} \\ x'_1 \sim p_{1:T}}} \mathbf{B}(\ell(f'_1, x'_1), \dots, \ell(f'_T, x'_T)) \right\} \\ &\quad + \sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \mathbb{E}_{\substack{f_1:T \sim q_{1:T} \\ x_1:T \sim p_{1:T}}} \mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)). \end{aligned} \tag{17}$$

We remark for the upper bound to hold it is enough to assume that  $\Phi_T$  contains *some* sequence that maps the payoffs to some element of  $S$ .

Consider the two terms in the above bound separately. The first term can be written as

$$\begin{aligned} &\sup_{p_1, q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T, q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \mathbb{E}_{\substack{f'_1:T \sim q_{1:T} \\ x'_1:T \sim p_{1:T}}} \left\{ \inf_{c \in S} \left\| c - \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) \right\| - \inf_{c' \in S} \left\| c' - \frac{1}{T} \sum_{t=1}^T \ell(f'_t, x'_t) \right\| \right\} \\ &\leq \sup_{p_1, q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T, q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \mathbb{E}_{\substack{f'_1:T \sim q_{1:T} \\ x'_1:T \sim p_{1:T}}} \left\{ \left\| \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) - \frac{1}{T} \sum_{t=1}^T \ell(f'_t, x'_t) \right\| \right\} \\ &\leq \sup_{p_1, q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T, q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \left\{ \left\| \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) - \frac{1}{T} \sum_{t=1}^T \ell(f'_t, x'_t) \right\| \right\} \end{aligned}$$

where in the first inequality we used  $\inf_a [C_1(a)] - \inf_a [C_2(a)] \leq \sup_a [C_1(a) - C_2(a)]$  along with a triangle inequality. This is now bounded by

$$2 \sup_{\mathbf{M}} \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=1}^T d_t \right\| \right]$$

where the supremum is over distributions  $\mathbf{M}$  of martingale difference sequences  $\{d_t\}_{t \in \mathbb{N}}$  such that each  $d_t \in \text{conv}(\mathcal{H} \cup -\mathcal{H})$ .

The second term in Eq. (17) is

$$\begin{aligned}
 & \sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \mathbb{E}_{\substack{f_{1:T} \sim q_{1:T} \\ x_{1:T} \sim p_{1:T}}} \mathbf{B}(\ell(f_1, x_1), \dots, \ell(f_T, x_T)) \\
 &= \sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \mathbb{E}_{\substack{f_{1:T} \sim q_{1:T} \\ x_{1:T} \sim p_{1:T}}} \inf_{c \in S} \left\| c - \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) \right\| \\
 &\leq \sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \mathbb{E}_{\substack{f_{1:T} \sim q_{1:T} \\ x_{1:T} \sim p_{1:T}}} \inf_{c \in S} \left\{ \left\| c - \frac{1}{T} \sum_{t=1}^T \ell(q_t, p_t) \right\| + \left\| \frac{1}{T} \sum_{t=1}^T \ell(q_t, p_t) - \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) \right\| \right\} \\
 &\leq \sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \left\{ \inf_{c \in S} \left\| c - \frac{1}{T} \sum_{t=1}^T \ell(q_t, p_t) \right\| + \mathbb{E}_{\substack{f_{1:T} \sim q_{1:T} \\ x_{1:T} \sim p_{1:T}}} \left\| \frac{1}{T} \sum_{t=1}^T \ell(q_t, p_t) - \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) \right\| \right\} \\
 &\leq \sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \left\{ \inf_{c \in S} \left\| c - \frac{1}{T} \sum_{t=1}^T \ell(q_t, p_t) \right\| \right\} \\
 &\quad + \sup_{p_1, q_1} \dots \sup_{p_T, q_T} \mathbb{E}_{\substack{f_{1:T} \sim q_{1:T} \\ x_{1:T} \sim p_{1:T}}} \left\| \frac{1}{T} \sum_{t=1}^T \ell(q_t, p_t) - \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) \right\| \tag{18}
 \end{aligned}$$

where the last inequality uses the fact that supremum is convex and infimum satisfies the following property:  $\inf_a [C_1(a) + C_2(a)] \leq [\inf_a C_1(a)] + [\sup_a C_2(a)]$ . By one shot approachability assumption, we can choose a particular response  $q_t$  (in the first term of Eq. (18)) for a given  $p_t$  to be the mixed strategy that satisfies  $\ell(q_t, p_t) \in S$ . Since  $S$  is a convex set, we conclude that

$$\frac{1}{T} \sum_{t=1}^T \ell(q_t, p_t) \in S$$

and the first term in Eq. (18) is zero. The second term is trivially upper bounded as

$$\begin{aligned}
 & \sup_{p_1, q_1} \dots \sup_{p_T, q_T} \mathbb{E}_{\substack{f_{1:T} \sim q_{1:T} \\ x_{1:T} \sim p_{1:T}}} \left\| \frac{1}{T} \sum_{t=1}^T \ell(q_t, p_t) - \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) \right\| \\
 &\leq \sup_{p_1, q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T, q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \left\| \frac{1}{T} \sum_{t=1}^T \ell(q_t, p_t) - \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) \right\| \\
 &\leq 2 \sup_{\mathbf{M}} \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=1}^T d_t \right\| \right].
 \end{aligned}$$

Combining the two upper bounds yields the desired result. ■

**Proof [of Theorem 17]** Consider the game where adversary plays from set  $\mathcal{X} = \mathcal{H}$ , the player plays from set  $\mathcal{F} = \{\pm 1\}$ , and  $S = \{0\}$ . Suppose the payoff is given by  $\ell(f, x) = f \cdot x$ . This game is clearly one-shot approachable since the player can always play  $\pm 1$  with equal probability to ensure that  $\ell(p, q) = 0$  irrespective of  $q$ .

Now consider the adversary strategy where adversary fixes a  $\mathcal{H}$  valued tree  $\mathbf{x}$  and at each time  $t$  picks a random  $\epsilon_t \in \{\pm 1\}$  and plays  $x_t = \epsilon_t \mathbf{x}_t(f_1 \cdot \epsilon_1, \dots, f_{t-1} \cdot \epsilon_{t-1})$  that is a random sign multiplied with the instance given by the path on the tree specified by  $f_1 \cdot \epsilon_1, \dots, f_{t-1} \cdot \epsilon_{t-1}$ . Further note that since  $\epsilon_t \in \{\pm 1\}$  are Rademacher random variables, we see that irrespective of choice of distribution from which  $f_t$  is drawn,  $f_t \cdot \epsilon_t$  is a Rademacher random variable conditioned on history. This shows that for the above prescribed adversary strategy, we have that for any  $\mathcal{X}$  valued tree  $\mathbf{x}$  and any two player strategies  $\{q_t^*\}$  and  $\{\bar{q}_t^*\}$  we have

$$\begin{aligned} & \mathbb{E}_{\substack{f_1 \sim q_1^* \\ \epsilon_1 \sim \text{Unif}\{\pm 1\}}} \cdots \mathbb{E}_{\substack{f_T \sim q_T^* \\ \epsilon_T \sim \text{Unif}\{\pm 1\}}} \left\| \frac{1}{T} \sum_{t=1}^T (f_t \cdot \epsilon_t) \mathbf{x}(f_1 \cdot \epsilon_1, \dots, f_{t-1} \cdot \epsilon_{t-1}) \right\| \\ &= \mathbb{E}_{\substack{f_1 \sim q_1^* \\ \epsilon_1 \sim \text{Unif}\{\pm 1\}}} \cdots \mathbb{E}_{\substack{f_T \sim \bar{q}_T^* \\ \epsilon_T \sim \text{Unif}\{\pm 1\}}} \left\| \frac{1}{T} \sum_{t=1}^T (f_t \cdot \epsilon_t) \mathbf{x}(f_1 \cdot \epsilon_1, \dots, f_{t-1} \cdot \epsilon_{t-1}) \right\| \\ &= \mathbb{E}_{\substack{f_1 \sim q_1^* \\ \epsilon_1 \sim \text{Unif}\{\pm 1\}}} \cdots \mathbb{E}_{\substack{f_{T-1} \sim q_{T-1}^* \\ \epsilon_{T-1} \sim \text{Unif}\{\pm 1\}}} \mathbb{E}_{\substack{f_T \sim \bar{q}_T^* \\ \epsilon_T \sim \text{Unif}\{\pm 1\}}} \left\| \frac{1}{T} \sum_{t=1}^T (f_t \cdot \epsilon_t) \mathbf{x}(f_1 \cdot \epsilon_1, \dots, f_{t-1} \cdot \epsilon_{t-1}) \right\| \\ \dots &= \mathbb{E}_{\substack{f_1 \sim \bar{q}_1^* \\ \epsilon_1 \sim \text{Unif}\{\pm 1\}}} \cdots \mathbb{E}_{\substack{f_T \sim \bar{q}_T^* \\ \epsilon_T \sim \text{Unif}\{\pm 1\}}} \left\| \frac{1}{T} \sum_{t=1}^T (f_t \cdot \epsilon_t) \mathbf{x}(f_1 \cdot \epsilon_1, \dots, f_{t-1} \cdot \epsilon_{t-1}) \right\| \end{aligned}$$

The first equality above is due to the fact that  $f_T \cdot \epsilon_T$  is a Rademacher random variable conditioned on  $f_1, \dots, f_{T-1}$  and  $\epsilon_1, \dots, \epsilon_{T-1}$  which means we can replace  $q_T^*$  with  $\bar{q}_T^*$ . The subsequent equalities are got similarly by replacing each  $q_t^*$  by  $\bar{q}_t^*$  one by one inside out by conditioning on  $f_1, \dots, f_{t-1}$  and  $\epsilon_1, \dots, \epsilon_{t-1}$ ; and replacing each  $q_t^*$  by  $\bar{q}_t^*$ . Hence we see that the adversary strategy is an equalizer strategy. Hence using Proposition 20 and picking the fixed  $f = 1$  we see that

$$\mathcal{V}_T \geq \sup_{\mathbf{x}} \mathbb{E}_{\epsilon \sim \text{Unif}\{\pm 1\}^T} \left[ \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \mathbf{x}(\epsilon) \right\| \right] \geq \frac{1}{2} \sup_{\mathbf{M}} \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=1}^T d_t \right\| \right]$$

where the last inequality is because the worst-case martingale difference sequence generated by random signs (Walsh Paley martingales) are lower bounded by the worst case martingale difference sequences within a factor of at most two Pisier (1975).  $\blacksquare$

## Appendix D. Proofs for Calibration (Section 4.3)

Let  $\delta > 0$  to be determined later. Let  $\|\cdot\|$  denote the  $\ell_1$  norm. Let  $C_\delta$  be the maximal  $2\delta$ -packing of  $\Delta(\mathcal{X})$  in this norm. Consider the calibration game defined in Example 4, augmented with the restriction that the player's choice belongs to  $C_\delta$  instead of  $\Delta(k)$ . The corresponding minimax expression with this restriction is clearly an upper bound on the value of the game defined in Example 4.

Observe that the first term in the Triplex Inequality of Theorem 2 is zero. The second term is upper bounded by a particular (sub)optimal response  $q_t$  being the point mass on  $p_t^\delta$ , the element of  $C_\delta$  closest to  $p_t$ . Note that any  $2\delta$  packing is also a  $2\delta$  cover. Thus, the second term becomes

$$\begin{aligned} & \sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \sup_{\phi \in \Phi_T} \left[ - \mathbb{E}_{\substack{x_{1:T} \sim p_{1:T} \\ f_{1:T} \sim q_{1:T}}} \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T)) \right] \\ &= \sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \sup_{\lambda > 0} \sup_{p \in \Delta(k)} \mathbb{E}_{\substack{x_{1:T} \sim p_{1:T} \\ f_{1:T} \sim q_{1:T}}} \left\| \frac{1}{T} \sum_{t=1}^T \ell_{\phi_p, \lambda}(f_t, x_t) \right\| \\ &\leq \sup_{p_1} \dots \sup_{p_T} \sup_{\lambda > 0} \sup_{p \in \Delta(k)} \mathbb{E}_{\substack{x_{1:T} \sim p_{1:T}}} \left\| \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{\|p_t^\delta - p\| \leq \lambda\} \cdot (p_t^\delta - x_t) \right\| \end{aligned}$$

which, in turn, is upper bounded via triangle inequality by

$$\begin{aligned} & \sup_{p_1} \dots \sup_{p_T} \sup_{\lambda > 0} \sup_{p \in \Delta(k)} \mathbb{E}_{\substack{x_{1:T} \sim p_{1:T}}} \left\| \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{\|p_t^\delta - p\| \leq \lambda\} \cdot (p_t^\delta - p_t) \right\| \\ &+ \sup_{p_1} \dots \sup_{p_T} \sup_{\lambda > 0} \sup_{p \in \Delta(k)} \mathbb{E}_{\substack{x_{1:T} \sim p_{1:T}}} \left\| \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{\|p_t^\delta - p\| \leq \lambda\} \cdot (p_t - x_t) \right\| \\ &\leq 2\delta + \sup_{p_1} \dots \sup_{p_T} \sup_{\lambda > 0} \sup_{p \in \Delta(k)} \mathbb{E}_{\substack{x_{1:T} \sim p_{1:T}}} \left\| \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{\|p_t^\delta - p\| \leq \lambda\} \cdot (p_t - x_t) \right\| \end{aligned}$$

Now note that for a given  $\lambda > 0$ ,  $p_1, \dots, p_T$  and  $p \in \Delta(k)$ , we have that  $\{\mathbf{1}\{\|p_t^\delta - p\| \leq \lambda\} \cdot (p_t - x_t)\}_{t \in \mathbb{N}}$  is a martingale difference sequence and so the second term in the triplex inequality is bounded as :

$$\sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \sup_{\phi \in \Phi_T} \left[ - \mathbb{E}_{\substack{x_{1:T} \sim p_{1:T} \\ f_{1:T} \sim q_{1:T}}} \mathbf{B}(\ell_{\phi_1}(f_1, x_1), \dots, \ell_{\phi_T}(f_T, x_T)) \right] \leq 2\delta + 2\sqrt{\frac{k}{T}}. \quad (19)$$

We now proceed to upper bounded the third term in the Triplex Inequality. Since  $-\mathbf{B}$  is a subadditive, by Theorem 3, we have that the third term is bounded by twice the sequential complexity

$$\begin{aligned} 2\mathfrak{R}_T(\ell, \Phi_T, -\mathbf{B}) &= 2 \sup_{\mathbf{f}, \mathbf{x}} \mathbb{E}_{\epsilon_{1:T}} \sup_{\phi \in \Phi_T} -\mathbf{B}\left(\epsilon_1 \ell_{\phi_1}(\mathbf{f}_1(\epsilon), \mathbf{x}_1(\epsilon)), \dots, \epsilon_T \ell_{\phi_T}(\mathbf{f}_T(\epsilon), \mathbf{x}_T(\epsilon))\right) \\ &= 2 \sup_{\mathbf{f}, \mathbf{x}} \mathbb{E}_{\epsilon_{1:T}} \sup_{\lambda > 0} \sup_{p \in \Delta(k)} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \mathbf{1}\{\|\mathbf{f}_t(\epsilon) - p\| \leq \lambda\} \cdot (\mathbf{f}_t(\epsilon) - \mathbf{x}_t(\epsilon)) \right\| \end{aligned}$$

where  $\mathbf{f}$  is a  $C_\delta$ -valued tree. Using the fact that  $\mathbf{f}$  is a discrete-valued tree, not a  $\Delta(k)$ -valued tree, we would like to pass from the supremum over  $\lambda > 0$  and  $p \in \Delta(k)$  to a supremum over finite discrete set in order to appeal to Lemma 4.

To this end, fix  $\mathbf{f}, \mathbf{x}$  and  $\epsilon_{1:T}$  and let us see how many genuinely different functions can we get by varying  $\lambda > 0$  and  $p \in \Delta(k)$ . This question boils down to looking at the size of the class

$$\mathcal{G} := \{g_{p,\lambda}(f) = \mathbf{1}\{\|f - p\| \leq \lambda\} : p \in \Delta(k), \lambda > 0\}$$

over the possible values of  $f \in C_\delta$ . Indeed, if  $g_{p,\lambda}(f) = g_{p',\lambda'}(f)$  for all  $f \in C_\delta$ , then

$$\frac{1}{T} \sum_{t=1}^T \mathbf{1}\{\|\mathbf{f}_t(\epsilon) - p\| \leq \lambda\} \cdot (\mathbf{f}_t(\epsilon) - \mathbf{x}_t(\epsilon)) = \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{\|\mathbf{f}_t(\epsilon) - p'\| \leq \lambda'\} \cdot (\mathbf{f}_t(\epsilon) - \mathbf{x}_t(\epsilon)).$$

We appeal to VC theory for bounding the size of  $\mathcal{G}$  over  $C_\delta$ . First, we claim that the VC dimension of  $\mathcal{G}$  is  $O(k^2)$ . Note that  $\mathcal{G}$  is the class of indicators over  $\ell_1$  balls of radius  $\lambda$  centered at  $p$  for various values of  $p, \lambda$ . A result of Goldberg and Jerrum Goldberg and Jerrum (1995) states that for a class  $\mathcal{G}$  of functions parametrized by a vector of length  $d$ , if for  $g \in \mathcal{G}$  and  $f \in \mathcal{F}$ ,  $\mathbf{1}\{g(f) = 1\}$  can be computed using  $m$  arithmetic operations, the VC dimension of  $\mathcal{G}$  is  $O(md)$ . In our case, the functions in  $\mathcal{G}$  are parametrized by  $k$  values and membership  $\|f - p\|_1 \leq \lambda$  can be established in  $O(k)$  operations. This yields  $O(k^2)$  bound on the VC dimension of  $\mathcal{G}$ . By Sauer-Shelah Lemma, the number of different labelings of the set  $C_\delta$  by  $\mathcal{G}$  is bounded by  $|C_\delta|^{c \cdot k^2}$  for some absolute constant  $c$ . We conclude that the effective number of different  $(p, \lambda)$  is finite. Let us remark that the VC upper bound is *not* used in place of the sequential Littlestone's dimension. It is only used to show that the set  $\Phi_T$  is finite, and such technique can be useful when the set of player's actions is finite.

Hence, there exists a finite set  $S$  of pairs  $(\lambda, p)$  with cardinality  $|S| \leq |C_\delta|^{c \cdot k^2}$  such that

$$\begin{aligned} 2\mathfrak{R}_T(\ell, \Phi_T, -\mathbf{B}) &\leq 2 \sup_{\mathbf{f}, \mathbf{x}} \mathbb{E}_{\epsilon_{1:T}} \sup_{\lambda > 0} \sup_{p \in \Delta(k)} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \mathbf{1}\{\|\mathbf{f}_t(\epsilon) - p\|_1 \leq \lambda\} \cdot (\mathbf{f}_t(\epsilon) - \mathbf{x}_t(\epsilon)) \right\|_1 \\ &= 2 \sup_{\mathbf{f}, \mathbf{x}} \mathbb{E}_{\epsilon_{1:T}} \max_{(p, \lambda) \in S} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \mathbf{1}\{\|\mathbf{f}_t(\epsilon) - p\|_1 \leq \lambda\} \cdot (\mathbf{f}_t(\epsilon) - \mathbf{x}_t(\epsilon)) \right\|_1 \\ &\leq 2 k^{1/2} \sup_{\mathbf{f}, \mathbf{x}} \mathbb{E}_\epsilon \max_{(p, \lambda) \in S} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \mathbf{1}\{\|\mathbf{f}_t(\epsilon) - p\|_1 \leq \lambda\} \cdot (\mathbf{f}_t(\epsilon) - \mathbf{x}_t(\epsilon)) \right\|_2 \end{aligned}$$

Now note that  $\|\cdot\|_2^2$  is  $(2, 2)$ -smooth and so applying Lemma 4 with  $G = \|\cdot\|_2$ ,  $\gamma = 2$ ,  $\eta = 2$ , we see that

$$\begin{aligned} 2\mathfrak{R}_T(\ell, \Phi_T, -\mathbf{B}) &\leq 2k^{1/2} \left( \frac{8 \log(2|S|)}{T} \right)^{1/2} \\ &\leq 2k^{1/2} \left( \frac{16ck^2 \log(|C_\delta|)}{T} \right)^{1/2} \\ &= c'k^{3/2} \left( \frac{\log(|C_\delta|)}{T} \right)^{1/2} \end{aligned}$$

for some small absolute constant  $c'$ .

Now note that the size of set  $C_\delta$  the  $2\delta$  packing of  $\Delta(k)$  is upper bounded by the size of the minimal  $\delta$  cover of  $\Delta(k)$  which can be bounded as  $|C_\delta| \leq (\frac{1}{\delta})^{k-1}$  and so we see that

$$2\mathfrak{R}_T(\ell, \Phi_T, -\mathbf{B}) \leq c' k^2 \left( \frac{\log(1/\delta)}{T} \right)^{1/2}.$$

Combining the above upper bound on the third term of triplex inequality and Equation 19 that bounds the second term of the triplex inequality (and since first term is anyway 0) we see that,

$$\mathcal{V}_T \leq 2\delta + 2\sqrt{\frac{k}{T}} + c' k^2 \left( \frac{\log(1/\delta)}{T} \right)^{1/2}.$$

Choosing  $\delta = 1/T$  concludes the proof.

## Appendix E. Proofs for Global Cost (Section 4.4)

**Proof [of Theorem 18]** The Triplex Inequality and Theorem 3 give

$$\begin{aligned} \mathcal{V}_T &\leq \sup_{p_1, q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ x_1 \sim p_1}} \dots \sup_{p_T, q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ x_T \sim p_T}} \mathbb{E}_{\substack{f'_1 \sim p_1 \\ x'_1 \sim p_1}} \left\| \frac{1}{T} \sum_{t=1}^T (\ell(f_t, x_t) - \ell(f'_t, x'_t)) \right\| \\ &\quad + \sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \sup_{f \in \mathcal{F}} \mathbb{E}_{\substack{f_{1:T} \sim q_{1:T} \\ x_{1:T} \sim p_{1:T}}} \left\{ \left\| \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) \right\| - \left\| \frac{1}{T} \sum_{t=1}^T \ell(f, x_t) \right\| \right\} \\ &\quad + 2 \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{1:T}} \sup_{f \in \mathcal{F}} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \ell(f, \mathbf{x}_t(\epsilon)) \right\|. \end{aligned}$$

Consider the first term in the Triplex Inequality. Observe that  $(\ell(f_t, x_t) - \ell(f'_t, x'_t))_{t=1}^T$  is a (vector valued) martingale difference sequence and so

$$\sup_{p_1, q_1} \mathbb{E}_{\substack{f_1, f'_1 \sim q_1 \\ x_1, x'_1 \sim p_1}} \dots \sup_{p_T, q_T} \mathbb{E}_{\substack{f_T, f'_T \sim q_T \\ x_T, x'_T \sim p_T}} \left\| \frac{1}{T} \sum_{t=1}^T (\ell(f_t, x_t) - \ell(f'_t, x'_t)) \right\| \leq 2 \sup_{\mathbf{M}} \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=1}^T d_t \right\| \right].$$

where the supremum is over distributions  $\mathbf{M}$  of martingale difference sequences  $\{d_t\}_{t \in \mathbb{N}}$  such that each  $d_t \in \text{conv}(\mathcal{H} \cup -\mathcal{H})$ .

Now, consider the second summand above:

$$\begin{aligned} &\sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \sup_{f \in \mathcal{F}} \mathbb{E}_{\substack{f_{1:T} \sim q_{1:T} \\ x_{1:T} \sim p_{1:T}}} \left\{ \left\| \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) \right\| - \left\| \frac{1}{T} \sum_{t=1}^T \ell(f, x_t) \right\| \right\} \\ &= \sup_{p_1} \inf_{q_1} \dots \sup_{p_T} \inf_{q_T} \left\{ \mathbb{E}_{\substack{f_{1:T} \sim q_{1:T} \\ x_{1:T} \sim p_{1:T}}} \left\| \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) \right\| - \inf_{f \in \mathcal{F}} \mathbb{E}_{x_{1:T} \sim p_{1:T}} \left\| \frac{1}{T} \sum_{t=1}^T \ell(f, x_t) \right\| \right\} \\ &\leq \sup_{p_1} \dots \sup_{p_T} \left\{ \mathbb{E}_{x_{1:T} \sim p_{1:T}} \left\| \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) \right\| - \inf_{f \in \mathcal{F}} \mathbb{E}_{x_{1:T} \sim p_{1:T}} \left\| \frac{1}{T} \sum_{t=1}^T \ell(f, x_t) \right\| \right\} \end{aligned}$$

where in the last step a (sub)optimal choice was made for  $q_t$ : the distribution  $q_t = \delta_{f_t}$  puts all the mass on  $f_t$  such that

$$\|\ell(f_t, p_t)\| = \inf_{f \in \mathcal{F}} \|\ell(f, p_t)\|.$$

Observe that by several applications of triangle and Jensen's inequalities,

$$\begin{aligned} & \mathbb{E}_{x_1:T \sim p_{1:T}} \left\| \frac{1}{T} \sum_{t=1}^T \ell(f_t, x_t) \right\| - \inf_{f \in \mathcal{F}} \mathbb{E}_{x_1:T \sim p_{1:T}} \left\| \frac{1}{T} \sum_{t=1}^T \ell(f, x_t) \right\| \\ & \leq \left\{ \left\| \frac{1}{T} \sum_{t=1}^T \ell(f_t, p_t) \right\| - \inf_{f \in \mathcal{F}} \left\| \frac{1}{T} \sum_{t=1}^T \ell(f, p_t) \right\| \right\} + \mathbb{E}_{x_1:T \sim p_{1:T}} \left\| \frac{1}{T} \sum_{t=1}^T (\ell(f_t, x_t) - \ell(f_t, p_t)) \right\| \end{aligned} \quad (20)$$

Under Assumption 1, along with the way we chose  $f_t$ , the first term in (20) becomes

$$\left\| \frac{1}{T} \sum_{t=1}^T \ell(f_t, p_t) \right\| - \inf_{f \in \mathcal{F}} \left\| \frac{1}{T} \sum_{t=1}^T \ell(f, p_t) \right\| \leq \frac{1}{T} \sum_{t=1}^T \|\ell(f_t, p_t)\| - \frac{1}{T} \sum_{t=1}^T \inf_{f \in \mathcal{F}} \|\ell(f, p_t)\| = 0.$$

We conclude that the second term in the Triplex Inequality can be upper bounded by

$$\sup_{p_1} \dots \sup_{p_T} \mathbb{E}_{x_1:T \sim p_{1:T}} \left\| \frac{1}{T} \sum_{t=1}^T (\ell(f_t, x_t) - \ell(f_t, p_t)) \right\|,$$

which, in turn, is no worse than the supremum over distributions  $\mathbf{M}$  of martingale difference sequences used to bound the first term.

This gives us the general upper bound on the value of the game:

$$\mathcal{V}_T \leq 4 \sup_{\mathbf{M}} \mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=1}^T d_t \right\| \right] + 2 \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_1:T} \sup_{f \in \mathcal{F}} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \ell(f, \mathbf{x}_t(\epsilon)) \right\|. \quad (21)$$

■

**Lemma 22** *Let  $\mathcal{F}$  be the probability simplex in any dimension. Let  $\|\cdot\|$  be any norm. The function*

$$x \mapsto \inf_{f \in \mathcal{F}} \|f \odot x\|,$$

*defined on the positive orthant, is concave.*

**Proof** Since the function above is absolutely homogeneous and continuous, all we need to prove is

$$\inf_{f \in \mathcal{F}} \|f \odot (x + y)\| \geq \inf_{f \in \mathcal{F}} \|f \odot x\| + \inf_{f \in \mathcal{F}} \|f \odot y\|.$$

for arbitrary  $x, y$ . That is, for arbitrary  $f', x, y$ ,

$$\|f' \odot (x + y)\| \geq \inf_{f \in \mathcal{F}} \|f \odot x\| + \inf_{f \in \mathcal{F}} \|f \odot y\|.$$

Define  $h, g \in \mathcal{F}$  as follows:

$$g_i = \frac{f'_i(1 + y_i/x_i)}{Z_g} \quad h_i = \frac{f'_i(1 + x_i/y_i)}{Z_h} ,$$

where

$$Z_g = \sum_i f'_i(1 + y_i/x_i) \quad Z_h = \sum_i f'_i(1 + x_i/y_i) .$$

Now, as we show below,  $1/Z_g + 1/Z_h \leq 1$ . Thus,

$$\begin{aligned} \|f' \odot (x + y)\| &\geq \frac{1}{Z_g} \|f' \odot (x + y)\| + \frac{1}{Z_h} \|f' \odot (x + y)\| \\ &= \|g \odot x\| + \|h \odot y\| \\ &\geq \inf_{f \in \mathcal{F}} \|f \odot x\| + \inf_{f \in \mathcal{F}} \|f \odot y\| . \end{aligned}$$

To finish the proof, note that, by Cauchy-Schwarz,

$$\left( \sum_i f'_i(1 + y_i/x_i) \right) \cdot \left( \sum_i f'_i \frac{x_i}{x_i + y_i} \right) \geq \left( \sum_i f'_i \right)^2 = 1 .$$

This shows,

$$\frac{1}{Z_g} \leq \sum_i f'_i \frac{x_i}{x_i + y_i} .$$

Similarly, we get

$$\frac{1}{Z_h} \leq \sum_i f'_i \frac{y_i}{x_i + y_i} .$$

Adding them, we get

$$\frac{1}{Z_g} + \frac{1}{Z_h} \leq \sum_i f'_i = 1$$

as claimed. This completes the proof. ■

# Neyman-Pearson classification under a strict constraint

**Philippe Rigollet**

*Operations Research  
& Financial Engineering  
Princeton University  
Princeton, NJ 08544 USA*

RIGOLLET@PRINCETON.EDU

**Xin Tong**

*Operations Research  
& Financial Engineering  
Princeton University  
Princeton, NJ 08544 USA*

XTONG@PRINCETON.EDU

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

Motivated by problems of anomaly detection, this paper implements the Neyman-Pearson paradigm to deal with asymmetric errors in binary classification with a convex loss. Given a finite collection of classifiers, we combine them and obtain a new classifier that satisfies simultaneously the two following properties with high probability: (i), its probability of type I error is below a pre-specified level and (ii), it has probability of type II error close to the minimum possible. The proposed classifier is obtained by minimizing an empirical objective subject to an empirical constraint. The novelty of the method is that the classifier output by this problem is shown to satisfy the original constraint on type I error. This strict enforcement of the constraint has interesting consequences on the control of the type II error and we develop new techniques to handle this situation. Finally, connections with chance constrained optimization are evident and are investigated.

**keywords:** binary classification, Neyman-Pearson paradigm, anomaly detection, empirical constraint, empirical risk minimization, chance constrained optimization.

## 1. Introduction

The Neyman-Pearson (NP) paradigm in statistical learning extends the objective of classical binary classification in that, while the latter focuses on minimizing classification error that is a weighted sum of type I and type II errors, where the weighting is proportional to the class priors , the former minimizes type II error with an upper bound  $\alpha$  on type I error. With slight abuse of language, in verbal discussion we do not distinguish type I/ II error from probability of type I/ II error. Motivations for the NP approach come from many practical problems, where the importance of type I error differs from that of type II error. Typical examples include medical diagnosis or anomaly detection.

In the learning context, as true errors are inaccessible, we cannot enforce almost surely the desired upper bound for type I error. The best we can hope is that a data dependent

classifier has type I error bounded with high probability. Henceforth, there are two goals in this project. The first is to design a learning procedure so that type I error of the learned classifier  $\hat{f}$  is upper bounded by a pre-specified level with pre-specified high probability; the second is to show that  $\hat{f}$  has good performance bounds for excess type II error.

This paper is organized as follows. In Section 2, the classical setup for binary classification is reviewed and the main notation is introduced. A parallel between binary classification and hypothesis testing is drawn in Section 3 with emphasis on the NP paradigm in both frameworks. The main propositions and theorems are stated in Section 4. Finally different extensions of the main results to a different sampling scheme and to chance constrained optimization are presented in Section 5. The proofs of the main results are gathered in Section 6.

In the rest of the paper, we denote by  $x_j$  the  $j$ -th coordinate of a vector  $x \in \mathbb{R}^d$ .

## 2. Binary classification

### 2.1. Classification risk and classifiers

Let  $(X, Y)$  be a random couple where  $X \in \mathcal{X} \subset \mathbb{R}^d$  is a vector of covariates and  $Y \in \{-1, 1\}$  is a label that indicates to which class  $X$  belongs. A *classifier*  $h$  is a mapping  $h : \mathcal{X} \rightarrow [-1, 1]$  whose sign returns the predicted class given  $X$ . An error occurs when  $-h(X)Y \geq 0$  and it is therefore natural to define the classification loss by  $\mathbb{I}(-h(X)Y \geq 0)$ , where  $\mathbb{I}(\cdot)$  denotes the indicator function.

The expectation of the classification loss with respect to the joint distribution of  $(X, Y)$  is called (*classification*) *risk* and is defined by

$$R(h) = \mathbb{P}(-h(X)Y \geq 0).$$

Clearly, the indicator function is not convex and for computation, a common practice is to replace it by a convex surrogate (see, e.g. Bartlett et al., 2006, and references therein).

To this end, we rewrite the risk function as

$$R(h) = \mathbb{E}[\varphi(-h(X)Y)], \quad (2.1)$$

where  $\varphi(z) = \mathbb{I}(z \geq 0)$ . Convex relaxation can be achieved by simply replacing the indicator function by a convex surrogate.

**Definition 2.1** A function  $\varphi : [-1, 1] \rightarrow \mathbb{R}^+$  is called a *convex surrogate* if it is non-decreasing, continuous and convex and if  $\varphi(0) = 1$ .

Commonly used examples of convex surrogates are the hinge loss  $\varphi(x) = (1 + x)_+$ , the logit loss  $\varphi(x) = \log_2(1 + e^x)$  and the exponential loss  $\varphi(x) = e^x$ .

For a given choice of  $\varphi$ , define the  $\varphi$ -risk

$$R_\varphi(h) = \mathbb{E}[\varphi(-Yh(X))].$$

Hereafter, we assume that  $\varphi$  is fixed and refer to  $R_\varphi$  as the risk. In our subsequent analysis, this convex relaxation will also be the ground to analyze a stochastic convex optimization problem subject to stochastic constraints. A general treatment of such problems can be found in subsection 5.2.

Because of overfitting, it is unreasonable to look for mappings minimizing empirical risk over all classifiers. Indeed, one could have a small empirical risk but a large true risk. Hence, we resort to regularization. There are in general two ways to proceed. The first is to restrict the candidate classifiers to a specific class  $\mathcal{H}$ , and the second is to change the objective function by, for example, adding a penalty term. The two approaches can be combined, and sometimes are obviously equivalent.

In this paper, we pursue the first idea by defining the class of candidate classifiers as follows. Let  $h_1, \dots, h_M, M \geq 2$  be a given collection of classifiers. In our setup, we allow  $M$  to be large. In particular, our results remain asymptotically meaningful as long as  $M = o(e^n)$ . Such classifiers are usually called base classifiers and can be constructed in a very naive manner. Typical examples include decision stumps or small trees. While the  $h_j$ 's may have no satisfactory classifying power individually, for over two decades, boosting type of algorithms have successfully exploited the idea that a suitable weighted majority vote among these classifiers may result in low classification risk (Schapire, 1990). Consequently, we restrict our search for classifiers to the set of functions consisting of convex combinations of the  $h_j$ 's:

$$\mathcal{H}^{\text{conv}} = \{h_\lambda = \sum_{j=1}^M \lambda_j h_j, \lambda \in \Lambda\},$$

where  $\Lambda$  denotes the flat simplex of  $\mathbb{R}^M$  and is defined by  $\Lambda = \{\lambda \in \mathbb{R}^M : \lambda_j \geq 0, \sum_{j=1}^M \lambda_j = 1\}$ . In effect, classification rules given by the sign of  $h \in \mathcal{H}^{\text{conv}}$  are exactly the set of rules produced by the weighted majority votes among the base classifiers  $h_1, \dots, h_M$ .

By restricting our search to classifiers in  $\mathcal{H}^{\text{conv}}$ , the best attainable  $\varphi$ -risk is called *oracle risk* and is abusively denoted by  $R_\varphi(\mathcal{H}^{\text{conv}})$ . As a result, we have  $R_\varphi(h) \geq R_\varphi(\mathcal{H}^{\text{conv}})$  for any  $h \in \mathcal{H}^{\text{conv}}$  and a natural measure of performance for a classifier  $h \in \mathcal{H}^{\text{conv}}$  is given by its excess risk defined by  $R_\varphi(h) - R_\varphi(\mathcal{H}^{\text{conv}})$ .

The excess risk of a data driven classifier  $h_n$  is a random quantity and we are interested in bounding it with high probability. Formally, the statistical goal of binary classification is to construct a classifier  $h_n$  such that the oracle inequality

$$R_\varphi(h_n) \leq R_\varphi(h_{\mathcal{H}^{\text{conv}}}) + \Delta_n(\mathcal{H}^{\text{conv}}, \delta) \quad (2.2)$$

holds with probability  $1 - \delta$ , where  $\Delta_n(\cdot, \cdot)$  should be as small as possible.

In the scope of this paper, we focus on candidate classifiers in the class  $\mathcal{H}^{\text{conv}}$ . Some of the following results such as Theorem 4.1 can be extended to more general classes of classifiers with known complexity such as classes with bounded VC-dimension, as for example in Cannon et al. (2002). However, our main argument for bounding type II error relies on Proposition 4.1 which, in turn, depends heavily on the convexity of the problem, and it is not clear how it can be extended to more general classes of classifiers.

## 2.2. The Neyman-Pearson paradigm

In classical binary classification, the risk function can be expressed as a convex combination of type I error  $R^-(h) = \mathbb{P}(-Yh(X) \geq 0 | Y = -1)$  and of type II error  $R^+(h) =$

$\mathbb{P}(-Yh(X) \geq 0 | Y = 1)$ :

$$R(h) = \mathbb{P}(Y = -1)R^-(h) + \mathbb{P}(Y = 1)R^+(h).$$

More generally, we can define the  $\varphi$ -type I and  $\varphi$ -type II errors respectively by

$$R_\varphi^-(h) = \mathbb{E}[\varphi(-Yh(X)) | Y = -1] \quad \text{and} \quad R_\varphi^+(h) = \mathbb{E}[\varphi(-Yh(X)) | Y = 1].$$

Following the NP paradigm, for a given class  $\mathcal{H}$  of classifiers, we seek to solve the constrained minimization problem:

$$\min_{\substack{h \in \mathcal{H} \\ R_\varphi^-(h) \leq \alpha}} R_\varphi^+(h), \quad (2.3)$$

where  $\alpha \in (0, 1)$ , the significance level, is a constant specified by the user.

NP classification is closely related to the NP approach to statistical hypothesis testing. We now recall a few key concepts about the latter. Many classical works have addressed the theory of statistical hypothesis testing, in particular Lehmann and Romano (2005) provides a thorough treatment of the subject.

Statistical hypothesis testing bears strong resemblance with binary classification if we assume the following model. Let  $P^-$  and  $P^+$  be two probability distributions on  $\mathcal{X} \subset \mathbb{R}^d$ . Let  $p \in (0, 1)$  and assume that  $Y \in \{-1, 1\}$  takes value 1 with probability  $p$  and value  $-1$  with probability  $1 - p$ . Assume further that the conditional distribution of  $X$  given  $Y$  is given by  $P^Y$ . Given such a model, the goal of statistical hypothesis testing is to determine whether  $X$  was generated from  $P^-$  or  $P^+$ . To that end, we construct a test  $\phi : \mathcal{X} \rightarrow [0, 1]$  and the conclusion of the test based on  $\phi$  is that  $X$  is generated from  $P^+$  with probability  $\phi(X)$  and from  $P^-$  with probability  $1 - \phi(X)$ . Note that randomness here comes from an exogenous randomization process such as flipping a biased coin. Two kinds of errors arise: type I error occurs when rejecting  $P^-$  when it is true, and type II error occurs when accepting  $P^-$  when it is false. The Neyman-Pearson paradigm in hypothesis testing amounts to choosing  $\phi$  that solves the following constrained optimization problem

$$\begin{aligned} & \text{maximize} && \mathbb{E}[\phi(X) | Y = 1], \\ & \text{subject to} && \mathbb{E}[\phi(X) | Y = -1] \leq \alpha, \end{aligned}$$

where  $\alpha \in (0, 1)$  is the significance level of the test. In other words, we specify a significance level  $\alpha$  on type I error, and minimize type II error. We call a solution to this problem *a most powerful test* of level  $\alpha$ . The Neyman-Pearson Lemma gives mild sufficient conditions for the existence of such a test.

**Theorem 2.1 (Neyman-Pearson Lemma)** *Let  $P^-$  and  $P^+$  be probability distributions possessing densities  $p^-$  and  $p^+$  respectively with respect to some measure  $\mu$ . Let  $\varphi_k(x) = \mathbb{I}(L(x) \geq k)$ , where the likelihood ratio  $L(x) = p^+(x)/p^-(x)$  and  $k$  is such that  $P^-(L(X) > k) \leq \alpha$  and  $P^-(L(X) \geq k) \geq \alpha$ . Then,*

- $\varphi_k$  is a level  $\alpha = \mathbb{E}[\varphi_k(X) | Y = -1]$  most powerful test.

- For a given level  $\alpha$ , the most powerful test of level  $\alpha$  is defined by

$$\phi(X) = \begin{cases} 1 & \text{if } L(X) > k \\ 0 & \text{if } L(X) < k \\ \frac{\alpha - P^-(L(X) > k)}{P^-(L(X) = k)} & \text{if } L(X) = k. \end{cases}$$

Notice that in the learning framework,  $\phi$  cannot be computed since it requires the knowledge of the likelihood ratio and of the distributions  $P^-$  and  $P^+$ . Therefore, it remains merely a theoretical propositions. Nevertheless, the result motivates the NP paradigm pursued here.

### 3. Neyman-Pearson classification via convex optimization

Recall that in NP classification with a convex surrogate  $\varphi$ , the goal is to solve the following optimization problem

$$\min_{\substack{h \in \mathcal{H} \\ R_\varphi^-(h) \leq \alpha}} R_\varphi^+(h). \quad (3.1)$$

This cannot be done directly as conditional distributions  $P^-$  and  $P^+$ , and hence  $R_\varphi^-$  and  $R_\varphi^+$ , are unknown. In statistical applications, information about these distributions is available through two i.i.d. samples  $X_1^-, \dots, X_{n^-}^-, n^- \geq 1$  and  $X_1^+, \dots, X_{n^+}^+, n^+ \geq 1$ , where  $X_i^- \sim P^-, i = 1, \dots, n^-$  and  $X_i^+ \sim P^+, i = 1, \dots, n^+$ . We do not assume that the two samples  $(X_1^-, \dots, X_{n^-}^-)$  and  $(X_1^+, \dots, X_{n^+}^+)$  are mutually independent. Presently the sample sizes  $n^-$  and  $n^+$  are assumed to be deterministic and will appear in the subsequent finite sample bounds. A different sampling scheme, where these quantities are random, is investigated in subsection 5.1.

#### 3.1. Previous results and new input

While the binary classification problem has been extensively studied, theoretical proposition on how to implement the NP paradigm remains scarce. To the best of our knowledge, Cannon et al. (2002) initiated the theoretical treatment of the NP classification paradigm and an early empirical study can be found in Casasent and Chen (2003). The framework of Cannon et al. (2002) is the following. Fix a constant  $\varepsilon_0 > 0$  and let  $\mathcal{H}$  be a given set of classifiers with finite VC dimension. They study a procedure that consists of solving the following relaxed empirical optimization problem

$$\min_{\substack{h \in \mathcal{H} \\ \hat{R}^-(h) \leq \alpha + \varepsilon_0/2}} \hat{R}^+(h), \quad (3.2)$$

where

$$\hat{R}^-(h) = \frac{1}{n^-} \sum_{i=1}^{n^-} \mathbb{I}(h(X_i^-) \geq 0), \quad \text{and} \quad \hat{R}^+(h) = \frac{1}{n^+} \sum_{i=1}^{n^+} \mathbb{I}(h(X_i^+) \leq 0)$$

denote the empirical type I and empirical type II errors respectively. Let  $\hat{h}$  be a solution to (3.2). Denote by  $h^*$  a solution to the original Neyman-Pearson optimization problem:

$$h^* \in \operatorname{argmin}_{\substack{h \in \mathcal{H} \\ R^-(h) \leq \alpha}} R^+(h), \quad (3.3)$$

The main result of Cannon et al. (2002) states that, simultaneously with high probability, the type II error  $R^+(\hat{h})$  is bounded from above by  $R^+(h^*) + \varepsilon_1$ , for some  $\varepsilon_1 > 0$  and the type I error of  $\hat{h}$  is bounded from above by  $\alpha + \varepsilon_0$ . In a later paper, Cannon et al. (2003) consider problem (3.2) for a data-dependent family of classifiers  $\mathcal{H}$ , and bound estimation errors accordingly. Several results for traditional statistical learning such as PAC bounds or oracle inequalities have been studied in Scott (2005) and Scott and Nowak (2005) in the same framework as the one laid down by Cannon et al. (2002). A noteworthy departure from this setup is Scott (2007) where sensible performance measures for NP classification that go beyond analyzing separately two kinds of errors are introduced. Moreover, Corollary 1 in Scott (2007) provides an oracle inequality for the type II error of a classifier that satisfies an strict constraint on the type I error. However, this result is not directly comparable to the present paper since the rate at which the type II error decreases is not explicitly controlled. This drawback is inherent to methods based on empirical risk minimization as opposed to convexified methods as discussed below. Finally, a related work is that of Blanchard et al. (2010) who develop a general solution to semi-supervised novelty detection by reducing it to NP classification. Recently, Han et al. (2008) transposed several results of Cannon et al. (2002) and Scott and Nowak (2005) to NP classification with convex loss.

The present work departs from previous literature in our treatment of type I error. As a matter of fact, the classifiers in all the papers mentioned above can only ensure that  $\mathbb{P}(R^-(\hat{h}) > \alpha + \varepsilon_0)$  is small, for some  $\varepsilon_0 > 0$ . However, it is our primary interest to make sure that  $R^-(\hat{h}) \leq \alpha$  with high probability, following the original principle of the Neyman-Pearson paradigm that type I error should be controlled by a pre-specified level  $\alpha$ . As will be illustrated, to control  $\mathbb{P}(R^-(\hat{h}) > \alpha)$ , it is necessary to have  $\hat{h}$  be a solution to some program with a strengthened constraint on empirical type I error. If our concern is only on type I error, we can just do so. However, we also want to control excess type II error simultaneously.

The difficulty was foreseen in the seminal paper Cannon et al. (2002), where it is claimed without justification that if we use  $\alpha' < \alpha$  for the empirical program, “it seems unlikely that we can control the estimation error  $R^+(\hat{h}) - R^+(h^*)$  in a distribution independent way”. We have analytically confirmed this opinion, but due to limited space we refer the interested reader to the full version of this paper (Rigollet and Tong, 2011).

To overcome this dilemma, we resort to a continuous convex surrogate as our loss function. In particular, we design a modified version of empirical risk minimization method such that the data-driven classifier  $\hat{h}$  has type I error bounded by  $\alpha$  with high probability. Moreover, we consider here a class  $\mathcal{H}$  that allows a different treatment of the empirical processes involved.

This new approach comes with new technical challenges which we summarize here. In the approach of Cannon et al. (2002) and of Scott and Nowak (2005), the relaxed constraint on the type I error is constructed such that the constraint  $\hat{R}^-(h) \leq \alpha + \varepsilon_0/2$  on type I error in (3.2) is satisfied by  $h^*$  with high probability, and that this classifier accommodates excess

type II error well. As a result, the control of type II error mainly follows as a standard exercise to control suprema of empirical processes. This is not the case here; we have to develop methods to control the optimum value of a convex optimization problem under a stochastic constraint. Such methods have consequences not only in NP classification but also on chance constrained optimization as explained in subsection 5.2.

### 3.2. Convexified NP classifier

To solve the problem of NP classification (2.3) where the distribution of the observations is unknown, we resort to empirical risk minimization. In view of the arguments presented in the previous subsection, we cannot simply replace the unknown true risk functions by their empirical counterparts. The treatment of the convex constraint should be done carefully and we proceed as follows.

For any classifier  $h$  and a given convex surrogate  $\varphi$ , define  $\hat{R}_\varphi^-$  and  $\hat{R}_\varphi^+$  to be the empirical counterparts of  $R_\varphi^-$  and  $R_\varphi^+$  respectively by

$$\hat{R}_\varphi^-(h) = \frac{1}{n^-} \sum_{i=1}^{n^-} \varphi(h(X_i^-)), \quad \text{and} \quad \hat{R}_\varphi^+(h) = \frac{1}{n^+} \sum_{i=1}^{n^+} \varphi(-h(X_i^+)).$$

Moreover, for any  $a > 0$ , let  $\mathcal{H}^{\varphi,a} = \{h \in \mathcal{H}^{\text{conv}} : R_\varphi^-(h) \leq a\}$  be the set of classifiers in  $\mathcal{H}^{\text{conv}}$  whose convexified type I errors are bounded from above by  $a$ , and let  $\mathcal{H}_{n^-}^{\varphi,a} = \{h \in \mathcal{H}^{\text{conv}} : \hat{R}_\varphi^-(h) \leq a\}$  be the set of classifiers in  $\mathcal{H}^{\text{conv}}$  whose empirical convexified type I errors are bounded by  $a$ . To make our analysis meaningful, we assume that  $\mathcal{H}^{\varphi,a} \neq \emptyset$ .

We are now in a position to construct a classifier in  $\mathcal{H}^{\text{conv}}$  according to the Neyman-Pearson paradigm. For any  $\tau > 0$  such that  $\tau \leq \alpha\sqrt{n^-}$ , define the convexified NP classifier  $\tilde{h}^\tau$  as any classifier that solves the following optimization problem

$$\min_{\substack{h \in \mathcal{H}^{\text{conv}} \\ \hat{R}_\varphi^-(h) \leq \alpha - \tau/\sqrt{n^-}}} \hat{R}_\varphi^+(h). \quad (3.4)$$

Note that this problem consists of minimizing a convex function subject to a convex constraint and can therefore be solved by standard algorithms such as (see, e.g., Boyd and Vandenberghe, 2004, and references therein). In the next section, we present a series of results on type I and type II errors of classifiers that include  $\tilde{h}^\tau$ .

## 4. Performance Bounds

### 4.1. Control of type I error

The first challenge is to identify classifiers  $h$  such that  $R_\varphi^-(h) \leq \alpha$  with high probability. This is done by enforcing its empirical counterpart  $\hat{R}_\varphi^-(h)$  be bounded from above by the quantity

$$\alpha_\tau = \alpha - \tau/\sqrt{n^-},$$

for a proper choice of positive constant  $\tau$ .

**Theorem 4.1** Fix constants  $\delta, \alpha \in (0, 1), L > 0$  and let  $\varphi : [-1, 1] \rightarrow \mathbb{R}^+$  be a given  $L$ -Lipschitz convex surrogate. Define

$$\tau = 4\sqrt{2}L\sqrt{\log\left(\frac{2M}{\delta}\right)}. \quad (4.1)$$

Then for any classifier  $h \in \mathcal{H}^{\text{conv}}$  that satisfies  $\hat{R}_\varphi^-(h) \leq \alpha_\tau$ , we have

$$R^-(h) \leq R_\varphi^-(h) \leq \alpha,$$

with probability at least  $1 - \delta$ . Equivalently

$$\mathbb{P}[\mathcal{H}_{n^-}^{\varphi, \alpha_\tau} \subset \mathcal{H}^{\varphi, \alpha}] \geq 1 - \delta. \quad (4.2)$$

#### 4.2. Simultaneous control of the two errors

Theorem 4.1 guarantees that any classifier that satisfies the strengthened constraint on the empirical  $\varphi$ -type I error will have  $\varphi$ -type I error and true type I error bounded from above by  $\alpha$ . We now check that the constraint is not too strong so that the type II error is overly deteriorated. Indeed, an extremely small  $\alpha_\tau$  would certainly ensure a good control of type I error but would deteriorate significantly the best achievable type II error. Below, we show not only that this is not the case for our approach but also that the convexified NP classifier  $\tilde{h}^\tau$  defined in subsection 3.2 with  $\tau$  defined in (4.1) suffers only a small degradation of its type II error compared to the best achievable. Analogues to classical binary classification, a desirable result is that with high probability,

$$R_\varphi^+(\tilde{h}^{\alpha_\tau}) - \min_{h \in \mathcal{H}^{\varphi, \alpha}} R_\varphi^+(h) \leq \tilde{\Delta}_n(\mathcal{F}), \quad (4.3)$$

where  $\tilde{\Delta}_n(\mathcal{F})$  goes to 0 as  $n = n^- + n^+ \rightarrow \infty$ .

The following proposition is pivotal to our argument.

**Proposition 4.1** Fix constant  $\alpha \in (0, 1)$  and let  $\varphi : [-1, 1] \rightarrow \mathbb{R}^+$  be a given continuous convex surrogate. Assume further that there exists  $\nu_0 > 0$  such that the set of classifiers  $\mathcal{H}^{\varphi, \alpha - \nu_0}$  is nonempty. Then, for any  $\nu \in (0, \nu_0)$ ,

$$\min_{h \in \mathcal{H}^{\varphi, \alpha - \nu}} R_\varphi^+(h) - \min_{h \in \mathcal{H}^{\varphi, \alpha}} R_\varphi^+(h) \leq \varphi(1) \frac{\nu}{\nu_0 - \nu}.$$

This proposition ensures that if the convex surrogate  $\varphi$  is continuous, strengthening the constraint on type I error does not deteriorate too much the optimal type II error. We should mention that the proof does not use the Lipschitz property of  $\varphi$ , but only that it is uniformly bounded by  $\varphi(1)$  on  $[-1, 1]$ . This proposition has direct consequences on chance constrained optimization as discussed in subsection 5.2.

The next theorem shows that the NP classifier  $\tilde{h}^\tau$  defined in subsection 3.2 is a good candidate to perform classification with the Neyman-Pearson paradigm. It relies on the following assumption which is necessary to verify the condition of Proposition 4.1.

**Assumption 1** There exists a positive constant  $\varepsilon < 1$  such that the set of classifiers  $\mathcal{H}^{\varphi, \varepsilon\alpha}$  is nonempty.

Note that this assumption can be tested using (4.2) for large enough  $n^-$ . Indeed, it follows from this inequality that with probability  $1 - \delta$ ,

$$\mathcal{H}_{n^-}^{\varphi, \varepsilon\alpha - \tau/\sqrt{n^-}} \subset \mathcal{H}^{\varphi, \varepsilon\alpha - \tau/\sqrt{n^-} + \tau/\sqrt{n^-}} = \mathcal{H}^{\varphi, \varepsilon\alpha}.$$

Thus, it is sufficient to check if  $\mathcal{H}_{n^-}^{\varphi, \varepsilon\alpha - \tau/\sqrt{n^-}}$  is nonempty for some  $\varepsilon > 0$ . Before stating our main theorem, we need the following definition. Under Assumption 1, let  $\bar{\varepsilon}$  denote the smallest  $\varepsilon$  such that  $\mathcal{H}^{\varphi, \varepsilon\alpha} \neq \emptyset$  and let  $n_0$  be the smallest integer such that

$$n_0 \geq \left( \frac{4\tau}{(1 - \bar{\varepsilon})\alpha} \right)^2. \quad (4.4)$$

**Theorem 4.2** *Let  $\varphi, \tau, \delta$  and  $\alpha$  be the same as in Theorem 4.1, and  $\tilde{h}^\tau$  denote any solution to (3.4). Moreover, let Assumption 1 hold and assume that  $n^- \geq n_0$  where  $n_0$  is defined in (4.4). Then, the following hold with probability  $1 - 2\delta$ ,*

$$R^-(\tilde{h}^\tau) \leq R_\varphi^-(\tilde{h}^\tau) \leq \alpha \quad (4.5)$$

and

$$R_\varphi^+(\tilde{h}^\tau) - \min_{h \in \mathcal{H}^{\varphi, \alpha}} R_\varphi^+(h) \leq \frac{4\varphi(1)\tau}{(1 - \bar{\varepsilon})\alpha\sqrt{n^-}} + \frac{2\tau}{\sqrt{n^+}}. \quad (4.6)$$

In particular, there exists a constant  $C > 0$  depending on  $\alpha, \varphi(1)$  and  $\bar{\varepsilon}$  such that (4.6) yields

$$R_\varphi^+(\tilde{h}^\tau) - \min_{h \in \mathcal{H}^{\varphi, \alpha}} R_\varphi^+(h) \leq C \left( \sqrt{\frac{\log(2M/\delta)}{n^-}} + \sqrt{\frac{\log(2M/\delta)}{n^+}} \right)$$

Note here that Theorem 4.2 is not exactly of the type (4.3). The right hand side of (4.6) goes to zero if both  $n^-$  and  $n^+$  go to infinity. Moreover, inequality (4.6) conveys a message that accuracy of the estimate depends on information from both classes of labeled data. This concern motivates us to consider a different sampling scheme, under which parallel results to Theorem 4.5 and Theorem 4.6 are developed, and relegated to Section 6.

## 5. Extensions

### 5.1. A Different Sampling Scheme

We now consider a model for observations that is more standard in statistical learning theory (Devroye et al., 1996; Boucheron et al., 2005, see, e.g.,).

Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be  $n$  independent copies of the random couple  $(X, Y) \in \mathcal{X} \times \{-1, 1\}$ . Denote by  $P_X$  the marginal distribution of  $X$  and by  $\eta(x) = \mathbb{E}[Y|X = x]$  the regression function of  $Y$  onto  $X$ . Denote by  $p$  the probability of positive label and observe that

$$p = \mathbb{P}[Y = 1] = \mathbb{E}(\mathbb{P}[Y = 1|X]) = \frac{1 + \mathbb{E}[\eta(X)]}{2}.$$

In what follows, we assume that  $P_X(\eta(X) = -1) \vee P_X(\eta(X) = 1) < 1$  so that  $p \in (0, 1)$ .

Let  $N^- = \text{card}\{Y_i : Y_i = -1\}$  be the random number of instances labeled  $-1$  and  $N^+ = n - N^- = \text{card}\{Y_i : Y_i = 1\}$ . In this setup, the NP classifier is defined as in

subsection 3.2 where  $n^-$  and  $n^+$  are replaced by  $N^-$  and  $N^+$  respectively. To distinguish this classifier from  $\tilde{h}^\tau$  previously defined, we denote the NP classifier obtained with this sampling scheme by  $\tilde{h}_n^\tau$ .

Let the event  $\mathcal{F}$  be defined by

$$\mathcal{F} = \{R_\varphi^-(\tilde{h}_n^\tau) \leq \alpha\} \cap \{R_\varphi^+(\tilde{h}_n^\tau) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R_\varphi^+(h) \leq \frac{4\varphi(1)\tau}{(1-\bar{\varepsilon})\alpha\sqrt{N^-}} + \frac{2\tau}{\sqrt{N^+}}\}.$$

Denote  $\mathcal{B}_{n^-} = \{Y_1 = \dots = Y_{n^-} = -1, Y_{n^-+1} = \dots = Y_n = 1\}$ . Although the event  $\mathcal{B}_{n^-}$  is different from the event  $\{N^- = n^-\}$ , symmetry leads to the following key observation:

$$\mathbb{P}(\mathcal{F}|N^- = n^-) = \mathbb{P}(\mathcal{F}|\mathcal{B}_{n^-}).$$

Therefore, under the conditions of Theorem 4.2, we find that for  $n^- \geq n_0$  the event  $\mathcal{F}$  satisfies

$$\mathbb{P}(\mathcal{F}|N^- = n^-) \geq 1 - 2\delta. \quad (5.1)$$

We obtain the following corollary of Theorem 4.2.

**Corollary 5.1** *Let  $\varphi, \tau, \delta$  and  $\alpha$  be the same as in Theorem 4.1, and  $\tilde{h}_n^\tau$  be the NP classifier obtained with the current sampling scheme. Then under Assumption 1, if  $n > 2n_0/(1-p)$ , where  $n_0$  is defined in (4.4), we have with probability  $(1-2\delta)(1-e^{-\frac{n(1-p)^2}{2}})$ ,*

$$R^-(\tilde{h}_n^\tau) \leq R_\varphi^-(\tilde{h}_n^\tau) \leq \alpha \quad (5.2)$$

and

$$R_\varphi^+(\tilde{h}_n^\tau) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R_\varphi^+(h) \leq \frac{4\varphi(1)\tau}{(1-\bar{\varepsilon})\alpha\sqrt{N^-}} + \frac{2\tau}{\sqrt{N^+}}. \quad (5.3)$$

Moreover, with probability  $1 - 2\delta - e^{-\frac{n(1-p)^2}{2}} - e^{-\frac{np^2}{2}}$ , we have simultaneously (5.2) and

$$R_\varphi^+(\tilde{h}_n^\tau) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R_\varphi^+(h) \leq \frac{4\sqrt{2}\varphi(1)\tau}{(1-\bar{\varepsilon})\alpha\sqrt{n(1-p)}} + \frac{2\sqrt{2}\tau}{\sqrt{np}}. \quad (5.4)$$

## 5.2. Chance constrained optimization

Implementing the Neyman-Pearson paradigm for the convexified binary classification bears strong connections with chance constrained optimization. A recent account of such problems can be found in Ben-Tal et al. (2009, Chapter 2) and we refer to this book for references and applications. A chance constrained optimization problem is of the following form:

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \mathbb{P}\{F(\lambda, \xi) \leq 0\} \geq 1 - \alpha, \quad (5.1)$$

where  $\xi$  is a random vector,  $\Lambda \subset \mathbb{R}^M$  is convex,  $\alpha$  is a small positive number and  $f$  is a deterministic real valued convex function. For simplicity, we take  $F$  to be scalar valued but extensions to vector valued functions and conic orders are considered in (see, e.g., Ben-Tal et al., 2009, Chapter 10). Moreover, it is standard to assume that  $F(\cdot, \xi)$  is convex almost surely.

Problem (5.1) may not be convex because the chance constraint  $\{\lambda \in \Lambda, : \mathbb{P}\{F(\lambda, \xi) \leq 0\} \geq 1 - \alpha\}$  is not convex in general and thus may not be tractable. To solve this problem, Prékopa (1995) and Lagoa et al. (2005) have derived sufficient conditions on the distribution of  $\xi$  for the chance constraint to be convex. On the other hand, Calafiore and Campi (2006) initiated a different treatment of the problem where no assumption on the distribution of  $\xi$  is made, in line with the spirit of statistical learning. In that paper, they introduced the so-called *scenario approach* based on a sample  $\xi_1, \dots, \xi_n$  of independent copies of  $\xi$ . The scenario approach consists of solving

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad F(\lambda, \xi_i) \leq 0, i = 1, \dots, n. \quad (5.2)$$

Calafiore and Campi (2006) showed that under certain conditions, if the sample size  $n$  is bigger than some  $n(\alpha, \delta)$ , then with probability  $1 - \delta$ , the optimal solution  $\hat{\lambda}^{sc}$  of (5.2) is feasible for (5.1). The authors did not address the control of the term  $f(\hat{\lambda}^{sc}) - f^*$  where  $f^*$  denotes the optimal objective value in (5.1).

In an attempt to overcome this limitation, a new *analytical approach* was introduced by (Nemirovski and Shapiro, 2006). It amounts to solving the following convex optimization problem

$$\min_{\lambda \in \Lambda, t \in \mathbb{R}^s} f(\lambda) \quad \text{s.t.} \quad G(\lambda, t) \leq 0, \quad (5.3)$$

in which  $t$  is some additional instrumental variable and where  $G(\cdot, t)$  is convex. The problem (5.3) provides a conservative convex approximation to (5.1), in the sense that every  $\lambda$  feasible for (5.3) is also feasible for (5.1). Nemirovski and Shapiro (2006) considered a particular class of conservative convex approximation where the key step is to replace  $P_\xi\{F(\lambda, \xi) \geq 0\}$  by  $\mathbb{E}\varphi(F(\lambda, \xi))$  in (5.1), where  $\varphi$  a nonnegative, nondecreasing, convex function that takes 1 at 0. Nemirovski and Shapiro (2006) discuss several choices of  $\varphi$  including hinge loss and exponential loss, with a focus on the latter that they name *Bernstein Approximation*.

The idea of a conservative convex approximation is also what we employ in our paper. Denote by  $P^-$  the conditional distribution of  $X$  given  $Y = -1$ . In a parallel form of (5.1), we cast our target problem as

$$\min_{\lambda \in \Lambda} R^+(\mathbf{h}_\lambda) \quad \text{s.t.} \quad P^-\{\mathbf{h}_\lambda(X) \leq 0\} \geq 1 - \alpha, \quad (5.4)$$

where  $\Lambda$  is the flat simplex of  $\mathbb{R}^M$ .

The problem (5.4) differs from (5.1) in that  $R^+(\mathbf{h}_\lambda)$  is not a convex function of  $\lambda$ . Replacing  $R^+(\mathbf{h}_\lambda)$  by  $R_\varphi^+(\mathbf{h}_\lambda)$  turns (5.4) into a standard chance constrained optimization problem:

$$\min_{\lambda \in \Lambda} R_\varphi^+(\mathbf{h}_\lambda) \quad \text{s.t.} \quad P^-\{\mathbf{h}_\lambda(X) \leq 0\} \geq 1 - \alpha. \quad (5.5)$$

However, there are two important differences in our setting, so that we cannot use directly Scenario Approach or Bernstein Approximation or other analytical approaches to (5.1). First,  $R_\varphi^+(\mathbf{h}_\lambda)$  is an *unknown* function of  $\lambda$ . Second, we assume minimum knowledge about  $P^-$ . On the other hand, chance constrained optimization techniques in previous literature assume knowledge about the distribution of the random vector  $\xi$ . For example, Nemirovski and Shapiro (2006) require that the moment generating function of the random vector  $\xi$  is efficiently computable to study the Bernstein Approximation.

Given a finite sample, it is not feasible to construct a strictly conservative approximation to the constraint in (5.5). Instead, what is possible is to ensure that if we learned  $\hat{f}_\lambda$  from the sample, this constraint is satisfied with high probability  $1 - \delta$ , i.e., the classifier is approximately feasible for (5.5). In retrospect, our approach to (5.5) is an innovative hybrid between the analytical approach based on convex surrogates and the scenario approach.

We do have structural assumptions on the scope of the problem. Let  $g_j, j \in \{1, \dots, M\}$  be arbitrary functions that take values in  $[-1, 1]$  and  $F(\lambda, \xi) = \sum_{j=1}^N \lambda_j g_j(\xi)$ . Consider a convexified version of (5.1):

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \mathbb{E}[\varphi(F(\lambda, \xi))] \leq \alpha, \quad (5.6)$$

where  $\varphi$  is a  $L$ -Lipschitz convex surrogate,  $L > 0$ . Suppose that we observe a sample  $(\xi_1, \dots, \xi_n)$  that are independent copies of  $\xi$ . Denote by  $f_\varphi^*$  the value of the objective at the optimum in (5.6). We propose to approximately solve the above problem by

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \sum_{i=1}^n \varphi(F(\lambda, \xi_i)) \leq n\alpha - \tau\sqrt{n},$$

for some  $\tau > 0$  to be defined. Denote by  $\tilde{\lambda}$  any solution to this problem. The following theorem summarizes our contribution to chance constrained optimization.

**Theorem 5.1** *Fix constants  $\delta, \alpha \in (0, 1), L > 0$  and let  $\varphi : [-1, 1] \rightarrow \mathbb{R}^+$  be a given  $L$ -Lipschitz convex surrogate. Define*

$$\tau = 4\sqrt{2}L \sqrt{\log\left(\frac{2M}{\delta}\right)}.$$

*Then, the following hold with probability at least  $1 - 2\delta$*

- (i)  $\tilde{\lambda}$  is feasible for (5.1).
- (ii) *If there exists  $\varepsilon \in (0, 1)$  such that the constraint  $\mathbb{E}[\varphi(F(\lambda, \xi))] \leq \varepsilon\alpha$  is feasible for some  $\lambda \in \Lambda$ , then for*

$$n \geq \left( \frac{4\tau}{(1-\varepsilon)\alpha} \right)^2,$$

*we have*

$$f(\tilde{\lambda}) - f_\varphi^* \leq \frac{4\varphi(1)\tau}{(1-\varepsilon)\alpha\sqrt{n}}.$$

The proof essentially follows that of Theorem 4.2 and we omit it. The limitations of Theorem 5.1 include rigid structural assumptions on the function  $F$  and on the set  $\Lambda$ . Also, we did not address the effect of replacing the indicator function by a convex surrogate; this investigation is beyond the scope of this paper.

## 6. Proofs

### 6.1. Proof of Theorem 4.1

We begin with the following lemma, which is extensively used in the sequel. Its proof relies on standard arguments to bound suprema of empirical processes. Recall that  $\{h_1, \dots, h_M\}$  is family of  $M$  classifiers such that  $h_j : \mathcal{X} \rightarrow [-1, 1]$  and that for any  $\lambda$  in the simplex  $\Lambda \subset \mathbb{R}^M$ ,  $h_\lambda$  denotes the convex combination defined by

$$h_\lambda = \sum_{j=1}^N \lambda_j h_j.$$

The following standard notation in empirical process theory will be used. Let  $X_1, \dots, X_n \in \mathcal{X}$  be  $n$  i.i.d random variables with marginal distribution  $P$ . Then for any measurable function  $f : \mathcal{X} \rightarrow \mathbb{R}$ , we write

$$P_n(f) = \frac{1}{n} \sum_{i=1}^n f(X_i) \quad \text{and} \quad P(f) = \mathbb{E}f(X) = \int f dP.$$

Moreover, the Rademacher average of  $f$  is defined as

$$R_n(f) = \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i),$$

where  $\varepsilon_1, \dots, \varepsilon_n$  are i.i.d. Rademacher random variables such that  $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$  for  $i = 1, \dots, n$ .

**Lemma 6.1** *Fix  $L > 0, \delta \in (0, 1)$ . Let  $X_1, \dots, X_n$  be  $n$  i.i.d random variables on  $\mathcal{X}$  with marginal distribution  $P$ . Moreover, let  $\varphi : [-1, 1] \rightarrow \mathbb{R}$  an  $L$ -Lipschitz function. Then, with probability at least  $1 - \delta$ , it holds*

$$\sup_{\lambda \in \Lambda} |(P_n - P)(\varphi \circ h_\lambda)| \leq \frac{4\sqrt{2}L}{\sqrt{n}} \sqrt{\log\left(\frac{2M}{\delta}\right)}.$$

PROOF. Define  $\bar{\varphi}(\cdot) \doteq \varphi(\cdot) - \varphi(0)$ , so that  $\bar{\varphi}$  is an  $L$ -Lipschitz function that satisfies  $\bar{\varphi}(0) = 0$ . Moreover, for any  $\lambda \in \Lambda$ , it holds

$$(P_n - P)(\varphi \circ h_\lambda) = (P_n - P)(\bar{\varphi} \circ h_\lambda).$$

Let  $\Phi : \mathbb{R} \rightarrow \mathbb{R}_+$  be a given convex increasing function. Applying successively the symmetrization and the contraction inequalities (see, e.g., Koltchinskii, 2011, Section 2), we find

$$\mathbb{E}\Phi\left(\sup_{\lambda \in \Lambda} |(P_n - P)(\bar{\varphi} \circ h_\lambda)|\right) \leq \mathbb{E}\Phi\left(2 \sup_{\lambda \in \Lambda} |R_n(\bar{\varphi} \circ h_\lambda)|\right) \leq \mathbb{E}\Phi\left(4L \sup_{\lambda \in \Lambda} |R_n(h_\lambda)|\right).$$

Observe now that  $\lambda \mapsto |R_n(h_\lambda)|$  is a convex function and Theorem 32.2 in Rockafellar (1997) entails that

$$\sup_{\lambda \in \Lambda} |R_n(h_\lambda)| = \max_{1 \leq j \leq M} |R_n(h_j)|.$$

We now use a Chernoff bound to control this quantity. To that end, fix  $s, t > 0$ , and observe that

$$\begin{aligned} \mathbb{P} \left( \sup_{\lambda \in \Lambda} |(P_n - P)(\varphi \circ h_\lambda)| > t \right) &\leq \frac{1}{\Phi(st)} \mathbb{E} \Phi \left( s \sup_{\lambda \in \Lambda} |(P_n - P)(\bar{\varphi} \circ h_\lambda)| \right) \\ &\leq \frac{1}{\Phi(st)} \mathbb{E} \Phi \left( 4Ls \max_{1 \leq j \leq M} |R_n(h_j)| \right). \end{aligned} \quad (6.7)$$

Moreover, since  $\Phi$  is increasing,

$$\begin{aligned} \mathbb{E} \Phi \left( 4Ls \max_{1 \leq j \leq M} |R_n(h_j)| \right) &= \mathbb{E} \max_{1 \leq j \leq M} \Phi(4Ls|R_n(h_j)|) \\ &\leq \sum_{j=1}^M \mathbb{E} [\Phi(4LsR_n(h_j)) \vee \Phi(-4LsR_n(h_j))] \\ &\leq 2 \sum_{j=1}^M \mathbb{E} \Phi(4LsR_n(h_j)). \end{aligned} \quad (6.8)$$

Now choose  $\Phi(\cdot) = \exp(\cdot)$ , then

$$\mathbb{E} \Phi(4LsR_n(h_j)) = \prod_{i=1}^n \mathbb{E} \cosh \left( \frac{4Ls h_j(X_i)}{n} \right) \leq \exp \left( \frac{8L^2 s^2}{n} \right),$$

where  $\cosh$  is the hyperbolic cosine function and where in the inequality, we used the fact that  $|h_j(X_i)| \leq 1$  for any  $i, j$  and  $\cosh(x) \leq \exp(x^2/2)$ . Together with (6.7) and (6.8), it yields

$$\mathbb{P} \left( \sup_{\lambda \in \Lambda} |(P_n - P)(\varphi \circ h_\lambda)| > t \right) \leq 2M \inf_{s>0} \exp \left( \frac{8L^2 s^2}{n} - st \right) \leq 2M \exp \left( -\frac{nt^2}{32L^2} \right).$$

Choosing

$$t = \frac{4\sqrt{2}L}{\sqrt{n}} \sqrt{\log \left( \frac{2M}{\delta} \right)},$$

completes the proof of the Lemma.  $\square$

We now proceed to the proof of Theorem 4.1. Note first that from the properties of  $\varphi$ ,  $R^-(h) \leq R_\varphi^-(h)$ . Next, we have for any data-dependent classifier  $h \in \mathcal{H}^{\text{conv}}$  such that  $\hat{R}_\varphi^-(h) \leq \alpha_\tau$ :

$$R_\varphi^-(h) \leq \hat{R}_\varphi^-(h) + \sup_{h \in \mathcal{H}^{\text{conv}}} |\hat{R}_\varphi^-(h) - R_\varphi^-(h)| \leq \alpha - \frac{\tau}{\sqrt{n^-}} + \sup_{h \in \mathcal{H}^{\text{conv}}} |\hat{R}_\varphi^-(h) - R_\varphi^-(h)|.$$

Lemma 6.1 implies that, with probability  $1 - \delta$

$$\sup_{h \in \mathcal{H}^{\text{conv}}} |\hat{R}_\varphi^-(h) - R_\varphi^-(h)| = \sup_{\lambda \in \Lambda} |(P_{n^-}^- - P^-)(\varphi \circ h_\lambda)| \leq \frac{\tau}{\sqrt{n^-}}.$$

The previous two displays imply that  $R_\varphi^-(h) \leq \alpha$  with probability  $1 - \delta$ , which completes the proof of Theorem 4.1.

## 6.2. Proof of Proposition 4.1

The proof of this proposition builds upon the following lemma.

**Lemma 6.2** *Let  $\gamma(\alpha) = \inf_{h_\lambda \in \mathcal{H}^{\varphi, \alpha}} R_\varphi^+(h_\lambda)$ , then  $\gamma$  is a non-increasing convex function on  $[0, 1]$ .*

PROOF. First, it is clear that  $\gamma$  is a non-increasing function of  $\alpha$  because for  $\alpha' > \alpha$ ,  $\{h_\lambda \in \mathcal{H}^{\text{conv}} : R_\varphi^-(h_\lambda) \leq \alpha\} \subset \{h_\lambda \in \mathcal{H}^{\text{conv}} : R_\varphi^-(h_\lambda) \leq \alpha'\}$ .

We now show that  $\gamma$  is convex. To that end, observe first that since  $\varphi$  is continuous on  $[-1, 1]$ , the set  $\{\lambda \in \Lambda : h_\lambda \in \mathcal{H}^{\varphi, \alpha}\}$  is compact. Moreover, the function  $\lambda \mapsto R_\varphi^+(h_\lambda)$  is convex. Therefore, there exists  $\lambda^* \in \Lambda$  such that

$$\gamma(\alpha) = \inf_{h_\lambda \in \mathcal{H}^{\varphi, \alpha}} R_\varphi^+(h_\lambda) = \min_{h_\lambda \in \mathcal{H}^{\varphi, \alpha}} R_\varphi^+(h_\lambda) = R_\varphi^+(h_{\lambda^*}).$$

Now, fix  $\alpha_1, \alpha_2 \in [0, 1]$ . From the above considerations, there exists  $\lambda_1, \lambda_2 \in \Lambda$  such that  $\gamma(\alpha_1) = R_\varphi^+(h_{\lambda_1})$  and  $\gamma(\alpha_2) = R_\varphi^+(h_{\lambda_2})$ . For any  $\theta \in (0, 1)$ , define the convex combinations  $\bar{\alpha}_\theta = \theta\alpha_1 + (1 - \theta)\alpha_2$  and  $\bar{\lambda}_\theta = \theta\lambda_1 + (1 - \theta)\lambda_2$ . Since  $\lambda \mapsto R_\varphi^-(h_\lambda)$  is convex, it holds

$$R_\varphi^-(h_{\bar{\lambda}_\theta}) \leq \theta R_\varphi^-(h_{\lambda_1}) + (1 - \theta) R_\varphi^-(h_{\lambda_2}) \leq \theta\alpha_1 + (1 - \theta)\alpha_2 = \bar{\alpha}_\theta,$$

so that  $h_{\bar{\lambda}_\theta} \in \mathcal{H}^{\varphi, \bar{\alpha}_\theta}$ . Hence,  $\gamma(\bar{\alpha}_\theta) \leq R_\varphi^+(h_{\bar{\lambda}_\theta})$ . Together with the convexity of  $\varphi$ , it yields

$$\gamma(\theta\alpha_1 + (1 - \theta)\alpha_2) \leq \theta R_\varphi^+(h_{\lambda_1}) + (1 - \theta) R_\varphi^+(h_{\lambda_2}) = \theta\gamma(\alpha_1) + (1 - \theta)\gamma(\alpha_2).$$

□

We now complete the proof of Proposition 4.1. For any  $x \in [0, 1]$ , let  $\gamma(x) = \inf_{h \in \mathcal{H}^{\varphi, x}} R_\varphi^+(h)$  and observe that the statement of the proposition is equivalent to

$$\gamma(\alpha - \nu) - \gamma(\alpha) \leq \varphi(1) \frac{\nu}{\nu_0 - \nu} \quad 0 < \nu < \nu_0. \quad (6.9)$$

Lemma 6.2 together with the assumption that  $\mathcal{H}^{\varphi, \alpha - \nu_0} \neq \emptyset$  imply that  $\gamma$  is a non-increasing convex real-valued function on  $[\alpha - \nu_0, 1]$  so that

$$\gamma(\alpha - \nu) - \gamma(\alpha) \leq \nu \sup_{g \in \partial\gamma(\alpha - \nu)} |g|,$$

where  $\partial\gamma(\alpha - \nu)$  denotes the sub-differential of  $\gamma$  at  $\alpha - \nu$ . Moreover, since  $\gamma$  is a non-increasing convex function on  $[\alpha - \nu_0, \alpha - \nu]$ , it holds

$$\gamma(\alpha - \nu_0) - \gamma(\alpha - \nu) \geq (\nu - \nu_0) \sup_{g \in \partial\gamma(\alpha - \nu)} |g|.$$

The previous two displays yield

$$\gamma(\alpha - \nu) - \gamma(\alpha) \leq \nu \frac{\gamma(\alpha - \nu_0) - \gamma(\alpha - \nu)}{\nu - \nu_0} \leq \nu \frac{\varphi(1)}{\nu - \nu_0}.$$

### 6.3. Proof of Theorem 4.2

Define the events  $\mathcal{E}^-$  and  $\mathcal{E}^+$  by

$$\begin{aligned}\mathcal{E}^- &= \bigcap_{h \in \mathcal{H}^{\text{conv}}} \left\{ |\hat{R}_\varphi^-(h) - R_\varphi^-(h)| \leq \frac{\tau}{\sqrt{n^-}} \right\}, \\ \mathcal{E}^+ &= \bigcap_{h \in \mathcal{H}^{\text{conv}}} \left\{ |\hat{R}_\varphi^+(h) - R_\varphi^+(h)| \leq \frac{\tau}{\sqrt{n^+}} \right\}.\end{aligned}$$

Lemma 6.1 implies

$$\mathbb{P}(\mathcal{E}^-) \wedge \mathbb{P}(\mathcal{E}^+) \geq 1 - \delta. \quad (6.10)$$

Note first Theorem 4.1 implies that (4.5) holds with probability  $1 - \delta$ . Observe now that the l.h.s of (4.6) can be decomposed as

$$R_\varphi^+(\tilde{h}^\tau) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R_\varphi^+(h) = A_1 + A_2 + A_3,$$

where

$$\begin{aligned}A_1 &= \left( R_\varphi^+(\tilde{h}^\tau) - \hat{R}_\varphi^+(\tilde{h}^\tau) \right) + \left( \hat{R}_\varphi^+(\tilde{h}^\tau) - \min_{h \in \mathcal{H}_{n^-}^{\varphi,\alpha_\tau}} R_\varphi^+(h) \right) \\ A_2 &= \min_{h \in \mathcal{H}_{n^-}^{\varphi,\alpha_\tau}} R_\varphi^+(h) - \min_{h \in \mathcal{H}_{n^-}^{\varphi,\alpha_{2\tau}}} R_\varphi^+(h) \\ A_3 &= \min_{h \in \mathcal{H}_{n^-}^{\varphi,\alpha_{2\tau}}} R_\varphi^+(h) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R_\varphi^+(h)\end{aligned}$$

To bound  $A_1$  from above, observe that

$$A_1 \leq \sup_{h \in \mathcal{H}_{n^-}^{\varphi,\alpha_\tau}} 2|\hat{R}_\varphi^+(h) - R_\varphi^+(h)| \leq 2 \sup_{h \in \mathcal{H}^{\text{conv}}} |\hat{R}_\varphi^+(h) - R_\varphi^+(h)|.$$

Therefore, on the event  $\mathcal{E}^+$  it holds

$$A_1 \leq \frac{2\tau}{\sqrt{n^+}}.$$

We now treat  $A_2$ . Note that  $A_2 \leq 0$  if  $\mathcal{H}^{\varphi,\alpha_{2\tau}} \subset \mathcal{H}_{n^-}^{\varphi,\alpha_\tau}$  and note that  $A_2 \leq 0$  on this event. But this event contains  $\mathcal{E}^-$  so that  $A_2 \leq 0$  on the event  $\mathcal{E}^-$ .

Finally, to control  $A_3$ , observe that under Assumption 1, Proposition 4.1 can be applied with  $\nu = 2\tau/\sqrt{n^-}$  and  $\nu_0 = (1 - \bar{\varepsilon})\alpha$ . Indeed, the assumptions of the theorem imply that  $\nu \leq \nu_0/2$ . It yields

$$A_3 \leq \frac{4\varphi(1)\tau}{(1 - \bar{\varepsilon})\alpha\sqrt{n^-}}.$$

Combining the bounds on  $A_1$ ,  $A_2$  and  $A_3$  obtained above, we find that (4.6) holds on the event  $\mathcal{E}^- \cap \mathcal{E}^+$  that has probability at least  $1 - 2\delta$  in view of (6.10).

The last statement of the theorem follows directly from the definition of  $\tau$ .

#### 6.4. Proof of Corollary 5.1

We will use the following Lemma to bound the left tail of a binomial distribution, whose proof we omit for this short version.

**Lemma 6.3** *Let  $N$  be a binomial random variables with parameters  $n \geq 1$  and  $q \in (0, 1)$ . Then, for any  $t > 0$  such that  $t \leq nq/2$ , it holds*

$$\mathbb{P}(N \geq t) \geq 1 - e^{-\frac{nq^2}{2}}.$$

Now prove (5.3),

$$\begin{aligned} \mathbb{P}(\mathcal{F}) &= \sum_{n^- = 0}^n \mathbb{P}(\mathcal{F}|N^- = n^-)\mathbb{P}(N^- = n^-) \\ &\geq \sum_{n^- = n_0}^n \mathbb{P}(\mathcal{F}|N^- = n^-)\mathbb{P}(N^- = n^-) \\ &\geq (1 - 2\delta)\mathbb{P}(N^- \geq n_0), \end{aligned}$$

where in the last inequality, we used (5.1). Applying now Lemma 6.3, we obtain

$$\mathbb{P}(N^- \geq n_0) \geq 1 - e^{-\frac{n(1-p)^2}{2}}.$$

Therefore,

$$\mathbb{P}(\mathcal{F}) \geq (1 - 2\delta)(1 - e^{-\frac{n(1-p)^2}{2}}),$$

which completes the proof of (5.3).

The proof of (5.4) follows by observing that

$$\left\{ R_\varphi^+(\tilde{h}_n^\tau) - \min_{h \in \mathcal{H}^{\varphi, \alpha}} R_\varphi^+(h) > \frac{4\sqrt{2}\varphi(1)\tau}{(1 - \bar{\varepsilon})\alpha\sqrt{n(1-p)}} + \frac{2\sqrt{2}\tau}{\sqrt{np}} \right\} \subset \mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3 = (\mathcal{A}_1 \cap \mathcal{A}_2^c) \cup \mathcal{A}_2 \cup \mathcal{A}_3,$$

where

$$\begin{aligned} \mathcal{A}_1 &= \left\{ R_\varphi^+(\tilde{h}_n^\tau) - \min_{h \in \mathcal{H}^{\varphi, \alpha}} R_\varphi^+(h) > \frac{4\varphi(1)\tau}{(1 - \bar{\varepsilon})\alpha\sqrt{N^-}} + \frac{2\tau}{\sqrt{N^+}} \right\} \subset \mathcal{F}^c, \\ \mathcal{A}_2 &= \{N^- < n(1-p)/2\}, \\ \mathcal{A}_3 &= \{N^+ < np/2\}. \end{aligned}$$

Since  $\mathcal{A}_2^c \subset \{N^- \geq n_0\}$ , we find

$$\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2^c) \leq \sum_{n^- \geq n_0} \mathbb{P}(\mathcal{F}^c|N^- = n^-)\mathbb{P}(N^- = n^-) \leq 2\delta.$$

Next, using Lemma 6.3, we get

$$\mathbb{P}(\mathcal{A}_2) \leq e^{-\frac{n(1-p)^2}{2}} \quad \text{and} \quad \mathbb{P}(\mathcal{A}_3) \leq e^{-\frac{np^2}{2}}.$$

Hence, we find

$$\mathbb{P} \left\{ R_\varphi^+(\tilde{h}_n^\tau) - \min_{h \in \mathcal{H}^{\varphi, \alpha}} R_\varphi^+(h) > \frac{4\sqrt{2}\varphi(1)\tau}{(1 - \bar{\varepsilon})\alpha\sqrt{n(1-p)}} + \frac{2\sqrt{2}\tau}{\sqrt{np}} \right\} \leq 2\delta + e^{-\frac{n(1-p)^2}{2}} + e^{-\frac{np^2}{2}},$$

which completes the proof of the corollary.

## References

- P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 2006.
- Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2009.
- G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *J. Mach. Learn. Res.*, 11:2973–3009, Nov 2010.
- Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005.
- Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.
- Giuseppe C. Calafiore and Marco C. Campi. The scenario approach to robust control design. *IEEE Trans. Automat. Control*, 51(5):742–753, 2006.
- A. Cannon, J. Howse, D. Hush, and C. Scovel. Learning with the neyman-pearsen and min-max criteria. *Technical Report LA-UR-02-2951*, 2002.
- A. Cannon, J. Howse, D. Hush, and C. Scovel. Simple classifiers. *Technical Report LA-UR-03-0193*, 2003.
- D. Casasent and X. Chen. Radial basis function neural networks for nonlinear fisher discrimination and neyman-pearsen classification. *Neural Networks*, 16(5-6):529 – 535, 2003.
- Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.
- M. Han, D. Chen, and Z. Sun. Analysis to neyman-pearsen classification with convex loss function. *Analysis in Theory and Applications*, 24(1):18–28, 2008.
- V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Berlin, 2011. Lectures from the 38th Summer School on Probability Theory held in Saint-Flour, July 2008.
- Constantino M. Lagoa, Xiang Li, and Mario Sznaier. Probabilistically constrained linear programs and risk-adjusted controller design. *SIAM J. Optim.*, 15(3):938–951 (electronic), 2005.
- E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.
- Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM J. Optim.*, 17(4):969–996, 2006.

NEYMAN-PEARSON CLASSIFICATION

- András Prékopa. *Stochastic programming*, volume 324 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1995.
- Philippe Rigollet and Xin Tong. Neyman-pearson classification, convexity and stochastic constraints. arXiv:1102.5750, February 2011.
- R. Tyrrell Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.
- R.E. Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.
- C. Scott. Comparison and design of neyman-pearson classifiers, 2005. Manuscript.
- C. Scott. Performance measures for neyman-pearson classification. *IEEE Transactions on Information Theory*, 53(8):2852–2863, 2007.
- C. Scott and R. Nowak. A neyman-pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, 2005.

RIGOLLET TONG

## Sequential Event Prediction with Association Rules

**Cynthia Rudin**

RUDIN@MIT.EDU

*MIT Sloan School of Management  
Massachusetts Institute of Technology  
77 Massachusetts Avenue Cambridge, MA 02139, USA*

**Benjamin Letham**

BLETHAM@MIT.EDU

*Operations Research Center  
Massachusetts Institute of Technology  
77 Massachusetts Avenue Cambridge, MA 02139, USA*

**Ansaf Salleb-Aouissi**

ANSAF@CCLS.COLUMBIA.EDU

*Center for Computational Learning Systems  
Columbia University  
475 Riverside Drive, New York, NY, 10115, USA*

**Eugene Kogan**

KOGAN.GENE@GMAIL.COM

*Sourcetone  
1295 5th Avenue Apt 14D, New York, NY 10029, USA*

**David Madigan**

MADIGAN@STAT.COLUMBIA.EDU

*Department of Statistics  
Columbia University  
1255 Amsterdam Avenue, New York, NY 10027, USA*

**Editor:** Sham Kakade, Ulrike von Luxburg

### Abstract

We consider a supervised learning problem in which data are revealed sequentially and the goal is to determine what will next be revealed. In the context of this problem, algorithms based on association rules have a distinct advantage over classical statistical and machine learning methods; however, there has not previously been a theoretical foundation established for using association rules in supervised learning. We present two simple algorithms that incorporate association rules, and provide generalization guarantees on these algorithms based on algorithmic stability analysis from statistical learning theory. We include a discussion of the strict minimum support threshold often used in association rule mining, and introduce an “adjusted confidence” measure that provides a weaker minimum support condition that has advantages over the strict minimum support. The paper brings together ideas from statistical learning theory, association rule mining and Bayesian analysis.

**Keywords:** statistical learning theory, algorithmic stability, association rules, sequence prediction

### 1. Introduction

Given a “sequence database” of past event sequences to learn from, we aim to predict the next event within a current event sequence. Consider for instance, the data generated by a customer placing items into the virtual basket of an online grocery store such as NYC’s

Fresh Direct, Peapod by Stop & Shop, or Roche Bros. The customer adds items one by one into the current basket, creating a sequence of events. The customer has identified him- or herself, so that all past orders are known. After each item selection, a confirmation screen contains a small list of recommendations for items that are not already in the basket. If the store can find patterns within the customer’s past purchases, it may be able to accurately recommend the next item that the customer will add to the basket. There are many other domains in which a sequence of events is repeated in a somewhat similar way, and predictions need to be made before each event. Another example is to predict each next symptom of a sick patient, given the patient’s past sequence of symptoms and treatments, and a database of the timeline of symptoms and treatments for other patients. In these examples, a subset of past events (for instance, a set of ingredients for a particular recipe, or a set of symptoms associated with a particular disease) can be useful in predicting the next event. We nickname the problem of predicting these sequentially revealed events based on subsets of past events “the online grocery store problem.”

In order to make predictions using subsets of data, we employ *association rules* (Agrawal et al., 1993). An association rule in this setting is an implication  $a \rightarrow b$  (such as *lettuce and carrots*  $\rightarrow$  *tomatoes*), where  $a$  is a subset of items, and  $b$  is a single item. Association rule mining has proven successful for applications in market basket analysis (cross selling, product placement, affinity promotion, see also Kohavi et al., 2004), mining gene expression data (Jiang and Gruenwald, 2005) and weblog analysis (Huang et al., 2002). The association rule approach has the distinct advantage in being able to directly model underlying conditional probabilities  $P(b|a)$  eschewing the linearity assumptions underlying many classical supervised classification, regression, and ranking methods. However, association rules are generally used as an exploratory tool, rather than a predictive tool (with some exceptions, such as the work of Liu et al., 1998; Veloso et al., 2008), and have not previously been established as a principled approach for supervised learning. Specifically, rule mining algorithms are generally used for finding patterns in databases without a hold-out test set.

Our main contribution is a framework and generalization analysis, in the context of the online grocery store problem, for supervised learning algorithms based on association rules. An important part of this analysis is how a fundamental property of a rule, namely the “support,” is incorporated into the generalization bounds. The “support” of an itemset for the online grocery store problem is the number of times that the itemset has appeared in the sequence database. For instance, the support of *lettuce* is the number of times lettuce has been purchased in the past. Typically in association rule mining, a strict minimum support threshold condition is placed on the support of itemsets within a rule, so that rules falling below the minimum support threshold are simply discarded. The idea of a condition on the support is not shared with other types of supervised learning algorithms, since they do not use subsets in the same way as rule mining. Thus a new aspect of generalization is explored in our framework in that it handles predictions created from subsets of data. In classical supervised learning paradigms, bounds scale only with the sample size, and a large sample is necessary for generalization. In the context of association rules, the minimum support threshold forces predictions to be made only when there are enough data. Thus, in the association rules framework, there are now two mechanisms for generalization: first a large sample, and second, a minimum support. These are separate mechanisms, in the sense that it is possible to generalize with a somewhat small sample size and a large minimum

support threshold, and it is also possible to generalize with a large sample size and no support threshold. We thus derive two types of bounds: large sample bounds, which scale with the sample size, and small sample bounds, which scale with the minimum support of rules. Using both large and small sample bounds (that is, the minimum of the two bounds) gives a complete picture. The large sample bound is of order  $\mathcal{O}(\sqrt{1/m})$  as in other supervised problems (classification, ranking, regression), where  $m$  denotes the number of event sequences in the database, that is, the number of past baskets ordered by the online grocery store customer.

Our bounds are derived using a specific notion of algorithmic stability called “pointwise hypothesis stability.” The original notions of algorithmic stability were invented in the 1970’s and have been revitalized recently (Devroye and Wagner, 1979; Bousquet and Elisseeff, 2002), the main idea being that algorithms may be better able to generalize if they are insensitive to small changes in the training data such as the removal or change of one training example. The pointwise hypothesis stability specifically considers the average change in loss that will occur at one of the training examples if that example is removed from the training set. Our generalization analysis uses conditions on the minimum support of rules in order to bound the pointwise hypothesis stability.

There are two algorithms considered in this work. At the core of each algorithm is a method for rank-ordering association rules where the list of possible rules is generated using the customer’s past purchase history and subsets of items within the current basket. These algorithms build off of the rule mining literature that has been developing since the early 1990’s (Agrawal et al., 1993) by using an application-specific rule mining method as a subroutine. Both of our algorithms are simple enough that they can be understood by users, customers, patients, managers, etc; an advantage of using association rules is that they are interpretable. Rules can provide a simple reason to the customer why an item might be relevant, or identify that a key ingredient is missing from a particular recipe. One of the algorithms considered in this work uses a fixed minimum support threshold to exclude rules whose itemsets occur rarely. Then the remaining rules are ranked according to the “confidence,” which for rule  $a \rightarrow b$  is the empirical probability that  $b$  will be in the basket given that  $a$  is in the basket. The right-hand sides of the highest ranked rules will be recommended by the algorithm. However, the use of a strict minimum support threshold is problematic for several well-known reasons, for instance it is known that important rules (“nuggets,” which are rare but strong rules) are often excluded by a minimum support threshold condition.

The other algorithm introduced in this work provides an alternative to the minimum support threshold, in that rules are ranked by an “adjusted” confidence, which is a simple Bayesian shrinkage estimator of the probability of a rule  $P(b|a)$ . The right-hand sides of rules with the highest adjusted confidence are recommended by the algorithm. For this algorithm, the generalization guarantee (or bias-variance tradeoff) is smoothly controlled by a parameter  $K$ , which provides only a weak (less restrictive) minimum support condition. The key benefits of an algorithm based on the adjusted confidence are that: 1) it allows the possibility of choosing very accurate (high confidence) rules that have appeared very few times in the training set (low support), and 2) given two rules with the same or similar prediction accuracy on the training set (confidence), the rule that appears more frequently

(higher support) achieves a higher adjusted confidence and is thus preferred over the other rule.

All of the bounds are tied to the measure of quality (the evaluation metric, or loss function) used for the algorithm. We would like to directly compare the performance of algorithms for various settings of the adjusted confidence's  $K$  parameter (and for the minimum support threshold  $\theta$ ). It is problematic to have the loss defined using the same  $K$  value as the algorithm, in that case we would be using a different method of evaluation for each setting of  $K$ , and we would not be able to directly compare performance across different settings of  $K$ . To allow a direct comparison, we select one reference value of the adjusted confidence, called  $K_r$  (r for "reference"), and the loss depends on  $K_r$  rather than on  $K$ . The bounds are written generally in terms of  $K_r$ . The special case  $K_r = 0$  is where the algorithm is evaluated with respect to the confidence measure. The small sample bound for the adjusted confidence algorithm has two terms: one that generally decreases with  $K$  (as the support increases, there is better generalization) and the other that decreases as  $K$  gets closer to  $K_r$  (better generalization as the algorithm is closer to the way it is being measured). These two terms are thus agreeing if  $K_r > K$  and competing if  $K_r < K$ . In practice, the choice of  $K$  can be determined in several ways:  $K$  can be manually determined (for instance by the customer), it can be set using side information by "empirical Bayes" as considered by McCormick et al. (2011), or it can be set via cross-validation on an extra hold-out set.

Section 2 describes the max confidence, min support algorithm that has the hard support threshold, and the adjusted confidence algorithm that has the soft threshold. Section 3 provides the generalization analysis. Section 4 provides experimental results. Section 5 contains a discussion and summary of relevant literature. The longer version of the work (Rudin et al., 2011) contains proofs, strengthened versions of some of the results presented here, analogous results for binary classification, and a discussion regarding the suitability of other methods, specifically regression, for solving the sequential prediction problem.

## 2. Derivation of Algorithms

We assume an interface similar to that of Fresh Direct, where users add items one by one into the basket. After each selection, a confirmation screen contains a handful of recommendations for items that are not already in the customer's basket. The customer's past orders are known. The set of items is  $\mathcal{X}$ , for instance  $\mathcal{X}=\{\text{apples, bananas, pears, etc}\}$ ;  $\mathcal{X}$  is the set of possible events. The customer has a past history of orders  $S$  which is a collection of  $m$  baskets,  $S = \{z_i\}_{i=1,\dots,m}$ ,  $z_i \subseteq \mathcal{X}$ ;  $S$  is the sequence database. The customer's current basket is usually denoted by  $B \subset \mathcal{X}$ ;  $B$  is the current sequence. An algorithm uses  $B$  and  $S$  to find rules  $a \rightarrow b$ , where  $a$  is in the basket and  $b$  is not in the basket. For instance, if *salsa* and *guacamole* are in the basket  $B$  and also if *salsa*, *guacamole* and *tortilla chips* were often purchased together in  $S$ , then the rule (*salsa* and *guacamole*)  $\rightarrow$  *tortilla chips* might be used to recommend *tortilla chips*.

The support of  $a$ , written  $\text{Sup}(a)$  or  $\#a$ , is the number of times in the past the customer has ordered itemset  $a$ ,  $\text{Sup}(a) := \#a := \sum_{i=1}^m \mathbf{1}_{[a \subseteq z_i]}$ . If  $a = \emptyset$ , meaning  $a$  contains no items, then  $\#a := \sum_i 1 = m$ . The confidence of a rule  $a \rightarrow b$  is  $\text{Conf}(a \rightarrow b) := f_{S,0}(a, b) := \frac{\#(a \cup b)}{\#a}$ , the fraction of times  $b$  is purchased given that  $a$  is purchased. It is an estimate

of the conditional probability of  $b$  given  $a$ . Ultimately an algorithm should order rules by conditional probability; however, the rules that possess the highest confidence values often have a left-hand side with small support, and their confidence values do not yield good estimates for the true conditional probabilities. We introduce the “adjusted” confidence as a remedy for this problem: the *adjusted confidence* for rule  $a \rightarrow b$  is:

$$f_{S,K}(a, b) := \frac{\#(a \cup b)}{\#a + K}.$$

The adjusted confidence for  $K = 0$  is equivalent to the confidence. The adjusted confidence is a particular Bayesian estimate of the confidence. Specifically, assuming a beta prior distribution for the confidence, the posterior mean is given by:

$$\hat{p} = \frac{L + \#(a \cup b)}{L + K + \#a},$$

where  $L$  and  $K$  denote the parameters of the beta prior distribution. The beta distribution is the “conjugate” prior distribution for a binomial likelihood. For the adjusted confidence we choose  $L = 0$ . This choice yields the benefits of the lower bounds derived in this section, and the stability properties described later. The prior for the adjusted confidence tends to bias rules towards the *bottom* of the ranked list. Any rule achieving a high adjusted confidence must overcome this bias.

Other choices for  $L$  and  $K$  are meaningful. A *collaborative filtering prior* would have  $L/(L + K)$  represent the probability of purchasing item  $b$  given that item  $a$  was purchased, calculated over a subset of other customers. A *revenue management prior* would have  $L$  and  $K$  be based on the item price, favoring expensive items.

A rule cannot have a high adjusted confidence unless it has both a large enough confidence and a large enough support on the left side. To see this, take  $f_{S,K}(a, b)$  large, meaning for some  $\eta$ , we have  $f_{S,K}(a, b) > \eta$ , implying:

$$\text{Conf}(a \rightarrow b) = f_{S,0}(a, b) > \eta \frac{\#a + K}{\#a} \geq \eta,$$

$$\text{Sup}(a) = \#a \geq (\#a + K) \left[ \frac{\#(a \cup b)}{\#a + K} \right] > (\#a + K)\eta, \text{ implying } \text{Sup}(a) = \#a > \frac{\eta K}{1 - \eta}. \quad (1)$$

And further, expression (1) implies:

$$\text{Sup}(a \cup b) = \#(a \cup b) > \eta(\#a + K) > \eta K / (1 - \eta).$$

Thus, rules attaining high values of adjusted confidence have a lower bound on confidence, and a lower bound on support of both the right- and left-hand sides, which means a better estimate of the conditional probability. As  $K$  increases, rules with low support are heavily penalized with respect to the adjusted confidence, so they tend not to be at the top of the list. On the other hand, such rules might be chosen when all other rules have low confidence. That is the main advantage of having no *firm* minimum support cutoff: “nuggets” that have fairly low support may filter to the top.

We now formally state the recommendation algorithms. Both algorithms use a subroutine for mining association rules to generate a set of candidate rules. One of the simplest

- 
- Input:**  $(S, B, \mathcal{X})$ , that is, past orders  $S = \{z_i\}_{i=1,\dots,m}$ ,  $z_i \subseteq \mathcal{X}$ , current basket  $B \subset \mathcal{X}$ , set of items  $\mathcal{X}$
- Output:** Set of all rules  $\{a_j \rightarrow b_j\}_j$  where  $b_j$  is a single item that is not in the basket  $B$ , and where  $a_j$  is either a subset of items in the basket  $B$ , or else it is the empty set. Also the left-hand side  $a_j$  must be allowed (meaning it is in  $A$ ). That is, output rules  $\{a_j \rightarrow b_j\}_j$  such that  $b_j \in \mathcal{X} \setminus B$  and  $a_j \subseteq B \subset \mathcal{X}$  with  $a_j \in A$ , or  $a_j = \emptyset$ .
- 

**Algorithm 1:** *Subroutine GenRules.*

- 
- Input:**  $(\theta, \mathcal{X}, S, B, \text{GenRules}, c)$ , that is, minimum threshold parameter  $\theta$ , set of items  $\mathcal{X}$ , past orders  $S = \{z_i\}_{i=1,\dots,m}$ ,  $z_i \subseteq \mathcal{X}$ , current basket  $B \subset \mathcal{X}$ , *GenRules* generates candidate rules  $\text{GenRules}(S, B, \mathcal{X}) = \{a_j \rightarrow b_j\}_j$ , number of recommendations  $c \geq 1$
- Output:** Recommendation List, which is a subset of  $c$  items in  $\mathcal{X}$
- Algorithm:**
- 1) Apply  $\text{GenRules}(S, B, \mathcal{X})$  to get rules  $\{a_j \rightarrow b_j\}_j$  where  $a_j$  is in the basket  $B$  and  $b_j$  is not.
  - 2) Compute score for each rule  $a_j \rightarrow b_j$  as  $\bar{f}_{S,\theta}(a_j, b_j) = f_{S,0}(a_j, b_j) = \frac{\#(a_j \cup b_j)}{\#a_j}$  when support  $\#a_j \geq \theta$ , and  $\bar{f}_{S,\theta}(a_j, b_j) = 0$  otherwise.
  - 3) Reorder rules by decreasing score.
  - 4) Find the top  $c$  rules with distinct right-hand sides, and let Recommendation List be the right-hand sides of these rules.
- 

**Algorithm 2:** Max Confidence, Min Support Algorithm.

such rule mining algorithms is *GenRules*, provided as Algorithm 1, which in practice should be made specific by using a rule mining algorithm that retrieves a set of rules tailored to the application. *GenRules* can be replaced by any algorithm for mining association rules; there is a vast literature on such algorithms since the field of association rule mining evolved on their development, e.g. Apriori (Agrawal et al., 1993). *GenRules* requires a set  $A$  which is the set of allowed left-hand sides of rules.

## 2.1. Max Confidence, Min Support Algorithm

The max confidence, min support algorithm, shown as Algorithm 2, is based on the idea of eliminating rules whose itemsets occur rarely, which is commonly done in the rule-mining literature. For this algorithm, the rules are ranked by confidence, and rules that do not achieve a predetermined fixed minimum support threshold are completely omitted. The algorithm recommends the right-hand sides from the top ranked rules. Specifically, if  $c$  items are to be recommended to the user, the algorithm picks the top ranked  $c$  distinct items.

It is common that the minimum support threshold is imposed on the right and left side  $\text{Sup}(a \cup b) \geq \theta$ ; however, as long as  $\text{Sup}(a)$  is large, we can get a reasonable estimate of  $P(b|a)$ . In that sense, it is sufficient (and less restrictive) to impose the minimum support

---

**Input:**  $(K, \mathcal{X}, S, B, GenRules, c)$ , that is, parameter  $K$ , set of items  $\mathcal{X}$ , past orders  $S = \{z_i\}_{i=1,\dots,m}$ ,  $z_i \subseteq \mathcal{X}$ , current basket  $B \subset \mathcal{X}$ ,  $GenRules$  generates candidate rules  $GenRules(S, B, \mathcal{X}) = \{a_j \rightarrow b_j\}_j$ , number of recommendations  $c \geq 1$

**Output:** Recommendation List, which is a subset of  $c$  items in  $\mathcal{X}$

**Algorithm:**

- 1) Apply  $GenRules(S, B, \mathcal{X})$  to get rules  $\{a_j \rightarrow b_j\}_j$  where  $a_j$  is in the basket  $B$  and  $b_j$  is not.
  - 2) Compute adjusted confidence of each rule  $a_j \rightarrow b_j$  as  $f_{S,K}(a_j, b_j) = \frac{\#(a_j \cup b_j)}{\#a_j + K}$ .
  - 3) Reorder rules by decreasing adjusted confidence.
  - 4) Find the top  $c$  rules with distinct right-hand sides, and let Recommendation List be the right-hand sides of these rules.
- 

**Algorithm 3:** Adjusted Confidence Algorithm.

threshold on the left side:  $\text{Sup}(a) \geq \theta$ . In this work, we only have a required minimum support on the left side. As a technical note, we might worry about the minimum support threshold being so high that there are no rules that meet the threshold. This is actually not a major concern because of the minimum support being imposed only on the left-hand side: as long as  $m \geq \theta$ , all rules  $\emptyset \rightarrow b$  meet the minimum support threshold. The thresholded confidence is denoted by  $\bar{f}_{S,\theta}$ :

$$\bar{f}_{S,\theta}(a, b) := f_{S,0}(a, b) \text{ if } \#a \geq \theta, \text{ and } \bar{f}_{S,\theta}(a, b) := 0 \text{ otherwise.}$$

## 2.2. Adjusted Confidence Algorithm

The adjusted confidence algorithm is shown as Algorithm 3. A chosen value of  $K$  is used to compute the adjusted confidence for each rule, and rules are then ranked according to adjusted confidence.

The definition of the adjusted confidence makes an implicit assumption that the order in which items were placed into previous baskets is irrelevant. It is easy to include a dependence on the order by defining a “directed” version of the adjusted confidence, and calculations can be adapted accordingly. The numerator of the adjusted confidence becomes the number of past orders where  $a$  is placed in the basket *before*  $b$ .

$$f_{S,K}^{(\text{directed})}(a, b) = \frac{\#\{(a \cup b) : b \text{ follows } a\}}{\#a + K}.$$

## 3. Generalization and Stability

Our main calculations show that each algorithm’s empirical error does not dramatically change by altering one of the training examples. These calculations will be used within algorithmic stability analysis (Rogers and Wagner, 1978; Devroye and Wagner, 1979; Bousquet and Elisseeff, 2002). Stability bounds depend on how the space of functions is searched by the algorithm (rather than the size of the function space). There are many different ways to measure the stability of an algorithm; the bounds presented here use pointwise hypothesis stability so that the bounds scale correctly with the number of training examples  $m$ . For

simplicity, the algorithm recommends only one item,  $c = 1$ . Section 3.2 provides bounds for the large sample asymptotic regime where neither the minimum support threshold  $\theta$  nor the choice of  $K$  matters. Then we consider the new small  $m$  regime in Section 3.3, starting with a bound that formally shows that minimum support thresholds lead to better generalization. From there, we present a small sample bound for the adjusted confidence.

If the top recommendation has a higher adjusted confidence than the next item added, the algorithm incurs an error. (Even if that item is added later on, the algorithm incurs an error at this timestep.) To measure the size of that error, we can use a 0-1 loss, indicating whether or not our algorithm gave the highest adjusted confidence to the next item added. However, the 0-1 loss does not capture how close our algorithm was to correctly predicting the next item, though this information might be useful in determining how well the algorithm will generalize. We approximate the 0-1 loss using a modified loss that decays linearly near the discontinuity. This modified loss allows us to consider differences in adjusted confidence, not just whether one is larger than another:

$$|(\text{adjusted conf. of highest-scoring-correct rule}) - (\text{adjusted conf. of highest-scoring-incorrect rule})|. \quad (2)$$

However, as discussed in the introduction, if we adjust the loss function's  $K$  value to match the adjusted confidence  $K$  value, then we cannot fairly compare the algorithm's performance using two different values of  $K$ . An illustration of this point is that for large  $K$ , all adjusted confidence values are  $\ll 1$ , and for small  $K$ , then the adjusted confidence can be  $\approx 1$ ; differences in adjusted confidence for small  $K$  cannot be directly compared to those for large  $K$ . Since we want to directly compare performance as  $K$  is adjusted, we fix an evaluation measure that is separate from the choice of  $K$ . Specifically, we use the difference in adjusted confidence values with respect to a reference  $K_r$ :

$$|(\{\text{adjusted conf.}\}_{K_r} \text{ of highest-scoring-correct rule}_K) - (\{\text{adjusted conf.}\}_{K_r} \text{ of highest-scoring-incorrect rule}_K)|. \quad (3)$$

The reference  $K_r$  is a parameter of the loss function, whereas  $K$  is a parameter of an algorithm. We set  $K_r = 0$  to measure loss using the difference in confidence, and  $K = 0$  for an algorithm that chooses rules according to the confidence. As  $K$  gets farther from  $K_r$ , the algorithm is more distant from the way it is being evaluated, which leads to worse generalization.

### 3.1. Notation

We have a training set of  $m$  baskets  $S = \{z_i\}_{1 \dots m}$  that are the customers past orders. The baskets are chosen randomly (iid) from a fixed (but unknown) probability distribution  $\mathcal{D}$  over possible baskets. The generalization bound will be a guarantee on performance for a new randomly chosen basket. A basket  $z$  consists of an ordered (permuted) set of items,  $z \in 2^{\mathcal{X}} \times \Pi$ , where  $2^{\mathcal{X}}$  is the set of all subsets of  $\mathcal{X}$ , and  $\Pi$  is the set of permutations over at most  $|\mathcal{X}|$  elements. Denote  $z \sim \mathcal{D}$  to mean that basket  $z$  is drawn randomly (iid) according to distribution  $\mathcal{D}$  over the space of possible items in baskets and permutations over those items,  $2^{\mathcal{X}} \times \Pi$ . The  $t^{th}$  item added to the basket is written  $z_{\cdot,t}$ , where the dot is just a

placeholder for the generic basket  $z$ . The  $t^{th}$  element of the  $i^{th}$  basket in the training set is written  $z_{i,t}$ . We define the number of items in basket  $z$  by  $T_z$  and overload notation by defining  $T_i$  to be the number of items in the  $i^{th}$  training basket,  $T_i := |z_i|$ . Recall that  $\text{GenRules}$  produces only rules whose left-hand sides are in an allowed set  $A$ .

For the adjusted confidence algorithm, given a basket  $z$  and a particular time  $t$ , the algorithm uses the training set  $S$  to compute the adjusted confidences  $f_{S,K}$ . A *highest-scoring-correct* rule is a highest scoring rule that has the next item  $z_{\cdot,t+1}$  on the right. The left side  $a_{SztK}^+$  of a highest-scoring-correct rule obeys:

$$a_{SztK}^+ \in \operatorname{argmax}_{a \subseteq \{z_{\cdot,1}, \dots, z_{\cdot,t}\}, a \in A} f_{S,K}(a, z_{\cdot,t+1}).$$

(If  $z_{\cdot,t+1}$  has never been purchased, the adjusted confidence for all rules  $a \rightarrow z_{\cdot,t+1}$  is 0, and we choose the maximizing rule to be  $\emptyset \rightarrow z_{\cdot,t+1}$ .) A highest-scoring-correct rule correctly recommends the next item, and it is a best rule for the algorithm to choose.

The algorithm incurs an error when it recommends an incorrect item. A *highest-scoring-incorrect* rule is a highest scoring rule that does not have  $z_{\cdot,t+1}$  on the right. It is denoted  $a_{SztK}^- \rightarrow b_{SztK}^-$ , and obeys:

$$[a_{SztK}^-, b_{SztK}^-] \in \operatorname{argmax}_{\substack{a \subseteq \{z_{\cdot,1}, \dots, z_{\cdot,t}\}, a \in A \\ b \in \mathcal{X} \setminus \{z_{\cdot,1}, \dots, z_{\cdot,t+1}\}}} f_{S,K}(a, b).$$

If there is more than one highest-scoring rule, one is chosen at random. (With the exception that all incorrect rules are tied at zero adjusted confidence, in which case the left side is taken as  $\emptyset$  and the right side is chosen randomly). The adjusted confidence algorithm determines  $a_{SztK}^+$ ,  $a_{SztK}^-$ , and  $b_{SztK}^-$ , whereas nature chooses  $z_{\cdot,t+1}$ .

If the adjusted confidence of the rule  $a_{SztK}^- \rightarrow b_{SztK}^-$  is larger than that of  $a_{SztK}^+ \rightarrow z_{\cdot,t+1}$ , it means that the algorithm recommended the wrong item. The loss function below counts the proportion of times this happens for each basket, and is defined with respect to  $K_r$ .

$$\ell_{0-1,K_r}(f_{S,K}, z) := \frac{1}{T_z} \sum_{t=0}^{T_z-1} \begin{cases} 1 & \text{if } f_{S,K_r}(a_{SztK}^+, z_{\cdot,t+1}) - f_{S,K_r}(a_{SztK}^-, b_{SztK}^-) \leq 0 \\ 0 & \text{otherwise.} \end{cases}$$

We will now define the true error. The true error is an expectation of the loss function with respect to  $\mathcal{D}$ , and is a random variable since the training set  $S$  is random,  $S \sim \mathcal{D}^m$ .

$$\text{TrueErr}(f_{S,K}, K_r) := \mathbb{E}_{z \sim \mathcal{D}} \ell_{0-1,K_r}(f_{S,K}, z).$$

We upper bound the true error by using a different loss  $\ell_{\gamma,K_r}$  that is a continuous upper bound on the 0-1 loss  $\ell_{0-1,K_r}$ . It is defined with respect to  $K_r$  and another parameter  $\gamma > 0$  as follows:

$$\ell_{\gamma,K_r}(f_{S,K}, z) := \frac{1}{T_z} \sum_{t=0}^{T_z-1} c_\gamma(f_{S,K_r}(a_{SztK}^+, z_{\cdot,t+1}) - f_{S,K_r}(a_{SztK}^-, b_{SztK}^-)), \text{ where}$$

$$c_\gamma(y) = \begin{cases} 1 & \text{for } y \leq 0 \\ 1 - y/\gamma & \text{for } 0 \leq y \leq \gamma \\ 0 & \text{for } y \geq \gamma. \end{cases}$$

As  $\gamma$  approaches 0, this loss approaches the 0-1 loss. Also,  $\ell_{0-1,K_r}(f_{S,K}, z) \leq \ell_{\gamma,K_r}(f_{S,K}, z)$ . We define  $\text{TrueErr}_\gamma$  using this loss:

$$\text{TrueErr}_\gamma(f_{S,K}, K_r) := \mathbb{E}_{z \sim \mathcal{D}} \ell_{\gamma,K_r}(f_{S,K}, z),$$

where  $\text{TrueErr} \leq \text{TrueErr}_\gamma$ . The first set of results below bound  $\text{TrueErr}$  by considering the difference between  $\text{TrueErr}_\gamma$  and its empirical counterpart that we define next.

We overload notation by replacing  $z$  for a generic basket with  $i$  for a training basket. The left-hand side  $a_{SitK}^+$  of a highest-scoring-correct rule for basket  $z_i$  at time  $t$  obeys :

$$a_{SitK}^+ \in \operatorname{argmax}_{\substack{a \subseteq \{z_{i,1}, \dots, z_{i,t}\}, a \in A \\ b \in \mathcal{X} \setminus \{z_{i,1}, \dots, z_{i,t+1}\}}} f_{S,K}(a, z_{i,t+1}),$$

similarly, a highest-scoring-incorrect rule for  $z_i$  at time  $t$  has:

$$[a_{SitK}^-, b_{SitK}^-] \in \operatorname{argmax}_{\substack{a \subseteq \{z_{i,1}, \dots, z_{i,t}\}, a \in A \\ b \in \mathcal{X} \setminus \{z_{i,1}, \dots, z_{i,t+1}\}}} f_{S,K}(a, b).$$

The empirical error is an average of the loss over the training baskets:

$$\text{EmpErr}_\gamma(f_{S,K}, K_r) := \frac{1}{m} \sum_{\text{baskets } i=1}^m \ell_{\gamma,K_r}(f_{S,K}, z_i).$$

For the max confidence, min support algorithm, we substitute  $\theta$  where  $K$  appears in the notation. Again we use  $z$  when referring to a randomly drawn basket and  $i$  to refer to a specific training basket  $z_i$ . For instance, for training basket  $z_i$ , we define  $a_{Sit\theta}^+$ ,  $a_{Sit\theta}^-$ , and  $b_{Sit\theta}^-$  by:

$$a_{Sit\theta}^+ \in \operatorname{argmax}_{\substack{a \subseteq \{z_{i,1}, \dots, z_{i,t}\}, a \in A}} \bar{f}_{S,\theta}(a, z_{i,t+1}),$$

$$[a_{Sit\theta}^-, b_{Sit\theta}^-] \in \operatorname{argmax}_{\substack{a \subseteq \{z_{i,1}, \dots, z_{i,t}\}, a \in A \\ b \in \mathcal{X} \setminus \{z_{i,1}, \dots, z_{i,t+1}\}}} \bar{f}_{S,\theta}(a, b), \text{ and similarly,}$$

$$\ell_{0-1,K_r}(\bar{f}_{S,\theta}, z_i) := \frac{1}{T_i} \sum_{t=0}^{T_i-1} \begin{cases} 1 & \text{if } f_{S,K_r}(a_{Sit\theta}^+, z_{i,t+1}) - f_{S,K_r}(a_{Sit\theta}^-, b_{Sit\theta}^-) \leq 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\ell_{\gamma,K_r}(\bar{f}_{S,\theta}, z_i) := \frac{1}{T_i} \sum_{t=0}^{T_i-1} c_\gamma(f_{S,K_r}(a_{Sit\theta}^+, z_{i,t+1}) - f_{S,K_r}(a_{Sit\theta}^-, b_{Sit\theta}^-))$$

and  $\text{TrueErr}(\bar{f}_{S,\theta}, K_r)$  and  $\text{TrueErr}_\gamma(\bar{f}_{S,\theta}, K_r)$  are defined analogously as expectations of the losses, and  $\text{EmpErr}_\gamma(\bar{f}_{S,\theta}, K_r)$  is again an average of the loss over the training baskets.

### 3.2. Generalization Analysis for Large $m$

The choice of minimum support threshold  $\theta$  or the choice of parameter  $K$  matters mainly in the regime where  $m$  is small. For the max confidence, min support algorithm, when  $m$  is large, then all itemsets that would be chosen by the customer have appeared more times than the minimum support threshold with high probability. For the adjusted confidence algorithm, when  $m$  is large, prediction ability is guaranteed as follows.

**Theorem 1** (*Generalization Bound for Adjusted Confidence Algorithm, Large  $m$* )  
 For set of rules  $A$  and  $K \geq 0$ ,  $K_r \geq 0$ , with probability  $1 - \delta$  (with respect to training set  $S \sim \mathcal{D}^m$ ),

$$\text{TrueErr}(f_{S,K}, K_r) \leq \text{EmpErr}_\gamma(f_{S,K}, K_r) + \sqrt{\frac{1}{\delta} \left[ \frac{1}{2m} + 6\beta \right]}$$

$$\text{where } \beta = \frac{2|\mathcal{A}|}{\gamma} \left[ \frac{1}{(m-1)p_{\min A} + K} + \frac{|K_r - K| \frac{m}{m+K}}{(m-1)p_{\min A} + K_r} \right] + \mathcal{O}\left(\frac{1}{m^2}\right),$$

and where  $\mathcal{A} = \{a \in A : P_z(a \subseteq z) > 0\}$  are the itemsets that have some probability of being chosen. Out of these, any itemset that is the least likely to be chosen has probability  $p_{\min A}$ :

$$p_{\min A} := \min_{a \in \mathcal{A}} P_{z \sim \mathcal{D}}(a \subseteq z).$$

A special case is where  $K_r = K = 0$ : the algorithm chooses the rule with maximum confidence, and accuracy is then judged by the difference in confidence values between the highest-scoring-incorrect rule and the highest-scoring-correct rule. The expression reduces to:

**Corollary 2** (*Generalization Bound for Maximum Confidence Setting, Large  $m$* )  
 With probability  $1 - \delta$  (with respect to  $S \sim \mathcal{D}^m$ ),

$$\text{TrueErr}(f_{S,0}, 0) \leq \text{EmpErr}_\gamma(f_{S,0}, 0) + \sqrt{\frac{1}{\delta} \left[ \frac{1}{2m} + \frac{12|\mathcal{A}|}{\gamma(m-1)p_{\min A}} \right]} + \mathcal{O}\left(\frac{1}{m^2}\right).$$

The use of the pointwise hypothesis stability within this proof is the key to providing a decay of order  $\sqrt{(1/m)}$ . Now that this bound is established, we move to the small sample case, where the minimum support is the force that provides generalization.

### 3.3. Generalization Analysis for Small $m$

The first small sample result is a general bound for the max confidence, min support algorithm, that is, Algorithm 2.

**Theorem 3** (*Generalization Bound for Max Confidence, Min Support*)  
 For  $\theta \geq 1$ ,  $K_r \geq 0$ , with probability  $1 - \delta$  (with respect to  $S \sim \mathcal{D}^m$ ),  $m > \theta$ ,

$$\text{TrueErr}(\bar{f}_{S,\theta}, K_r) \leq \text{EmpErr}_\gamma(\bar{f}_{S,\theta}, K_r) + \sqrt{\frac{1}{\delta} \left[ \frac{1}{2m} + 6\beta \right]}$$

$$\text{where } \beta = \frac{2}{\gamma} \left[ \frac{1}{\theta} + K_r \left( \frac{1}{\theta + K_r} \right) \left( 1 + \frac{1}{\theta} \right) \right].$$

Figure 1 shows  $\beta$  as a function of  $\theta$  for several different values of  $K_r$ . The special case of interest is when  $K_r = 0$ , so that the loss is judged with respect to differences in confidence, as follows:

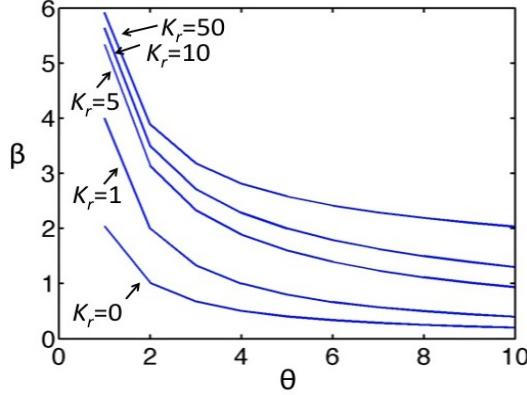


Figure 1:  $\beta$  vs.  $\theta$  from Theorem 3, with  $\gamma = 1$ . The different curves are different values of  $K_r = 0, 1, 5, 10, 50$  from bottom to top.

**Corollary 4** (*Generalization Bound for Max Confidence, Min Support,  $K_r = 0$* )  
 For  $\theta \geq 1$ , with probability  $1 - \delta$  (with respect to  $S \sim \mathcal{D}^m$ ),  $m > \theta$ ,

$$\text{TrueErr}(\bar{f}_{S,\theta}, 0) \leq \text{EmpErr}_\gamma(\bar{f}_{S,\theta}, 0) + \sqrt{\frac{1}{\delta} \left[ \frac{1}{2m} + \frac{12}{\theta\gamma} \right]}.$$

It is common to use a minimum support threshold that is a fraction of  $m$ , for instance,  $\theta = 0.1 \times m$ . In that case, the bound again scales with  $\sqrt{(1/m)}$ . Note that there is no generalization guarantee when  $\theta = 0$ ; the minimum support threshold enables generalization in the small  $m$  case.

Now we discuss the adjusted confidence algorithm for small  $m$  setting. In the proof of the following theorem, if we were to use the definitions established above, the bound does not simplify beyond a certain point and is difficult to read at an intuitive level. From that bound, it would not be easy to see what are the important quantities for the learning process, and how they scale. In what follows, we redefine the loss function slightly, so that it approximates a 0-1 loss from below instead of from above. This provides a concise and intuitive bound. Almost the same proof can be used to create both versions of the bound, only the last steps are different.

Define a *highest-scoring* rule  $a_{SztK}^* \rightarrow b_{SztK}^*$  as a rule that achieves the maximum adjusted confidence, over all of the possible rules. It will either be equal to  $a_{SztK}^+ \rightarrow z_{\cdot,t+1}$  or  $a_{SztK}^- \rightarrow b_{SztK}^-$ , depending on which has the larger adjusted confidence:

$$[a_{SztK}^*, b_{SztK}^*] \in \underset{\substack{a \subseteq \{z_{\cdot,1}, \dots, z_{\cdot,t}\}, a \in A \\ b \in \mathcal{X} \setminus \{z_{\cdot,1}, \dots, z_{\cdot,t}\}}}{\operatorname{argmax}} f_{S,K}(a, b).$$

Note that  $b_{SztK}^*$  can be equal to  $z_{\cdot,t+1}$  whereas  $b_{SztK}^-$  cannot. The notation for  $a_{SitK}^*$  and  $b_{SitK}^*$  is similar, and the new loss is:

$$\ell_{0-1,K_r}^{\text{new}}(f_{S,K}, z) := \frac{1}{T_z} \sum_{t=0}^{T_z-1} \begin{cases} 1 & \text{if } f_{S,K_r}(a_{SztK}^+, z_{\cdot,t+1}) - f_{S,K_r}(a_{SztK}^*, b_{SztK}^*) < 0 \\ 0 & \text{otherwise.} \end{cases}$$

By definition, the difference  $f_{S,K_r}(a_{SztK}^+, z_{\cdot,t+1}) - f_{S,K_r}(a_{SztK}^*, b_{SztK}^*)$  can never be strictly positive. The continuous approximation is:

$$\ell_{\gamma, K_r}^{\text{new}}(f_{S,K}, z) := \frac{1}{T_z} \sum_{t=0}^{T_z-1} c_{\gamma}^{\text{new}}(f_{S,K_r}(a_{SztK}^+, z_{\cdot,t+1}) - f_{S,K_r}(a_{SztK}^*, b_{SztK}^*)), \text{ where}$$

$$c_{\gamma}^{\text{new}}(y) = \begin{cases} 1 & \text{for } y \leq -\gamma \\ -y/\gamma & \text{for } -\gamma \leq y \leq 0 \\ 0 & \text{for } y \geq 0. \end{cases}$$

As  $\gamma$  approaches 0, the  $c_{\gamma}$  loss approaches the 0-1 loss. We define  $\text{TrueErr}_{\gamma}^{\text{new}}$  and  $\text{EmpErr}_{\gamma}^{\text{new}}$ :

$$\text{TrueErr}_{\gamma}^{\text{new}}(f_{S,K}, K_r) := \mathbb{E}_{z \sim \mathcal{D}} \ell_{\gamma, K_r}^{\text{new}}(f_{S,K}, z),$$

$$\text{EmpErr}_{\gamma}^{\text{new}}(f_{S,K}, K_r) := \frac{1}{m} \sum_{i=1}^m \ell_{\gamma, K_r}^{\text{new}}(f_{S,K}, z_i).$$

The minimum support threshold condition we used earlier is replaced by a weaker condition on the support. This weaker condition has the benefit of allowing more rules to be used in order to achieve a better empirical error; however, it is more difficult to get a generalization guarantee. This support condition is derived from the fact that the adjusted confidence of the highest-scoring rule  $a_{SitK}^* \rightarrow b_{SitK}^*$  exceeds that of the highest-scoring-correct rule  $a_{SitK}^+ \rightarrow z_{i,t+1}$ , which exceeds that of the marginal rule  $\emptyset \rightarrow z_{i,t+1}$ :

$$\frac{\#a_{SitK}^*}{\#a_{SitK}^* + K} \geq \frac{\#(a_{SitK}^* \cup b_{SitK}^*)}{\#a_{SitK}^* + K} \geq \frac{\#(a_{SitK}^+ \cup z_{i,t+1})}{\#a_{SitK}^+ + K} \geq \frac{\#z_{i,t+1}}{m + K}. \quad (4)$$

This leads to a lower bound on the support  $\#a_{SitK}^*$ :

$$\#a_{SitK}^* \geq K \left( \frac{\#z_{i,t+1}}{m + K - \#z_{i,t+1}} \right). \quad (5)$$

This is not a hard minimum support threshold, yet since the support generally increases as  $K$  increases, the bound will give a better guarantee for large  $K$ . Note that in the original notation, we would replace the condition (4) with  $\frac{\#a_{SitK}^-}{\#a_{SitK}^- + K} \geq \frac{\#(a_{SitK}^- \cup b_{SitK}^-)}{\#a_{SitK}^- + K} \geq \frac{\#b_{SitK}^-}{m + K}$  and proceed with analogous steps in the proof.

**Theorem 5** (*Generalization Bound for Adjusted Confidence Algorithm, Small  $m$* )

$$\begin{aligned} \text{TrueErr}_{\gamma}^{\text{new}}(f_{S,K}, K_r) &\leq \text{EmpErr}_{\gamma}^{\text{new}}(f_{S,K}, K_r) + \sqrt{\frac{1}{\delta} \left[ \frac{1}{2m} + 6\beta \right]} \text{ where} \\ \beta &= \frac{2}{\gamma} \frac{1}{K} \left( 1 - \frac{(m-1)p_{\min}}{m+K} \right) \\ &\quad + \frac{2}{\gamma} |K_r - K| \mathbb{E}_{\zeta \sim \text{Bin}(m-1, p_{\min})} \frac{1}{K \left( \frac{\zeta}{m+K-\zeta-1} \right) + K_r} \left( \frac{m}{m+K} + \frac{1}{K} \left( 1 - \frac{\zeta}{m+K} \right) \right), \end{aligned}$$

and where  $Q = \{x \in \mathcal{X} : P_{z \sim \mathcal{D}}(x \in z) > 0\}$  are the items that have some probability of being chosen by the customer. Out of these, any item that is the least likely to be chosen has probability  $p_{\min}$ :

$$p_{\min} := \min_{x \in Q} P_{z \sim \mathcal{D}}(x \in z).$$

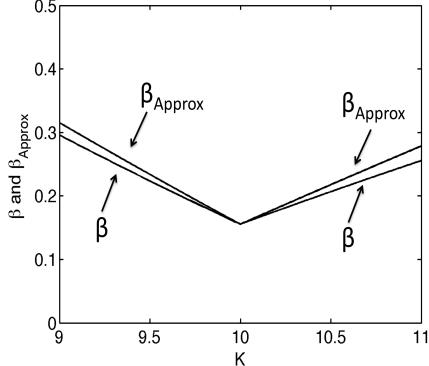


Figure 2:  $\beta$  and  $\beta_{\text{Approx}}$  vs  $K$ , where  $K_r = 10$ ,  $p_{\min} = 0.3$ ,  $m = 20$ ,  $\gamma = 1$ .

The stability  $\beta$  has two main terms. The first term decreases generally as  $1/K$ . The second term arises from the error in measuring loss with  $K_r$  rather than  $K$ . In order to interpret  $\beta$ , consider the following approximation to the expectation in the bound, which assumes that  $m$  is large and that  $m \gg K \gg 0$ , and that  $\zeta \approx mp_{\min}$ :

$$\beta \approx \frac{2}{\gamma} \frac{1}{K} \left( 1 - \frac{(m-1)p_{\min}}{m+K} \right) + \frac{2}{\gamma} |K_r - K| \frac{1}{K \frac{p_{\min}}{1-p_{\min}} + K_r}. \quad (6)$$

Intuitively, if either  $K$  is close to  $K_r$  or  $p_{\min}$  is large (close to 1) then this term becomes small. Figure 2 shows an example plot of  $\beta$  and the approximation using (6), which we denote by  $\beta_{\text{Approx}}$ .

One can observe that if  $K_r > K$ , then both terms tend to improve (decrease) with increasing  $K$ . When  $K_r < K$ , then the two terms can compete as  $K$  increases.

### 3.4. Summary of Bounds

We have provided probabilistic guarantees on performance that show the following: 1) For large  $m$ , the association rule-based algorithms for sequential event prediction have a performance guarantee of the same order as for classical supervised learning problems (classification, ranking, regression, density estimation). 2) For small  $m$ , the minimum support threshold guarantees generalization (at the expense of removing important rules). 3) The adjusted confidence provides a much weaker support threshold, allowing important rules to be used, while still being able to generalize. 4) All generalization guarantees depend on the way the goodness of the algorithm is measured (the choice of  $K_r$  in the loss function). There are two terms in the small sample size bound for the adjusted confidence. Generally, one term decreases (becomes better) as  $K$  increases, and the other term decreases as  $K$  gets closer to  $K_r$ .

## 4. Experiments

All datasets chosen for these experiments are publicly available from the UCI machine learning repository (Frank and Asuncion, 2010), and from the IBM Quest Market-Basket Synthetic Data Generator (Agrawal and Srikant, 1994). To obtain formatted market-basket

- 
- Input:**  $(S, B, \mathcal{X})$ , that is, past orders  $S = \{z_i\}_{i=1,\dots,m}$ ,  $z_i \subseteq \mathcal{X}$ , current basket  $B \subset \mathcal{X}$ , set of items  $\mathcal{X}$
- Output:** Set of all rules where  $a_j$  is an item in the basket  $B$  (or the empty set) and  $b_j$  is not in  $B$ . That is, rules  $\{a_j \rightarrow b_j\}_j$  such that  $b_j \in \mathcal{X} \setminus B$  and either  $a_j \in B$  or  $a_j = \emptyset$ .
- 

**Algorithm 4:** Subroutine *GenRules*, simplest version that considers only “marginal” rules.

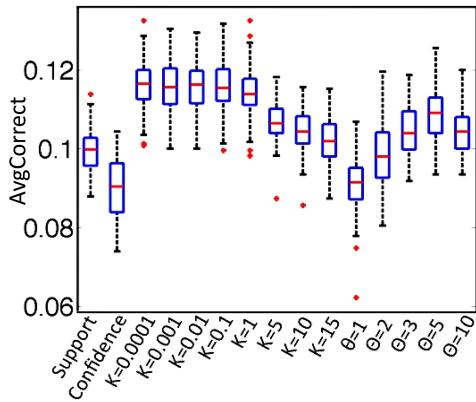


Figure 3: Boxplots of AvgCorrect values for Mushroom dataset.

Algorithm	mean $\pm$ standard dev.
Support	$0.0996 \pm 0.0051$
Confidence	$0.0902 \pm 0.0075$
$K=0.0001$	<b><math>0.1164 \pm 0.0061</math></b>
$K=0.001$	<b><math>0.1158 \pm 0.0062</math></b>
$K=0.01$	<b><math>0.1161 \pm 0.0061</math></b>
$K=0.1$	<b><math>0.116 \pm 0.0058</math></b>
$K=1$	$0.1142 \pm 0.0062$
$K=5$	$0.1069 \pm 0.0052$
$K=10$	$0.1044 \pm 0.0054$
$K=15$	$0.1024 \pm 0.0053$
$\theta=1$	$0.0909 \pm 0.007$
$\theta=2$	$0.0986 \pm 0.0077$
$\theta=3$	$0.1048 \pm 0.0064$
$\theta=5$	$0.1088 \pm 0.0069$
$\theta=10$	$0.1042 \pm 0.0057$

Figure 4: Means and standard deviations for Mushroom dataset. Bold indicates no significant difference from the best algorithm.

data, categorical data were converted into binary features (one feature per category). Each feature represents an item, and each example represents a basket. The feature value (0 or 1) indicates the presence of an item. Training baskets and test baskets were chosen randomly without replacement from the full dataset. Since these data do not come naturally with a time ordering, items in the basket were randomly permuted to attain an order. At each iteration, rules were formed from one item or the empty item on the left, and one item on the right (See *GenRules* as Algorithm 4). Recommendations of one item were made using the following 15 algorithms: highest support, highest confidence, highest adjusted confidence for eight  $K$  levels, max confidence, min support algorithm for five support threshold levels  $\theta$ . All 15 algorithms were evaluated by the average fraction of correct recommendations (AvgCorrect) per basket. As recommendations were made, it was common to have ties where multiple items are equally good to recommend, in which case the tie was broken at random; AvgCorrect is similar to  $\ell_{0-1,K}$  except for this way of dealing with ties.

The parameters of the experiment are: number of training baskets (20 in all cases), number of test baskets (100 in all cases), values of  $K$  for the adjusted confidence algorithm (0.0001, 0.001, 0.01, 0.1, 1, 5, 10, 15), and values of  $\theta$  for the max confidence, min support algorithm (1, 2, 3, 5, 10). Note that two of these algorithms are the same: the

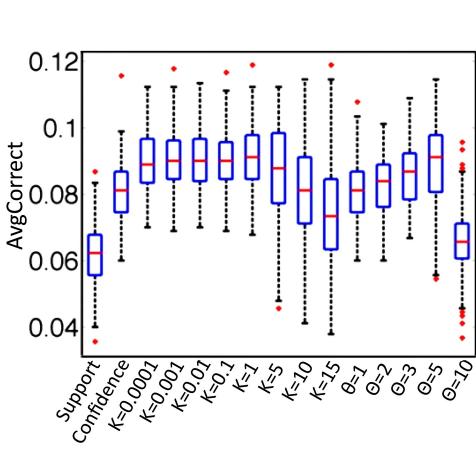


Figure 5: Boxplots of AvgCorrect values for Nursery dataset.

Algorithm	mean $\pm$ standard dev.
Support	0.0619 $\pm$ 0.0098
Confidence	0.081 $\pm$ 0.0094
$K=0.0001$	0.0898 $\pm$ 0.0091
$K=0.001$	<b>0.0902 <math>\pm</math> 0.0093</b>
$K=0.01$	<b>0.0902 <math>\pm</math> 0.0085</b>
$K=0.1$	<b>0.0903 <math>\pm</math> 0.0095</b>
$K=1$	<b>0.0909 <math>\pm</math> 0.0096</b>
$K=5$	0.0869 $\pm$ 0.0139
$K=10$	0.0804 $\pm$ 0.0154
$K=15$	0.0747 $\pm$ 0.0154
$\theta=1$	0.0811 $\pm$ 0.0088
$\theta=2$	0.0819 $\pm$ 0.0094
$\theta=3$	0.0858 $\pm$ 0.0095
$\theta=5$	0.0883 $\pm$ 0.0137
$\theta=10$	0.0654 $\pm$ 0.0111

Figure 6: Means and standard deviations for Nursery dataset.

max confidence algorithm is the same as the max confidence, min support algorithm for  $\theta=1$ . Datasets are: Car Evaluation (25 items, 1728 baskets), Chess King-Rook vs King-Pawn, (75 items, 3196 baskets), MONK’s problems (19 items, 1711 baskets) Mushroom (119 items, 8124 baskets), Nursery (32 items, 12960 baskets), Plants (70 items, 34781 baskets), T20I18D10KN22CR50 (22 items, 10000 baskets). Each experiment (training, test, evaluation for all 15 algorithms) was performed 100 times, (totaling  $100 \times 100 \times 15 = 150,000$  test basket evaluations per dataset, for each of 7 datasets). In Figures 3 through 6, the distribution of AvgCorrect values for each algorithm on datasets Mushroom and Nursery is shown via boxplot, along with the mean and standard deviation of AvgCorrect values for each algorithm. Bold indicates that the mean is not significantly different from that of the algorithm with the largest mean value; that is, bold indicates the highest scores. Similar experimental results for the other datasets are in the longer version (Rudin et al., 2011).

Figure 7 summarizes the results of all of the experiments by totaling the number of datasets for which each algorithm achieved one of the highest scores. The best performing algorithms were  $K = 0.01$  and  $K = 0.1$ , both algorithms achieving one of the top scores for 6 out of 7 of the datasets. The single dataset for which these algorithms did not achieve one the best scores was the very dense dataset T20I18D10KN22CR50, where the algorithms requiring a higher support (the max support algorithm, and also the adjusted confidence algorithm for  $K = 5, 10$ , and  $15$ ) achieved the highest AvgCorrect score. In that case, the  $K = 0.01$  and  $K = 0.1$  algorithms still performed better than the max confidence, min support algorithms for the parameters we tried.

The adjusted confidence algorithm with a very small  $K$  is similar to using the max confidence algorithm, except that whenever there is a tie, the tie is broken in favor of the rule with largest support. It seems that in most of the datasets we chose, this type of algorithm performed the best, which indicates two things. First, that for some datasets, increasing  $K$  too much can have the same effect as a too-large minimum support threshold: large values of  $K$  could potentially remove the best rules, leading to too much bias, where

the algorithm cannot explain enough of the variance in the data. Second, when comparing rules, it is important not to break ties at random as in the max confidence, min support algorithm, but instead to use the support of the rules. Another observation is that the performance levels of the adjusted confidence algorithm vary less than those of the max confidence, min support algorithm. In other words, our experiments indicate that a less-than-perfect choice of  $K$  for the adjusted confidence algorithm is likely to perform better than a less-than-perfect choice of  $\theta$  for the max confidence, min support algorithm.

## 5. Related Work and Ongoing Work

The usefulness of association rules and their impact on even a wider range of practical applications remains limited due to problems arising from the minimum support threshold. Most prior work relies on this strong requirement; exceptions include works of Li et al. (1999); Koh (2008) and DuMouchel and Pregibon (2001). Some work (Cohen et al., 2001; Wang et al., 2001) aims to find high confidence rules, ignoring the support altogether. Association rules are generally used as an exploratory tool rather than a predictive tool, which is in contrast with our work. On the other hand, it is clearly possible to use the adjusted confidence as an “interestingness” measure for database exploration. Lin et al. (2002) also construct a recommender system using rules, having a minimum confidence threshold and then an adjustable minimum support threshold. Lawrence et al. (2001) provide a recommender system for a grocery store, but the setting differs from ours in that they always recommend items that have never been previously purchased.

In terms of Bayesian analysis, DuMouchel and Pregibon (2001) present a Bayesian approach to the identification of interesting itemsets. While not a rule mining algorithm per se, the approach could be extended to produce rules. Breese et al. (1998) present a number of different algorithms for collaborative filtering, including two Bayesian approaches. One of their Bayesian approaches clusters users while the other constructs a Bayesian network. Condliff et al. (1999) present a hierarchical Bayesian approach to collaborative filtering that “borrows strength” across users. Neither Breese et al. nor Condliff et al. focus on repeated purchases but both present ideas that may have relevance to future versions of our approach.

In current work, we are designing a Bayesian framework that estimates  $K$  for the adjusted confidence by “borrowing strength” across both users and items (McCormick et al., 2011). We are also looking at different approaches to the online grocery store problem,

Algorithm	No. datasets
Support	1
Confidence	1
$K=0.0001$	4
$K=0.001$	5
$K=0.01$	6
$K=0.1$	6
$K=1$	2
$K=5$	2
$K=10$	2
$K=15$	2
$\theta=1$	1
$\theta=2$	1
$\theta=3$	1
$\theta=5$	0
$\theta=10$	1

Figure 7: Summary of experiments: for each algorithm, the number of datasets where it performed comparably with the best algorithm.

where we allow the predictions to alter the sequence in which items are placed into the basket (Letham et al., 2011).

## 6. Conclusion

This work synthesizes tools from several fields to analyze association rules in a supervised learning framework. This analysis is necessarily different from that of classical supervised learning analysis; association rules provide two mechanisms for generalization: a large sample, and a minimum support of rules. We considered two simple algorithms, both that create a bound on the support, regulating a tradeoff between accuracy on the training set and generalization ability. We have also demonstrated that the adjusted confidence introduced here has several advantages over the minimum support threshold that is commonly considered in association rule mining.

## Acknowledgments

C. Rudin is also at the Center for Computational Learning Systems, Columbia University. This work was performed partly while E. Kogan was at Fresh Direct. We would like to acknowledge support for this project from the National Science Foundation under grant IIS-1053407.

## References

- Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proc. 20<sup>th</sup> Int'l Conf. Very Large Data Bases, (VLDB)*, pages 487–499. Morgan Kaufmann, 1994.
- Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD Int'l Conf. on Management of Data*, pages 207–216, 1993.
- Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.
- John S. Breese, David Heckerman, and Carl Myers Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. Uncertainty in Artificial Intelligence (UAI)*, pages 43–52, 1998.
- Edith Cohen, Mayur Datar, Shinji Fujiwara, Aristides Gionis, Piotr Indyk, Rajeev Motwani, Jeffrey D. Ullman, and Cheng Yang. Finding interesting associations without support pruning. *IEEE Trans. Knowl. Data Eng.*, 13(1):64–78, 2001.
- Michelle Keim Condliff, David D. Lewis, David Madigan, and Christian Posse. Bayesian mixed-effects models for recommender systems. In *ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation*, 1999.
- Luc Devroye and T. J. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.

- William DuMouchel and Daryl Pregibon. Empirical bayes screening for multi-item associations. In *Proc. ACM SIGKDD Int'l Conf. on Knowl. Discovery and Data Mining*, pages 67–76, 2001.
- A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL <http://archive.ics.uci.edu/ml>.
- Xiangji Huang, Aijun An, Nick Cercone, and Gary Promhouse. Discovery of interesting association rules from Livelink web log data. In *Proc. IEEE Int'l Conf. on Data Mining (ICDM)*, 2002.
- Xiang-Rong Jiang and Le Gruenwald. Microarray gene expression data association rules mining based on BSC-tree and FIS-tree. *Data & Knowl. Eng.*, 53(1):3–29, 2005.
- Yun Sing Koh. Mining non-coincidental rules without a user defined support threshold. In *Advances in Knowl. Discovery and Data Mining, 12<sup>th</sup> Pacific-Asia Conf., (PAKDD)*, pages 910–915, 2008.
- Ron Kohavi, Llew Mason, Rajesh Parekh, and Zijian Zheng. Lessons and challenges from mining retail e-commerce data. *Machine Learning*, 57(1-2):83–113, 2004.
- R.D. Lawrence, G.S. Almasi, V. Kotlyar, M.S. Viveros, and S.S. Duri. Personalization of supermarket product recommendations. *Data Mining and Knowledge Discovery*, 5(1-2): 11–32, 2001.
- Ben Letham, Cynthia Rudin, and David Madigan. A supervised ranking approach to sequential event prediction. In Preparation, 2011.
- Jinyan Li, Xiuzhen Zhang, Guozhu Dong, Kotagiri Ramamohanarao, and Qun Sun. Efficient mining of high confidence association rules without support thresholds. In *Proc. Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 406–411, 1999.
- Weiyang Lin, Sergio A. Alvarez, and Carolina Ruiz. Efficient adaptive-support association rule mining for recommender systems. *Data Mining and Knowledge Discovery*, 6(1): 83–105, 2002.
- Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In *Proceedings of the 4<sup>th</sup> International Conference on Knowledge Discovery and Data Mining (KDD)*, 1998.
- Tyler McCormick, Cynthia Rudin, and David Madigan. A hierarchical model for association rule mining of sequential events: An approach to automated medical symptom prediction. *SSRN eLibrary*, 2011. URL <http://ssrn.com/paper=1736062>.
- W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *Annals of Statistics*, 6(3):506–514, 1978.
- Cynthia Rudin, Benjamin Letham, Eugene Kogan, and David Madigan. A learning theory framework for association rules and sequential events. In Preparation, 2011.

Adriano Veloso, Humberto Mossri de Almeida, Marcos André Gonçalves, and Wagner Meira Jr. Learning to rank at query-time using association rules. In *Proc. Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 267–274, 2008.

Ke Wang, Yu He, David W. Cheung, and Francis Y. L. Chin. Mining confident rules without support requirement. In *Proc. Conf. on Information and Knowledge Management (CIKM)*, pages 89–96, 2001.

# Optimal aggregation of affine estimators

**Joseph Salmon**

*LPMA*

*Université Paris Diderot Paris 7*

NAME@MATH.JUSSIEU.FR

**Arnak Dalalyan**

*Université Paris Est / ENPC*

*LIGM/ IMAGINE*

NAME@IMAGINE.ENPC.FR

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We consider the problem of combining a (possibly uncountably infinite) set of affine estimators in non-parametric regression model with heteroscedastic Gaussian noise. Focusing on the exponentially weighted aggregate, we prove a PAC-Bayesian type inequality that leads to sharp oracle inequalities in discrete but also in continuous settings. The framework is general enough to cover the combinations of various procedures such as least square regression, kernel ridge regression, shrinking estimators and many other estimators used in the literature on statistical inverse problems. As a consequence, we show that the proposed aggregate provides an adaptive estimator in the exact minimax sense without neither discretizing the range of tuning parameters nor splitting the set of observations. We also illustrate numerically the good performance achieved by the exponentially weighted aggregate.

**Keywords:** List of keywords

## 1. Introduction

There is a growing empirical evidence of superiority of aggregated statistical procedures, also referred to as *blending*, *stacked generalization*, or *ensemble methods*, with respect to “pure” ones. Since their introduction in the 1990’s, the most famous aggregation procedures such as *Boosting* (Freund, 1990), *Bagging* (Breiman, 1996) or *Random Forest* (Amit and Geman, 1997) were successfully used in practice for a large variety of applications. Moreover, the most recent Machine Learning competitions such as the Pascal VOC or the Netflix challenge were won by procedures combining different types of classifiers / predictors / estimators. It is therefore of central interest to understand from a theoretical point of view what kind of aggregation strategies should be used for getting the best possible combination of the available statistical procedures.

### 1.1. Historical remarks and motivation

In the statistical literature, to the best of our knowledge, the lecture notes of Nemirovski (2000) was the first work concerned by the theoretical analysis of aggregation procedures. It was followed by a paper by Juditsky and Nemirovski (2000), as well as by a series of papers by Catoni (see Catoni (2004) for a comprehensive account) and Yang (2000, 2003, 2004). For the regression model, a significant progress was achieved by Tsybakov (2003) with introducing the no-

tion of optimal rates of aggregation and proposing aggregation-rate-optimal procedures for the tasks of linear, convex and model selection aggregation. This point was further developed by Lounici (2007); Rigollet and Tsybakov (2007); Lecué (2007); Bunea et al. (2007), especially in the context of high dimension with sparsity constraints.

From a practical point of view, an important limitation of the previously cited results on the aggregation is that they are valid under the assumption that the aggregated procedures are deterministic (or random, but independent of the data used for the aggregation). In the Gaussian sequence model, a breakthrough was reached by Leung and Barron (2006). Building on a very elegant but not very well known result of George (1986), they established sharp oracle inequalities for the exponentially weighted aggregate (EWA) under the condition that the aggregated estimators are obtained from the data vector by orthogonally projecting it on some linear subspaces. Dalalyan and Tsybakov (2007, 2008), established the validity of Leung and Barron's result under more general (non Gaussian) noise distributions provided that the constituent estimators are independent of the data used for the aggregation. A natural question arises whether a similar result can be proved for a larger family of constituent estimators containing projection estimators and deterministic ones as specific examples. The main aim of the present paper is to answer this question by considering families of affine estimators.

Our interest in affine estimators is motivated by several reasons. First of all, affine estimators encompass many popular estimators such as the smoothing splines, the Pinsker estimator Pinsker (1980); Efromovich and Pinsker (1996), the local polynomial estimators, the non-local means Buades et al. (2005); Salmon and Le Pennec (2009), etc. For instance, it is known that if the underlying (unobserved) signal belongs to a Sobolev ball, then the (linear) Pinsker estimator is asymptotically minimax up to the optimal constant, while the best projection estimator is only rate-minimax. A second motivation is that—as proved by Juditsky and Nemirovski (2009)—the set of signals that are well estimated by linear estimators is very rich. It contains, for instance, sampled smooth functions, sampled modulated smooth functions and sampled harmonic functions (cf. Juditsky and Nemirovski (2009) for precise definitions). It is worth noting that oracle inequalities for the penalized empirical risk minimizer were also established by Golubev (2010), and for the model selection by Arlot and Bach (2009); Baraud et al. (2010).

In the present work, we establish sharp oracle inequalities in the statistical model of heteroscedastic regression, under various conditions on the constituent estimators assumed to be affine functions of the data. We assume that the design is deterministic and that the noise is Gaussian with a given covariance matrix. Our results provide theoretical guarantees of optimality, in terms of the expected loss, for the exponentially weighted aggregate. They have the advantage of covering in a unified fashion the particular cases of deterministic estimators considered by Dalalyan and Tsybakov (2008) and of projection estimators treated by Leung and Barron (2006).

## 1.2. Notation

Throughout this work, we focus on the heteroscedastic regression model with Gaussian additive noise. More precisely, we assume that we are given a vector  $\mathbf{Y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$  obeying the model:

$$y_i = f_i + \xi_i, \quad \text{for } i = 1, \dots, n, \tag{1}$$

where  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\top$  is a centered Gaussian random vector,  $f_i = \mathbf{f}(x_i)$  where  $\mathbf{f}$  is an unknown function  $\mathcal{X} \rightarrow \mathbb{R}$  and  $x_1, \dots, x_n \in \mathcal{X}$  are deterministic points. Here, no assumption is made on the set  $\mathcal{X}$ . Our objective is to recover the vector  $\mathbf{f} = (f_1, \dots, f_n)$ , often referred to as *signal*, based on the data  $y_1, \dots, y_n$ . In our work, the noise covariance matrix  $\Sigma = \mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top]$  is assumed to be diagonal (so it can be written  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ ), with a known upper bound on the spectral norm  $\|\Sigma\|$ . In our case,  $\|\Sigma\| = \max_{i=1, \dots, n} \sigma_i^2$ . We measure the performance of an estimator  $\hat{\mathbf{f}}$  by its expected empirical quadratic loss:  $r = \mathbb{E}(\|\mathbf{f} - \hat{\mathbf{f}}\|_n^2)$  where  $\|\mathbf{f} - \hat{\mathbf{f}}\|_n^2 = \frac{1}{n} \sum_{i=1}^n (f_i - \hat{f}_i)^2$ . We also denote by  $\langle \cdot | \cdot \rangle_n$  the corresponding empirical inner product.

In this paper, we only focus on *affine estimators*  $\hat{\mathbf{f}}_\lambda$ , *i.e.*, estimators that can be written as affine transforms of the data  $\mathbf{Y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ . Using the convention that all vectors are one-column matrices, affine estimators can be defined by

$$\hat{\mathbf{f}}_\lambda = A_\lambda \mathbf{Y} + \mathbf{b}_\lambda, \quad (2)$$

where the  $n \times n$  real matrix  $A_\lambda$  and the vector  $\mathbf{b}_\lambda \in \mathbb{R}^n$  are deterministic. This means that the entries of  $A_\lambda$  and  $\mathbf{b}_\lambda$  may depend on the points  $x_1, \dots, x_n$  but not on the data vector  $\mathbf{Y}$ . It is well-known that the quadratic risk of the estimator (2) is given by

$$r_\lambda = \mathbb{E}(\|\mathbf{f} - \hat{\mathbf{f}}_\lambda\|_n^2) = \|(A_\lambda - I_{n \times n})\mathbf{f} + \mathbf{b}_\lambda\|_n^2 + \frac{\text{Tr}(A_\lambda \Sigma A_\lambda^\top)}{n} \quad (3)$$

and that  $\hat{r}_\lambda$ , defined by

$$\hat{r}_\lambda = \|\mathbf{Y} - \hat{\mathbf{f}}_\lambda\|_n^2 + \frac{2}{n} \text{Tr}(\Sigma A_\lambda) - \frac{1}{n} \sum_{i=1}^n \sigma_i^2 \quad (4)$$

is an unbiased estimator of  $r_\lambda$  (direct application of Stein's Lemma, cf. Appendix).

Let us describe now different families of linear and affine estimators successfully used in the statistical literature (cf., for instance, Arlot and Bach (2009)). Our results apply to all these families and lead to a procedure that behaves nearly as well as the best one of the family.

**Ordinary least squares** Let  $\{\mathcal{S}_\lambda : \lambda \in \Lambda\}$  be a set of linear subspaces of  $\mathbb{R}^n$ . A well known family of affine estimators, successfully used in the context of model selection by Barron et al. (1999), is the set of orthogonal projections onto  $\mathcal{S}_\lambda$ . In the case of a family of linear regression models with design matrices  $X_\lambda$ , one has  $A_\lambda = X_\lambda (X_\lambda^\top X_\lambda)^{-1} X_\lambda^\top$ .

**Diagonal filters** Another set of common estimators are the so called diagonal filters  $\hat{\mathbf{f}} = A\mathbf{Y}$ , where  $A$  is a diagonal matrix  $A = \text{diag}(a_1, \dots, a_n)$ . Popular examples include:

- Ordered projections :  $a_k = \mathbb{1}_{(k \leq \lambda)}$  for some integer  $\lambda$  (where  $\mathbb{1}_{(\cdot)}$  is the indicator function). Those weights are also called truncated SVD or spectral cut-off. In this case the natural parametrization is  $\Lambda = \{1, \dots, n\}$ , indexing the number of elements conserved.
- Block projections:  $a_k = \mathbb{1}_{(k \leq w_1)} + \sum_{j=1}^{m-1} \lambda_j \mathbb{1}_{(w_j \leq k \leq w_{j+1})}$ ,  $k = 1, \dots, n$ , where  $\lambda_j \in \{0, 1\}$ . Here the natural parametrization is  $\Lambda = \{0, 1\}^{m-1}$ , indexing subsets of  $\{1, m-1\}$ .

- Tikhonov-Philipps filter:  $a_k = \frac{1}{1+(k/w)^\alpha}$ , where  $w, \alpha > 0$ . The set  $\Lambda = (\mathbb{R}_+^*)^2$  indexes continuously the smoothing parameters.
- Pinsker filter:  $a_k = (1 - \frac{k^\alpha}{w})_+$ , where  $x_+ = \max(x, 0)$  and  $w, \alpha > 0$ . In this case also  $\Lambda = (\mathbb{R}_+^*)^2$ .

**Kernel ridge regression** Assume that we have a positive definite kernel  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  and we aim at estimating the true function  $f$  in the associated reproducing kernel Hilbert space  $(\mathcal{H}_k, \|\cdot\|_k)$ . The kernel ridge estimator is obtained by minimizing the criterion  $\|\mathbf{Y} - \mathbf{f}\|_n^2 + \lambda \|\mathbf{f}\|_k^2$  w.r.t.  $\mathbf{f} \in \mathcal{H}_k$  (see (Shawe-Taylor and Cristianini, 2000, page 118)). Denoting by  $K$  the  $n \times n$  kernel-matrix with element  $K_{i,j} = k(x_i, x_j)$ , the unique solution  $\hat{\mathbf{f}}$  is a linear estimate of the data,  $\hat{\mathbf{f}} = A_\lambda \mathbf{Y}$ , with  $A_\lambda = K(K + n\lambda I_{n \times n})^{-1}$ , where  $I_{n \times n}$  is the identity matrix of size  $n \times n$ .

**Multiple Kernel learning** As proposed in Arlot and Bach (2009), it is also possible to handle the case of several kernels  $k_1, \dots, k_M$ , with associated positive definite matrices  $K_1, \dots, K_M$ . For a parameter  $\lambda = (\lambda_1, \dots, \lambda_M) \in \Lambda = \mathbb{R}_+^M$  one can define the estimators  $\hat{\mathbf{f}}_\lambda = A_\lambda \mathbf{Y}$  with

$$A_\lambda = \left( \sum_{m=1}^M \lambda_m K_m \right) \left( \sum_{m=1}^M \lambda_m K_m + nI_{n \times n} \right)^{-1}. \quad (5)$$

It is worth mentioning that the formulation in Eq.(5) can be linked to the group Lasso Yuan and Lin (2006) and to the multiple kernel Lanckriet et al. (2003/04) — see Bach (2008); Arlot and Bach (2009) for more details.

### 1.3. Organization of the paper

In Section 2, we introduce EWA and state a PAC-Bayes type bound assessing the optimality of EWA in combining affine estimators. As a consequence, we provide in Section 3 sharp oracle inequalities in various set-ups: ranging from finite to continuous families of constituent estimators and including the sparsity scenario. In Section 4, we apply our main results to prove that combining Pinsker's type filters with EWA leads to an asymptotically sharp adaptive procedure over the Sobolev ellipsoids. Section 5 is devoted to a numerical comparison of EWA with other classical filters (soft thresholding, blockwise shrinking, etc.), and illustrates the potential benefits of the aggregation. Some concluding remarks are presented in Section 6, while technical proofs are postponed to the Appendix.

## 2. Aggregation of estimators: main result

In this section we describe the statistical framework for aggregating estimators and we also introduce the exponentially weighted aggregate. The task of aggregation consists in estimating  $f$  by a suitable combination of the elements of a family of *constituent estimators*  $\mathcal{F}_\Lambda = (\hat{\mathbf{f}}_\lambda)_{\lambda \in \Lambda} \in \mathbb{R}^n$ . The target objective of the aggregation is to build an aggregate  $\hat{\mathbf{f}}_{\text{aggr}}$ , not necessarily in the family  $\mathcal{F}_\Lambda$ , that mimics the performance of the best constituent estimator. It is called *oracle* because of its dependence on the unknown function  $f$ . We assume that  $\Lambda$  is a measurable subset of  $\mathbb{R}^M$ , for some  $M \in \mathbb{N}$ .

The theoretical tool commonly used for evaluating the quality of an aggregation procedure is the oracle inequality (OI), generally written in the following form:

$$\mathbb{E}\|\hat{\mathbf{f}}_{\text{aggr}} - \mathbf{f}\|_n^2 \leq C_n \inf_{\lambda \in \Lambda} (\mathbb{E}\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|_n^2) + R_n, \quad (6)$$

with *residual* term  $R_n$  tending to zero, and *leading constant*  $C_n$  being bounded. The OIs with leading constant one are of central theoretical interest since they allow to bound the excess risk and to assess the aggregation-rate-optimality. The residual term  $R_n$  depends on the complexity (size) of the family  $\mathcal{F}_\Lambda$ , as on the amount of noise, measured in term of variance in our context.

## 2.1. Exponentially Weighted Aggregate (EWA)

Let  $r_\lambda = \mathbb{E}(\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|_n^2)$  denote the risk of the estimator  $\hat{\mathbf{f}}_\lambda$ , for any  $\lambda \in \Lambda$ , and let  $\hat{r}_\lambda$  be an estimator of  $r_\lambda$ . The precise form of  $\hat{r}_\lambda$  strongly depends on the nature of the constituent estimators. For any probability distribution  $\pi$  over the set  $\Lambda$  and for any  $\beta > 0$ , we define the probability measure of exponential weights,  $\hat{\pi}$ , by the following formula:

$$\hat{\pi}(d\lambda) = \theta(\lambda)\pi(d\lambda) \quad \text{with} \quad \theta(\lambda) = \frac{\exp(-n\hat{r}_\lambda/\beta)}{\int_\Lambda \exp(-n\hat{r}_\omega/\beta)\pi(d\omega)}. \quad (7)$$

The corresponding exponentially weighted aggregate, henceforth denoted by  $\hat{\mathbf{f}}_{\text{EWA}}$ , is the expectation of the  $\hat{\mathbf{f}}_\lambda$  w.r.t. the probability measure  $\hat{\pi}$ :

$$\hat{\mathbf{f}}_{\text{EWA}} = \int_\Lambda \hat{\mathbf{f}}_\lambda \hat{\pi}(d\lambda). \quad (8)$$

It is convenient and customary to use the terminology of Bayesian statistics: the measure  $\pi$  is called *prior*, the measure  $\hat{\pi}$  is called *posterior* and the aggregate  $\hat{\mathbf{f}}_{\text{EWA}}$  is then the *posterior mean*. The parameter  $\beta$  will be referred to as the *temperature parameter*. In the framework of aggregating statistical procedures, the use of such an aggregate can be traced back to George (1986).

The interpretation of the weights  $\theta(\lambda)$  is simple: they up-weight estimators all the more that their performance, measured in terms of the risk estimate  $\hat{r}_\lambda$ , is good. The temperature parameter reflects the confidence we have in this criterion: if the temperature is small ( $\beta \approx 0$ ) the distribution concentrates on the estimators achieving the smallest value for  $\hat{r}_\lambda$ , assigning almost zero weights to the other estimators. On the other hand, if  $\beta \rightarrow +\infty$  then the probability distribution over  $\Lambda$  is simply the prior  $\pi$ , and the data do not modify our confidence in the estimators. It should also be noted that averaging w.r.t. the posterior  $\hat{\pi}$  is not the only way of constructing an estimator of  $\mathbf{f}$ , some alternative estimators based on  $\hat{\pi}$  have been studied, see for instance Zhang (2006); Audibert (2009).

## 2.2. Main result

To state our main result, we denote by  $\mathcal{P}_\Lambda$  the set of all probability measures on  $\Lambda$  and by  $\mathcal{K}(p, p')$  the Kullback-Leibler divergence between two probability measures  $p, p' \in \mathcal{P}_\Lambda$ :

$$\mathcal{K}(p, p') = \begin{cases} \int_\Lambda \log\left(\frac{dp}{dp'}(\lambda)\right)p(d\lambda) & \text{if } p \ll p', \\ +\infty & \text{otherwise.} \end{cases}$$

**Theorem 1 (PAC Bayesian Bound)** *If either one of the following conditions is satisfied:*

**C<sub>1</sub>:** *The matrices  $A_\lambda$  are orthogonal projections (i.e., symmetric and idempotent) and the vectors  $\mathbf{b}_\lambda$  satisfy  $A_\lambda \mathbf{b}_\lambda = 0$ , for all  $\lambda \in \Lambda$ .*

**C<sub>2</sub>:** *The matrices  $A_\lambda$  are all symmetric, positive semidefinite and satisfy  $A_\lambda A_{\lambda'} = A_{\lambda'} A_\lambda$ ,  $A_\lambda \Sigma = \Sigma A_\lambda$  for all  $\lambda, \lambda' \in \Lambda$ . All the vectors  $\mathbf{b}_\lambda$  are zero.*

*Then, the risk of the aggregate  $\hat{\mathbf{f}}_{\text{EWA}}$  defined by Equations (7), (8) and (4) satisfies the inequality*

$$r_{\text{EWA}} = \mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) \leq \inf_{p \in \mathcal{P}_\Lambda} \left( \int_{\Lambda} \mathbb{E}(\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|_n^2) p(d\lambda) + \frac{\beta}{n} \mathcal{K}(p, \pi) \right) \quad (9)$$

*provided that  $\beta \geq \alpha \|\Sigma\|$ , where  $\alpha = 4$  if C<sub>1</sub> holds true and  $\alpha = 8$  if C<sub>2</sub> holds true.*

All the proofs of our results are given in the appendix, at the end of the paper.

Note also that the result of Theorem 1 applies to the estimator  $\hat{\mathbf{f}}_{\text{EWA}}$  that uses the full knowledge of the covariance matrix  $\Sigma$ . Indeed, even if for the choice of  $\beta$  only an upper bound on the spectral norm of  $\Sigma$  is required, the entire matrix  $\Sigma$  enters in the definition of the unbiased risks  $\hat{r}_\lambda$  that is used for defining  $\hat{\mathbf{f}}_{\text{EWA}}$ . The exponentially weighted aggregate  $\hat{\mathbf{f}}_{\text{EWA}}$  is easily extended to handle the more realistic situation where an unbiased estimate  $\hat{\Sigma}$ , independent of  $\mathbf{Y}$ , of the covariance matrix  $\Sigma$  is available. Simply replace  $\Sigma$  by  $\hat{\Sigma}$  in the definition of the unbiased risk estimate (4). When the estimators  $\hat{\mathbf{f}}_\lambda$  satisfy  $\pi$ -a.e. condition C<sub>1</sub> or C<sub>2</sub>, choosing  $\beta = \alpha \|\hat{\Sigma}\|$ , it can be checked that a claim similar to Theorem 1 remains valid.

Another observation is that using the extension of Stein's lemma presented in (Dalalyan and Tsybakov, 2008, Lemma 1), a result similar to Theorem 1 can be established for some specific non Gaussian noise distributions, provided that the components of the noise vector are independent.

### 3. Sharp oracle inequalities

In this section, we discuss consequences of the main result for specific choices of prior measures. Some of them are closely related to the oracle inequalities presented in Dalalyan and Tsybakov (2007, 2008); Alquier and Lounici (2010); Rigollet and Tsybakov (2011) especially when dealing with the sparsity scenario in the high dimensional framework.

#### 3.1. Discrete oracle inequality

In order to demonstrate that Inequality (9) can be reformulated in terms of an OI as defined by (6), let us consider the simple case when the prior  $\pi$  is discrete. That is, we assume that  $\pi(\Lambda_0) = 1$  for a countable set  $\Lambda_0 \subset \Lambda$ . Without loss of generality, we assume that  $\Lambda_0 = \mathbb{N}$ . Then, the following result holds true.

**Proposition 2** *If either one of the conditions C<sub>1</sub> and C<sub>2</sub> (cf. Theorem 1) is fulfilled and  $\pi$  is supported by  $\mathbb{N}$ , then the aggregate  $\hat{\mathbf{f}}_{\text{EWA}}$  defined by Equations (7), (8) and (4) satisfies the inequality*

$$\mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) \leq \inf_{j \in \mathbb{N}: \pi_j > 0} \left( \mathbb{E} \|\hat{\mathbf{f}}_j - \mathbf{f}\|_n^2 + \frac{\beta \log(1/\pi_j)}{n} \right) \quad (10)$$

provided that  $\beta \geq \alpha \|\Sigma\|$ , where  $\alpha = 4$  if  $\mathbf{C}_1$  holds true and  $\alpha = 8$  if  $\mathbf{C}_2$  holds true.

**Proof** It suffices to apply Theorem 1 and to bound the RHS from above by the minimum over all Dirac measures  $p = \delta_j$  with  $j$  such that  $\pi_j > 0$ .  $\blacksquare$

### 3.2. Continuous oracle inequality

It may be useful in practice to combine a family of affine estimators indexed by an open subset of  $\mathbb{R}^M$ , for some integer  $M > 0$ , for instance when the aim is to build an estimator that is nearly as accurate as the best kernel estimator with fixed kernel and varying bandwidth. In order to state an oracle inequality in such a “continuous” setup, let us denote by  $d_2(\lambda, \Lambda)$  the largest real  $\tau > 0$  such that the ball centered at  $\lambda$  with radius  $\tau$  is included in  $\Lambda$ . In what follows,  $\text{Leb}(\cdot)$  stands for the Lebesgue measure.

**Proposition 3** *Let  $\Lambda \subset \mathbb{R}^M$  be an open and bounded set and let  $\pi$  be the uniform probability on  $\Lambda$ . Assume that the mapping  $\lambda \mapsto r_\lambda$  is Lipschitz continuous, i.e.,  $|r_{\lambda'} - r_\lambda| \leq L_r \|\lambda' - \lambda\|_2$ ,  $\forall \lambda, \lambda' \in \Lambda$ . Under the conditions  $\mathbf{C}_1$  or  $\mathbf{C}_2$  aggregate  $\hat{\mathbf{f}}_{\text{EWA}}$  satisfies the inequality*

$$\mathbb{E}\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2 \leq \inf_{\lambda \in \Lambda} \left\{ \mathbb{E}\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|_n^2 + \frac{\beta M}{n} \log\left(\frac{\sqrt{M}}{2 \min(n^{-1}, d_2(\lambda, \Lambda))}\right) \right\} + \frac{L_r + \beta \log(\text{Leb}(\Lambda))}{n}. \quad (11)$$

for every  $\beta \geq \alpha \|\Sigma\|$  where  $\alpha = 4$  if  $\mathbf{C}_1$  holds true and  $\alpha = 8$  if  $\mathbf{C}_2$  holds true.

### 3.3. Sparsity oracle inequality

The continuous oracle inequality stated in previous subsection is well adapted to the case where the dimension  $M$  of  $\Lambda$  is small compared to the sample size  $n$  (or, more precisely, the signal to noise ratio  $n / \max_i \sigma_i^2$ ). If this is not the case, the choice of the prior should be done more carefully. For instance, consider the case of a set  $\Lambda \subset \mathbb{R}^M$  with large  $M$  under the sparsity scenario: there is a sparse vector  $\lambda^* \in \Lambda$  such that the risk of  $\hat{\mathbf{f}}_{\lambda^*}$  is small. Then, it is natural to choose a prior  $\pi$  that promotes the sparsity of  $\lambda$ . This can be done in the same vein as in Dalalyan and Tsybakov (2007, 2008), by means of the heavy tailed prior:

$$\pi(d\lambda) \propto \prod_{j=1}^M \frac{1}{(1 + |\lambda_j/\tau|^2)^2} \mathbb{1}_\Lambda(\lambda) d(\lambda), \quad (12)$$

where  $\tau > 0$  is a tuning parameter.

**Proposition 4** *Let  $\Lambda = \mathbb{R}^M$  and let  $\pi$  be defined by (12). Assume that the mapping  $\lambda \mapsto r_\lambda$  is continuously differentiable and, for some  $M \times M$  matrix  $\mathcal{M}$ , satisfies:*

$$r_\lambda - r_{\lambda'} - \nabla r_{\lambda'}^\top (\lambda - \lambda') \leq (\lambda - \lambda')^\top \mathcal{M}(\lambda - \lambda'), \quad \forall \lambda, \lambda' \in \Lambda. \quad (13)$$

*If either one of the conditions  $\mathbf{C}_1$  and  $\mathbf{C}_2$  (cf. Theorem 1) is fulfilled, then the aggregate  $\hat{\mathbf{f}}_{\text{EWA}}$  defined by Equations (7), (8) and (4) satisfies the inequality*

$$\mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) \leq \inf_{\lambda \in \mathbb{R}^M} \left\{ \mathbb{E}\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|_n^2 + \frac{4\beta}{n} \sum_{j=1}^M \log\left(1 + \frac{|\lambda_j|}{\tau}\right) \right\} + \text{Tr}(\mathcal{M})\tau^2 \quad (14)$$

provided that  $\beta \geq \alpha \|\Sigma\|$ , where  $\alpha = 4$  if  $\mathbf{C}_1$  holds true and  $\alpha = 8$  if  $\mathbf{C}_2$  holds true.

Let us discuss here some consequences of this sparsity oracle inequality. First of all, let us remark that in most cases  $\text{Tr}(\mathcal{M})$  is on the order of  $M$  and the choice  $\tau = \sqrt{\beta/(nM)}$  ensures that the last term in the RHS of Eq. (14) decreases at the parametric rate  $1/n$ . This is the choice we recommend for practical applications.

Assume now that we are given a large number of linear estimators  $\hat{\mathbf{g}}_1 = G_1 \mathbf{Y}, \dots, \hat{\mathbf{g}}_M = G_M \mathbf{Y}$  satisfying, for instance, condition **C<sub>2</sub>**. We will focus on matrices  $G_j$  having a spectral norm bounded by one (it is well known that the failure of this condition makes the linear estimator inadmissible, cf. Cohen (1966)). Assume furthermore that our aim is to propose an estimator that mimics the behavior of the best possible convex combination of a pair of estimators chosen among  $\hat{\mathbf{g}}_1, \dots, \hat{\mathbf{g}}_M$ . This task can be accomplished in the framework of the present paper by setting  $\Lambda = \mathbb{R}^M$  and  $\hat{\mathbf{f}}_\lambda = \lambda_1 \hat{\mathbf{g}}_1 + \dots + \lambda_M \hat{\mathbf{g}}_M$ , where  $\lambda = (\lambda_1, \dots, \lambda_M)$ . If the collection  $\{\hat{\mathbf{g}}_i\}$  satisfies condition **C<sub>2</sub>**, then it is also the case for the collection of their linear combinations  $\{\hat{\mathbf{f}}_\lambda\}$ . Moreover, the mapping  $\lambda \mapsto r_\lambda$  is quadratic with the Hessian matrix  $\nabla^2 r_\lambda$  given by the entries  $2\langle G_j \mathbf{f} | G_{j'} \mathbf{f} \rangle_n + \frac{2}{n} \text{Tr}(G_{j'} \Sigma G_j)$ ,  $j, j' = 1, \dots, M$ . This implies that Inequality (13) holds with  $\mathcal{M}$  being the Hessian divided by 2. Therefore, setting  $\sigma = (\sigma_1, \dots, \sigma_n)$ , we get  $\text{Tr}(\mathcal{M}) \leq \|\sum_{j=1}^M G_j^2\| \|\mathbf{f}\|_n^2 + \|\sigma\|_n^2 \leq M(\|\mathbf{f}\|_n^2 + \|\sigma\|_n^2)$ , where the norm of a matrix is understood as its largest singular value. Applying Proposition 4 with  $\tau = \sqrt{\beta/(nM)}$ , we get for  $\beta \geq 8\|\Sigma\|$ ,

$$\mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) \leq \inf_{\alpha, j, j'} \mathbb{E}\|\alpha \hat{\mathbf{g}}_j + (1-\alpha) \hat{\mathbf{g}}_{j'} - \mathbf{f}\|_n^2 + \frac{8\beta}{n} \log\left(1 + \sqrt{\frac{Mn}{\beta}}\right) + \frac{\beta}{n}(\|\mathbf{f}\|_n^2 + \|\sigma\|_n^2), \quad (15)$$

where the inf is taken over all  $\alpha \in [0, 1]$  and  $j, j' \in \{1, \dots, M\}$ . This shows that, using EWA with a sufficiently large temperature, one can achieve the best possible risk over the convex combinations of a pair of linear estimators—selected from a large (but finite) family—at the price of a residual term that decreases at the parametric rate up to a log factor.

### 3.4. Oracle inequalities for varying-block-shrinkage estimators

Let us consider now the problem of aggregation of two-block shrinkage estimators. It means that the constituent estimators have the following form: for  $\lambda = (a, b, k) \in [0, 1]^2 \times \{1, \dots, n\} := \Lambda$ ,  $\hat{\mathbf{f}}_\lambda = A_\lambda \mathbf{Y}$  where  $A_\lambda = \text{diag}(a \mathbf{1}(i \leq k) + b \mathbf{1}(i > k))$ ,  $i = 1, \dots, n$ . Let us choose the prior  $\pi$  as the uniform probability distribution on the set  $\Lambda$ .

**Proposition 5** *Let  $\hat{\mathbf{f}}_{\text{EWA}}$  be the exponentially weighted aggregate having as constituent estimators two-block shrinkage estimators  $A_\lambda \mathbf{Y}$ . If  $\Sigma$  is a diagonal matrix, then for any  $\lambda \in \Lambda$  and for any  $\beta \geq 8\|\Sigma\|$ ,*

$$\mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) \leq \mathbb{E}(\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|_n^2) + \frac{\beta}{n} \left\{ 1 + \log\left(\frac{n^2 \|\mathbf{f}\|_n^2 + n \text{Tr}(\Sigma)}{12\beta}\right) \right\}. \quad (16)$$

The proof of this result can be found in Dalalyan and Salmon (2011).

In the case  $\Sigma = I_{n \times n}$ , this result is comparable to (Leung, 2004, page 20, Theorem 2.49), which states that in the model of homoscedastic regression ( $\Sigma = I_{n \times n}$ ), the EWA acting on two-block positive-part James-Stein shrinkage estimators satisfies, for any  $k = 3, \dots, n-3$ , and for  $\beta = 8$ , the oracle inequality

$$\mathbb{E}(\|\hat{\mathbf{f}}_{\text{Leung}} - \mathbf{f}\|_n^2) \leq \mathbb{E}(\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|_n^2) + \frac{9}{n} + \frac{8}{n} \min_{K>0} \left\{ K \vee \left( \log \frac{n-6}{K} - 1 \right) \right\}. \quad (17)$$

#### 4. Application to minimax adaptive estimation

In the celebrated paper Pinsker (1980) proved that in the model (1) the minimax risk over ellipsoids can be asymptotically attained by a linear estimator. Let us denote by  $\theta_k(\mathbf{f}) = \langle \mathbf{f} | \varphi_k \rangle_n$  the coefficients of the (orthogonal) discrete sine transform of  $\mathbf{f}$ , hereafter denoted by  $\mathcal{D}\mathbf{f}$ . Pinsker's result—restricted to Sobolev ellipsoids  $\mathcal{F}(\alpha, R) = \{\mathbf{f} \in \mathbb{R}^n : \sum_{k=1}^n k^{2\alpha} \theta_k(\mathbf{f})^2 \leq R\}$  and to the homoscedastic noise ( $\Sigma = \sigma^2 I_{n \times n}$ )—states that, as  $n \rightarrow \infty$ , the equivalences

$$\inf_{\hat{\mathbf{f}}} \sup_{\mathbf{f} \in \mathcal{F}(\alpha, R)} \mathbb{E}(\|\hat{\mathbf{f}} - \mathbf{f}\|_n^2) \sim \inf_A \sup_{\mathbf{f} \in \mathcal{F}(\alpha, R)} \mathbb{E}(\|A\mathbf{Y} - \mathbf{f}\|_n^2) \quad (18)$$

$$\sim \inf_{w>0} \sup_{\mathbf{f} \in \mathcal{F}(\alpha, R)} \mathbb{E}(\|A_{\alpha, w}\mathbf{Y} - \mathbf{f}\|_n^2) \quad (19)$$

hold (Tsybakov, 2009, Theorem 3.2), where the first inf is taken over all possible estimators  $\hat{\mathbf{f}}$  and  $A_{\alpha, w} = \mathcal{D}^\top \text{diag}((1 - k^\alpha / w)_+ ; k = 1, \dots, n) \mathcal{D}$  is the Pinsker filter in the discrete sine basis. In simple words, this implies that the (asymptotically) minimax estimator can be chosen from the quite narrow class of linear estimators with Pinsker's filter. However, it should be emphasized that the minimax linear estimator depends on the parameters  $\alpha$  and  $R$ , that are generally unknown. An (adaptive) estimator, that does not depend on  $(\alpha, R)$  and is asymptotically minimax over a large scale of Sobolev ellipsoids has been proposed by Efromovich and Pinsker (1984). The next result, that is a direct consequence of Theorem 1, shows that EWA with linear constituent estimators is also asymptotically sharp adaptive over Sobolev ellipsoids.

**Proposition 6** *Let  $\lambda = (\alpha, w) \in \Lambda = \mathbb{R}_+^2$  and consider the prior*

$$\pi(d\lambda) = \frac{2n_\sigma^{-\alpha/(2\alpha+1)}}{(1 + n_\sigma^{-\alpha/(2\alpha+1)} w)^3} e^{-\alpha} d\alpha dw, \quad (20)$$

where  $n_\sigma = n/\sigma^2$ . Then, in model (1) with homoscedastic errors, the aggregate  $\hat{\mathbf{f}}_{\text{EWA}}$  based on the temperature  $\beta = 8\sigma^2$  and the constituent estimators  $\hat{\mathbf{f}}_{\alpha, w} = A_{\alpha, w}\mathbf{Y}$  (with  $A_{\alpha, w}$  being the Pinsker filter) is adaptive in the exact minimax sense<sup>1</sup> on the family of classes  $\{\mathcal{F}(\alpha, R) : \alpha > 0, R > 0\}$ .

It is worth noting that the exact minimax adaptivity property of our estimator  $\hat{\mathbf{f}}_{\text{EWA}}$  is achieved without any tuning parameter. All previously proposed methods that are provably adaptive in exact minimax sense depend on some parameters such as the lengths of blocks for blockwise Stein and Efromovich-Pinsker estimators or the step of discretization and the maximal value of bandwidth Cavalier et al. (2002). Another nice property of the estimator  $\hat{\mathbf{f}}_{\text{EWA}}$  is that it does not require any pilot estimator based on the data splitting device Efromovich (1996); Yang (2004).

#### 5. Experiments

In this section we present some numerical experiments on synthetic data, by focusing only on the case of homoscedastic Gaussian noise ( $\Sigma = \sigma^2 I_{n \times n}$ ) with known variance. Following the philosophy of reproducible research, a toolbox is made available freely for download at: [www.math.jussieu.fr/~salmon/code/index\\_codes.php](http://www.math.jussieu.fr/~salmon/code/index_codes.php)

1. see (Tsybakov, 2009, Definition 3.8)

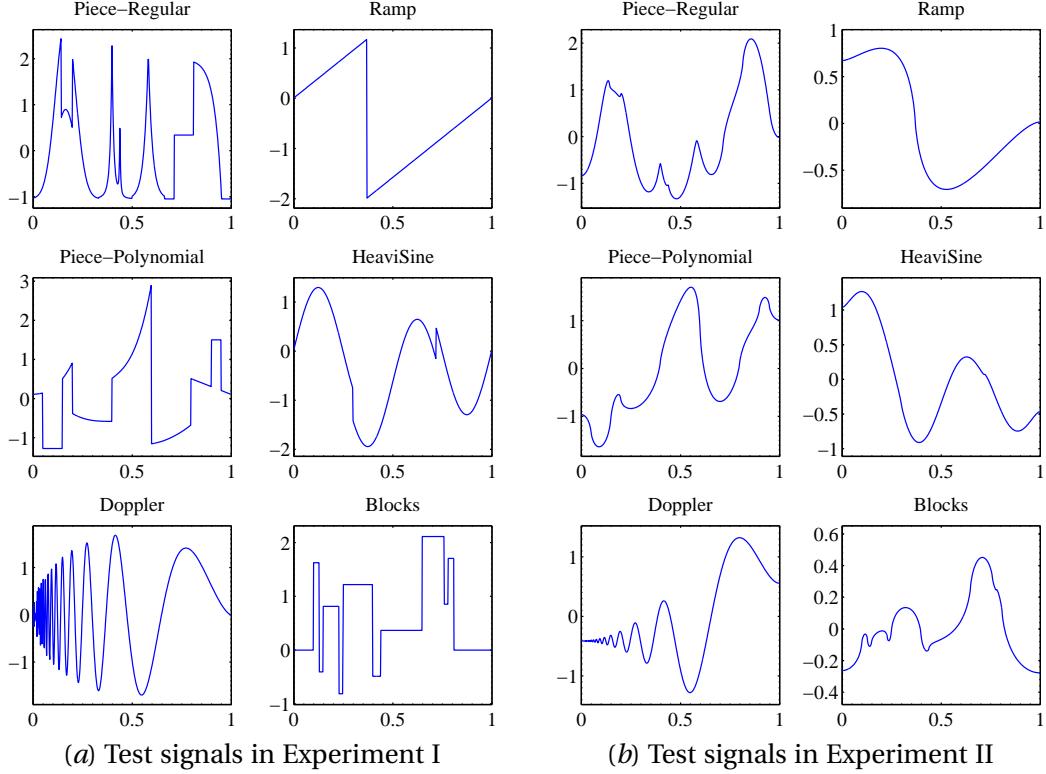


Figure 1: Test signals used in our experiment: Piece-Regular, Ramp, Piece-Polynomial, HeaviSine, Doppler and Blocks. (a) non-smooth (Experiment I) and (b) smooth (Experiment II).

We evaluate different estimation routines on several 1D signals, introduced by Donoho and Johnstone (1994, 1995) and considered as benchmark in literature on signal processing. The six signals we retained for our experiments because of their diversity are depicted in Figure 1. Since all these signals are non-smooth, we have also carried out experiments on their smoothed versions obtained by taking an antiderivative, see Figure 1. In what follows, the experiment on non-smooth signals will be referred to as Experiment I, whereas the experiment on their smoothed counterparts will be referred to as Experiment II. In both cases, prior to applying estimation routines, we normalize the (true) sampled signal to have an empirical norm equal to one and use the Discrete Sine Transform (DST) denoted by  $\boldsymbol{\theta}(\mathbf{Y}) = (\theta_1(\mathbf{Y}), \dots, \theta_n(\mathbf{Y}))^\top$ .

The four estimation routines—including EWA—used in our experiments are detailed below:

**Soft Thresholding (ST), Donoho and Johnstone (1994):** For a given threshold parameter  $t$ , the soft thresholding estimator of the vector of DST coefficients  $\theta_k(\mathbf{f})$  is defined by

$$\hat{\theta}_k = \text{sgn}(\theta_k(\mathbf{Y}))(|\theta_k(\mathbf{Y})| - \sigma t)_+. \quad (21)$$

In our experiments, we use the threshold minimizing the estimated unbiased risk defined via Stein's lemma. This procedure is referred to as SURE-shrink in Donoho and Johnstone (1995).

**Blockwise James-Stein (BJS) shrinkage, Cai (1999):** The set of indices  $\{1, \dots, n\}$  is partitioned into  $N = \lceil n/\log(n) \rceil$  non-overlapping blocks  $B_1, B_2, \dots, B_N$  of equal size  $L$ . (If  $n$  is not a multiple of  $N$ , the last block may be of smaller size than all the others.) The corresponding blocks of true coefficients  $\theta_{B_k}(\mathbf{f}) = (\theta_j(\mathbf{f}))_{j \in B_k}$  are estimated by shrinking the blocks of noisy coefficients  $\theta_{B_k}(\mathbf{Y})$ :

$$\widehat{\theta}_{B_k} = \left( 1 - \frac{\lambda L \sigma^2}{S_k^2(\mathbf{Y})} \sigma \right)_+ \theta_{B_k}(\mathbf{Y}), \quad k = 1, \dots, N \quad (22)$$

where  $S_k^2(\mathbf{Y}) = \|\theta_{B_k}(\mathbf{Y})\|_2^2$  and  $\lambda = 4.50524$  as in Cai (1999).

**Unbiased risk estimate (URE) minimization, Golubev (1992); Cavalier et al. (2002):** it consists in using a Pinsker filter, as defined in Section 4, with a data-driven choice of parameters  $\alpha$  and  $w$ . This choice is done by minimizing an unbiased estimate of the risk over a suitably chosen grid for the values of  $\alpha$  and  $w$ . Here, we use geometric grids ranging from 0.1 to 100 for  $\alpha$  and from 1 to  $n$  for  $w$ . The bi-dimensional grid used in all the experiments has  $100 \times 100$  elements. We refer to Cavalier et al. (2002) for the closed-form formula of the unbiased risk estimator.

**EWA on Pinsker's filters:** We consider the same finite family of linear smoothers—defined by Pinsker's filters—as in the URE routine described above. According to Proposition 2, this leads to an estimator which is nearly as accurate as the best Pinsker's estimator in the given finite family.

To report the result of our experiments, we have also computed the best linear smoother based on a Pinsker filter chosen among the candidates that we used for defining URE and EWA routines. By best smoother we mean the one minimizing the squared error, which can be computed since we know the ground truth. This pseudo-estimator will be referred to as oracle. The results summarized in Table 1 for Experiment I and Table 2 for Experiment II correspond to the average over 1000 trials of the mean squared error (MSE) from which we subtract the MSE of the oracle and multiply the resulting difference by the sample size. We report the results for  $\sigma = 0.33$  and for  $n \in \{2^8, 2^9, 2^{10}, 2^{11}\}$ .

Simulations show that EWA and URE have very comparable performances and are significantly more accurate than Soft Thresholding and Block James-Stein (see Table 1) for every size  $n$  of signals considered. The improvement is particularly important when the signal has large peaks (cf. Figure 2) or discontinuities (cf. Figure 3). In most cases, the EWA method also outperforms URE, but this difference is much less pronounced. One can also observe that in the case of smooth signals, the difference of the MSEs between EWA and the oracle, multiplied by  $n$ , remains nearly constant when  $n$  varies. This is in perfect agreement with our theoretical results in which the residual term decreases to zero inversely proportionally to the sample size.

Of course, soft thresholding and blockwise James-Stein procedures have been designed for being applied to the wavelet transform of a Besov smooth function, rather than to the Fourier transform of a Sobolev-smooth function. However, the point here is not to demonstrate the superiority of EWA as compared to ST and BJS procedures. The point is to stress the importance of having sharp adaptivity up to optimal constant and not simply adaptivity in the sense of rate of convergence. Indeed, the procedures ST and BJS are provably rate-adaptive when applied to

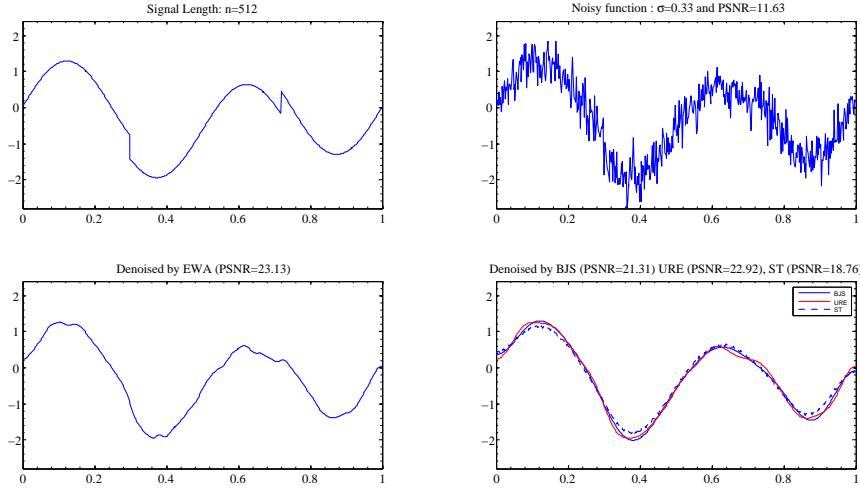


Figure 2: Heavisine. The first row is the true signal (left) and a noisy version corrupted by Gaussian noise with standard deviation  $\sigma = 0.33$  (right). The second row gives denoised version obtained by EWA (left), BJS, ST and URE (right). The PSNR is computed by the formula  $\text{PSNR} = 10 \log_{10} (\max(f)^2 / \text{MSE})$ .

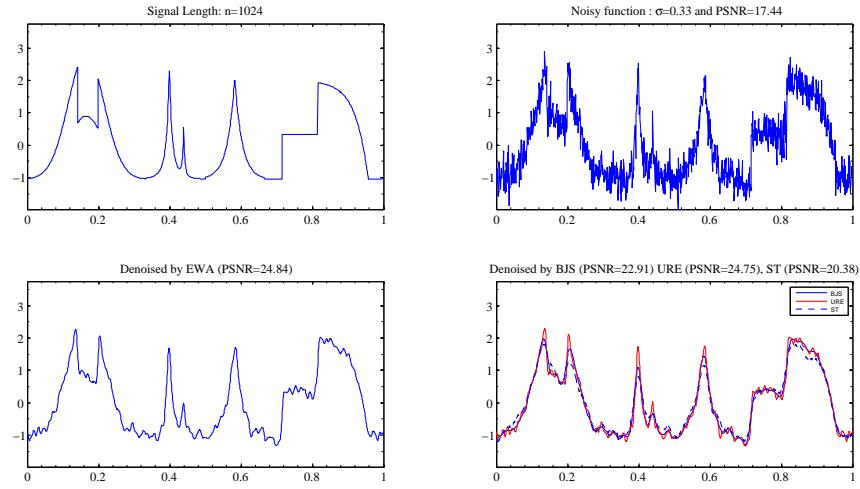


Figure 3: Piece-Regular. The first row is the true signal (left) and a noisy version corrupted by Gaussian noise with standard deviation  $\sigma = 0.33$  (right). The second row gives denoised version obtained by EWA (left) and by BJS, ST and URE (right). The PSNR is computed by the formula  $\text{PSNR} = 10 \log_{10} (\max(f)^2 / \text{MSE})$ .

Fourier transform of a Sobolev-smooth function, but they are not sharp adaptive—they do not attain the optimal constant—whereas EWA and URE do attain.

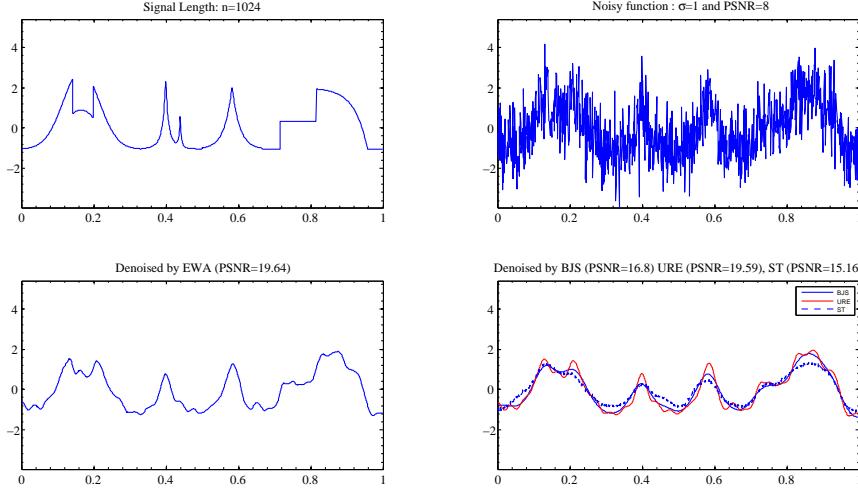


Figure 4: Piece-Regular. The first row is the true signal (left) and a noisy version corrupted by Gaussian noise with standard deviation  $\sigma = 1$  (right). The second row gives denoised version obtained by EWA (left) and by BJS, ST and URE (right). The PSNR is computed by the formula  $\text{PSNR} = 10 \log_{10} (\max(f)^2 / \text{MSE})$ .

## 6. Summary and future work

In this paper, we have addressed the problem of aggregating a set of affine estimators in the context of regression with fixed design and heteroscedastic noise. Under some assumptions on the constituent estimators, we have proven that the EWA with a suitably chosen temperature parameter satisfies PAC-Bayesian type inequality, from which different types of oracle inequalities have been deduced. All these inequalities are with leading constant one and with rate-optimal residual term. As a by-product of our results, we have shown that EWA applied to the family of Pinsker's estimators produces an estimator, which is adaptive in the exact minimax sense. Next in our agenda is carrying out an experimental evaluation of the proposed aggregate using the approximation schemes described by Dalalyan and Tsybakov (2009), Rigollet and Tsybakov (2011) and Alquier and Lounici (2010). It will also be interesting to extend the results of this work to the case of the unknown noise variance in the same vein as in Giraud (2008).

## Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under the grant PARCIMONIE.

## References

- P. Alquier and K. Lounici. Pac-bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Statist.*, 5:127–145, 2010.

<i>n</i>	EWA	URE	BJS	ST	EWA	URE	BJS	ST
<b>Blocks</b>								
256	0.051 (0.42)	0.245 (0.39)	9.617 (1.78)	4.846 (1.29)	0.062 (0.35)	0.212 (0.31)	13.233 (2.11)	6.036 (1.23)
512	-0.052 (0.35)	0.302 (0.50)	13.807 (2.16)	9.256 (1.70)	-0.100 (0.30)	0.205 (0.39)	17.080 (2.29)	12.620 (1.75)
1024	-0.050 (0.36)	0.299 (0.46)	19.984 (2.68)	17.569 (2.17)	-0.107 (0.35)	0.270 (0.41)	21.862 (2.92)	23.006 (2.35)
2048	-0.007 (0.42)	0.362 (0.57)	28.948 (3.31)	30.447 (2.96)	-0.150 (0.34)	0.234 (0.42)	28.733 (3.19)	38.671 (3.02)
<b>HeaviSine</b>								
256	-0.060 (0.19)	0.247 (0.42)	1.155 (0.57)	3.966 (1.12)	-0.069 (0.32)	0.248 (0.40)	8.883 (1.76)	4.879 (1.20)
512	-0.079 (0.19)	0.215 (0.39)	2.064 (0.86)	5.889 (1.36)	-0.105 (0.30)	0.237 (0.37)	12.147 (2.28)	9.793 (1.64)
1024	-0.059 (0.23)	0.240 (0.36)	3.120 (1.20)	8.685 (1.64)	-0.092 (0.34)	0.291 (0.46)	15.207 (2.18)	16.798 (2.13)
2048	-0.051 (0.25)	0.278 (0.48)	4.858 (1.42)	12.667 (2.03)	-0.059 (0.34)	0.283 (0.54)	21.543 (2.47)	27.387 (2.77)
<b>Ramp</b>								
256	0.038 (0.37)	0.294 (0.47)	6.933 (1.54)	5.644 (1.20)	0.017 (0.37)	0.203 (0.37)	12.201 (1.81)	3.988 (1.19)
512	0.010 (0.36)	0.293 (0.51)	9.712 (1.76)	9.977 (1.67)	-0.078 (0.35)	0.312 (0.49)	17.765 (2.72)	9.031 (1.62)
1024	-0.002 (0.30)	0.300 (0.45)	13.656 (2.25)	16.790 (2.06)	-0.026 (0.38)	0.321 (0.48)	23.321 (2.96)	17.565 (2.28)
2048	0.007 (0.34)	0.312 (0.50)	19.113 (2.68)	27.315 (2.61)	-0.007 (0.41)	0.314 (0.49)	31.550 (3.05)	29.461 (2.95)
<b>Piece-Polynomial</b>								

Table 1: Comparison of several adaptive methods on the six (non-smooth) signals of interest. For each signal length  $n$  and each method, we give the average value of  $n \times (\text{MSE} - \text{MSE}_{\text{Oracle}})$  and the corresponding standard deviation below, for 1000 replications of the experiment. Negative values indicate that in some cases the EWA procedure has a smaller risk than that of the best linear estimator used for the aggregation, which is possible since the EWA itself is not a linear estimator.

- Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Comput.*, 9:1545–1588, October 1997.
- S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. In *NIPS*, pages 46–54, 2009.
- J-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4):1591–1646, 2009.
- F. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.
- Y. Baraud, Ch. Giraud, and S. Huet. Estimator selection in the gaussian setting. *submitted*, 2010.

<i>n</i>	EWA	URE	BJS	ST	EWA	URE	BJS	ST
<b>Blocks</b>								
256	0.387 (0.43)	0.216 (0.40)	0.216 (0.24)	2.278 (0.98)	0.214 (0.23)	0.237 (0.40)	1.608 (0.73)	2.777 (1.04)
512	0.170 (0.20)	0.209 (0.41)	0.650 (0.25)	3.193 (1.07)	0.165 (0.20)	0.250 (0.44)	1.200 (0.48)	3.682 (1.24)
1024	0.162 (0.18)	0.226 (0.41)	1.282 (0.44)	4.507 (1.28)	0.147 (0.19)	0.229 (0.45)	1.842 (0.86)	5.043 (1.43)
2048	0.120 (0.17)	0.220 (0.37)	1.574 (0.55)	6.107 (1.55)	0.138 (0.20)	0.229 (0.40)	1.864 (1.07)	6.584 (1.58)
<b>HeaviSine</b>								
256	0.217 (0.16)	0.207 (0.42)	1.399 (0.54)	2.496 (0.96)	0.269 (0.27)	0.279 (0.49)	2.120 (1.09)	2.053 (0.95)
512	0.206 (0.18)	0.221 (0.43)	0.024 (0.26)	3.045 (1.10)	0.216 (0.20)	0.248 (0.45)	2.045 (1.17)	2.883 (1.13)
1024	0.179 (0.18)	0.200 (0.50)	0.113 (0.27)	3.905 (1.27)	0.183 (0.20)	0.228 (0.41)	1.251 (0.70)	3.780 (1.37)
2048	0.162 (0.15)	0.189 (0.37)	0.421 (0.27)	5.019 (1.53)	0.145 (0.19)	0.223 (0.42)	1.650 (1.12)	4.992 (1.42)
<b>Ramp</b>								
256	0.162 (0.16)	0.200 (0.38)	0.339 (0.24)	2.770 (1.00)	0.215 (0.25)	0.257 (0.48)	1.486 (0.68)	2.649 (1.01)
512	0.150 (0.18)	0.215 (0.38)	0.425 (0.23)	3.658 (1.20)	0.170 (0.20)	0.243 (0.46)	1.865 (0.84)	3.683 (1.20)
1024	0.146 (0.18)	0.211 (0.39)	0.935 (0.33)	4.815 (1.35)	0.179 (0.20)	0.236 (0.47)	1.547 (1.02)	5.017 (1.38)
2048	0.141 (0.20)	0.221 (0.43)	1.316 (0.42)	6.432 (1.54)	0.165 (0.20)	0.210 (0.39)	2.246 (1.15)	6.628 (1.70)
<b>Piece-Polynomial</b>								

Table 2: Comparison of several adaptive methods on the six smoothed signals of interest. For each signal length  $n$  and each method, we give the average value of  $n(\text{MSE} - \text{MSE}_{\text{Oracle}})$  and the corresponding standard deviation below, for 1000 replications of the experiment.

- A. R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.
- L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996.
- A. Buades, B. Coll, and J-M. Morel. A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.*, 4(2):490–530, 2005.
- F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.
- T. T. Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.*, 27(3):898–924, 1999. ISSN 0090-5364.
- O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004.

- L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874, 2002.
- A. Cohen. All admissible linear estimates of the mean vector. *The Annals of Mathematical Statistics*, 37(2):458–463, 1966. ISSN 00034851.
- A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. Technical Report arXiv:1104.3969v2 [math.ST], April 2011.
- A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp oracle inequalities and sparsity. In *COLT*, pages 97–111, 2007.
- A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2):39–61, 2008.
- A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. In *COLT*, 2009.
- D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.
- D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224, 1995.
- S. Y. Efromovich. On nonparametric regression for IID observations in a general setting. *Ann. Statist.*, 24(3):1125–1144, 1996.
- S. Y. Efromovich and M. S. Pinsker. A self-training algorithm for nonparametric filtering. *Avtomat. i Telemekh.*, 1(11):58–65, 1984.
- S. Y. Efromovich and M. S. Pinsker. Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica*, 6(4):925–942, 1996.
- Y. Freund. Boosting a weak learning algorithm by majority. In *Proceedings of the third annual workshop on Computational learning theory*, COLT, pages 202–216, 1990.
- E. I. George. Minimax multiple shrinkage estimation. *Ann. Statist.*, 14(1):188–205, 1986.
- Ch. Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14(4):1089–1107, 2008.
- G. K. Golubev. Nonparametric estimation of smooth densities of a distribution in  $L_2$ . *Problemy Peredachi Informatsii*, 28(1):52–62, 1992.
- Y. Golubev. On universal oracle inequalities related to high-dimensional linear models. *Ann. Statist.*, 38(5):2751–2780, 2010.
- A. B. Juditsky and A. S. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712, 2000.
- A. B. Juditsky and A. S. Nemirovski. Nonparametric denoising of signals with unknown local structure. I. Oracle inequalities. *Appl. Comput. Harmon. Anal.*, 27(2):157–179, 2009.

- G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72 (electronic), 2003/04.
- G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007.
- G. Leung. *Information Theory and Mixing Least Squares Regression*. PhD thesis, Yale University, 2004.
- G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory*, 52(8):3396–3410, 2006.
- K. Lounici. Generalized mirror averaging and  $D$ -convex aggregation. *Math. Methods Statist.*, 16(3):246–259, 2007.
- A. S. Nemirovski. *Topics in non-parametric statistics*, volume 1738 of *Lecture Notes in Math.* Springer, Berlin, 2000.
- M. S. Pinsker. Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Peredachi Inf.*, 16(2):52–68, 1980.
- Ph. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007.
- Ph. Rigollet and A. B. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–471, 2011.
- J. Salmon and E. Le Pennec. NL-Means and aggregation procedures. In *ICIP*, pages 2977–2980, 2009.
- J. Shawe-Taylor and N. Cristianini. *An introduction to support vector machines : and other kernel-based learning methods*. Cambridge University Press, 2000.
- C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.
- A. B. Tsybakov. Optimal rates of aggregation. In *COLT*, pages 303–313, 2003.
- A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer, New York, 2009.
- Y. Yang. Combining different procedures for adaptive regression. *J. Multivariate Anal.*, 74(1):135–161, 2000.
- Y. Yang. Regression with multiple candidate models: selecting or mixing? *Statist. Sinica*, 13(3):783–809, 2003.
- Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004.
- M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.

T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Inform. Theory*, 52(4):1307–1321, 2006. ISSN 1557-9654.

## Appendix

### Appendix A. Stein's Lemma with heteroscedastic noise

To define the EWA estimator, we first need to determine an unbiased risk estimate for any of the constituent estimators. We adapt a systematic method based on Stein's Lemma to the heteroscedastic framework. We recall this lemma given in Stein (1981), for our setting:

**Stein's Lemma 1** *With the model (1), if the estimator  $\hat{\mathbf{f}}$  is almost everywhere differentiable in  $\mathbf{Y}$  and if each  $\partial_{y_i} \hat{f}_i$  has finite first moment, then*

$$\hat{r} = \|\mathbf{Y} - \hat{\mathbf{f}}\|_n^2 + \frac{2}{n} \sum_{i=1}^n \sigma_i^2 \partial_{y_i} \hat{f}_i - \frac{1}{n} \sum_{i=1}^n \sigma_i^2, \quad (23)$$

is an unbiased estimate of  $r$ , ie.  $\mathbb{E}\hat{r} = r$ .

**Proof** For any  $i = 1, \dots, n$ , one has

$$\mathbb{E}(Y_i - \hat{f}_i)^2 = \mathbb{E}(Y_i - f_i)^2 + \mathbb{E}(f_i - \hat{f}_i)^2 + 2\mathbb{E}[(Y_i - f_i)(f_i - \hat{f}_i)],$$

The following identity is the classical Stein Lemma (cf. Tsybakov (2009) p.157), based on integration by parts:

$$\mathbb{E}[(Y_i - f_i)\hat{f}_i] = \sigma_i^2 \mathbb{E}[\partial_{y_i} f_i]. \quad (24)$$

where the differentiation is according to  $Y_i$ . Using the last two displays, one has:

$$\mathbb{E}\|\mathbf{Y} - \hat{\mathbf{f}}\|_n^2 = \mathbb{E}\|\mathbf{Y} - \mathbf{f}\|_n^2 + \mathbb{E}\|\mathbf{f} - \hat{\mathbf{f}}\|_n^2 - \frac{2}{n} \mathbb{E} \sum_{i=1}^n \sigma_i^2 \partial_{y_i} f_i, \quad (25)$$

leading to the announced unbiased risk estimate. ■

### Appendix B. Main Result

Now, we can apply Stein's Lemma for any estimator  $\hat{f}_\lambda$ , so that we can build  $\hat{r}_\lambda$  for any  $\lambda \in \Lambda$ . In this paper, we only focus on *affine estimators*  $\hat{f}_\lambda$ , i.e., estimators that can be written as affine transforms of the data  $\mathbf{Y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ . Affine estimators can be defined by

$$\hat{f}_\lambda = A_\lambda \mathbf{Y} + \mathbf{b}_\lambda, \quad (26)$$

where the  $n \times n$  real matrix  $A_\lambda$  and the vector  $\mathbf{b}_\lambda \in \mathbb{R}^n$  are deterministic. This means that the entries of  $A_\lambda$  and  $\mathbf{b}_\lambda$  may depend on the design points  $x_1, \dots, x_n$  but not on the data vector  $\mathbf{Y}$ .

It is easy to check that the the divergence term  $\sum_{i=1}^n \sigma_i^2 \partial_{y_i} \hat{\mathbf{f}}_i$  in Stein's Lemma is simply  $\text{Tr}(\Sigma A_\lambda)$  for affine estimators. Then  $\hat{r}_\lambda$ , defined by

$$\hat{r}_\lambda = \|\mathbf{Y} - \hat{\mathbf{f}}_\lambda\|_n^2 + \frac{2}{n} \text{Tr}(\Sigma A_\lambda) - \frac{1}{n} \sum_{i=1}^n \sigma_i^2 \quad (27)$$

is an unbiased estimator of  $r_\lambda$ .

In order to state our main result, we denote by  $\mathcal{P}_\Lambda$  the set of all probability measures on  $\Lambda$  and by  $\mathcal{K}(p, p')$  the Kullback-Leibler divergence between two probability measures  $p, p' \in \mathcal{P}_\Lambda$ .

$$\mathcal{K}(p, p') = \begin{cases} \int_\Lambda \log\left(\frac{dp}{dp'}(\lambda)\right) p(d\lambda) & \text{if } p \ll p', \\ +\infty & \text{otherwise.} \end{cases}$$

**Theorem 1** *If either one of the following conditions is satisfied:*

**C<sub>1</sub>:** *The matrices  $A_\lambda$  are orthogonal projections (i.e., symmetric and idempotent) and the vectors  $\mathbf{b}_\lambda$  satisfy  $A_\lambda \mathbf{b}_\lambda = 0$ , for all  $\lambda \in \Lambda$ .*

**C<sub>2</sub>:** *The matrices  $A_\lambda$  are all symmetric, positive semidefinite and satisfy  $A_\lambda A_{\lambda'} = A_{\lambda'} A_\lambda$ ,  $A_\lambda \Sigma = \Sigma A_\lambda$  for all  $\lambda, \lambda' \in \Lambda$ . All the vectors  $\mathbf{b}_\lambda$  are zero.*

Then, the aggregate  $\hat{\mathbf{f}}_{\text{EWA}}$  defined by Equations (7), (8) and (4) satisfies the inequality

$$\mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) \leq \inf_{p \in \mathcal{P}_\Lambda} \left( \int_\Lambda \mathbb{E}\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|_n^2 p(d\lambda) + \frac{\beta}{n} \mathcal{K}(p, \pi) \right)$$

provided that  $\beta \geq \alpha \|\Sigma\|$ , where  $\alpha = 4$  if C<sub>1</sub> holds true and  $\alpha = 8$  if C<sub>2</sub> holds true.

**Proof** [when C<sub>2</sub> is satisfied] According to Stein's lemma, the quantity

$$\hat{r}_{\text{EWA}} = \|\mathbf{Y} - \hat{\mathbf{f}}_{\text{EWA}}\|_n^2 + \frac{2}{n} \sum_{i=1}^n \sigma_i^2 \partial_{y_i} \hat{f}_{\text{EWA}, i} - \frac{1}{n} \sum_{i=1}^n \sigma_i^2 \quad (28)$$

is an unbiased estimate of the risk  $r_{\text{EWA}} = \mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2)$ . Using simple algebra, one checks that

$$\|\mathbf{Y} - \hat{\mathbf{f}}_{\text{EWA}}\|_n^2 = \int_\Lambda \left( \|\mathbf{Y} - \hat{\mathbf{f}}_\lambda\|_n^2 - \|\hat{\mathbf{f}}_\lambda - \hat{\mathbf{f}}_{\text{EWA}}\|_n^2 \right) \theta(\lambda) \pi(d\lambda). \quad (29)$$

By interchanging the integral and differential operators, we get the following expression for the derivatives of  $\hat{f}_{\text{EWA}, i}$ :

$$\partial_{y_i} \hat{f}_{\text{EWA}, i} = \int_\Lambda (\partial_{y_i} \hat{f}_{\lambda, i}) \theta(\lambda) \pi(d\lambda) + \int_\Lambda \hat{f}_{\lambda, i} (\partial_{y_i} \theta(\lambda)) \pi(d\lambda). \quad (30)$$

Let us defined  $A_{\text{EWA}} \triangleq \int_\Lambda A_\lambda \theta(\lambda) \pi(d\lambda)$ . With this notation, the last equality, combined with Equations (4), (28), (29) and the fact that  $\sum_{i=1}^n \sigma_i^2 \partial_{y_i} \hat{f}_{\lambda, i} = \text{Tr}(\Sigma A_\lambda)$ , implies that

$$\hat{r}_{\text{EWA}} = \int_\Lambda (\hat{r}_\lambda - \|\hat{\mathbf{f}}_\lambda - \hat{\mathbf{f}}_{\text{EWA}}\|_n^2) \theta(\lambda) \pi(d\lambda) + \frac{2}{n} \sum_{i=1}^n \sigma_i^2 \int_\Lambda \hat{f}_{\lambda, i} (\partial_{y_i} \theta(\lambda)) \pi(d\lambda).$$

Taking into account that  $\int_{\Lambda} \hat{f}_{\text{EWA},i} (\partial_{y_i} \theta(\lambda)) \pi(d\lambda) = \hat{f}_{\text{EWA},i} \partial_{y_i} (\int_{\Lambda} \theta(\lambda) \pi(d\lambda)) = 0$ , we come up with the following expression for the unbiased risk estimate:

$$\hat{r}_{\text{EWA}} = \int_{\Lambda} \left( \hat{r}_{\lambda} - \|\hat{f}_{\lambda} - \hat{f}_{\text{EWA}}\|_n^2 + 2 \langle \nabla_Y \log \theta(\lambda) | \Sigma(\hat{f}_{\lambda} - \hat{f}_{\text{EWA}}) \rangle_n \right) \theta(\lambda) \pi(d\lambda) \quad (31)$$

$$= \int_{\Lambda} \left( \hat{r}_{\lambda} - \|\hat{f}_{\lambda} - \hat{f}_{\text{EWA}}\|_n^2 - 2n\beta^{-1} \langle \nabla_Y \hat{r}_{\lambda} | \Sigma(\hat{f}_{\lambda} - \hat{f}_{\text{EWA}}) \rangle_n \right) \theta(\lambda) \pi(d\lambda). \quad (32)$$

Note that, so far, the precise form of the constituent estimators has not been exploited. This form is important for computing  $\nabla_Y \hat{r}_{\lambda}$ . In view of Equations (26) and (4), as well as the assumptions  $A_{\lambda}^T = A_{\lambda}$  and  $\mathbf{b}_{\lambda} \equiv 0$  (holding thanks to  $\mathbf{C}_2$ ), we get

$$\nabla_Y \hat{r}_{\lambda} = \frac{2}{n} (I_{n \times n} - A_{\lambda})^T (I_{n \times n} - A_{\lambda}) Y - \frac{2}{n} (I_{n \times n} - A_{\lambda})^T \mathbf{b}_{\lambda} = \frac{2}{n} (I_{n \times n} - A_{\lambda})^2 Y. \quad (33)$$

In what follows, we use the shorthand  $I = I_{n \times n}$ . Using this notation and Eq. (33), we get

$$\hat{r}_{\text{EWA}} = \int_{\Lambda} \left( \hat{r}_{\lambda} - \|\hat{f}_{\lambda} - \hat{f}_{\text{EWA}}\|_n^2 - \frac{4}{\beta} \langle (I - A_{\lambda})^2 Y | \Sigma(A_{\lambda} - A_{\text{EWA}}) Y \rangle_n \right) \theta(\lambda) \pi(d\lambda). \quad (34)$$

Recall now that for any pair of commuting matrices  $P$  and  $Q$  the identity  $(I - P)^2 = (I - Q)^2 + 2(I - \frac{P+Q}{2})(Q - P)$  holds true. Applying this formula to  $P = A_{\lambda}$  and  $Q = A_{\text{EWA}}$  we get the following expression:  $\langle (I - A_{\lambda})^2 Y | \Sigma(A_{\lambda} - A_{\text{EWA}}) Y \rangle_n = \langle (I - A_{\text{EWA}})^2 Y | \Sigma(A_{\lambda} - A_{\text{EWA}}) Y \rangle_n - 2 \langle (I - \frac{\hat{A} + A_{\lambda}}{2})(A_{\text{EWA}} - A_{\lambda}) Y | \Sigma(A_{\text{EWA}} - A_{\lambda}) Y \rangle_n$ . When one integrates over  $\Lambda$  with respect to the measure  $\theta \cdot \pi$ , the term of the first scalar product in the RHS of the last equation vanishes. On the other hand, positive semidefiniteness of matrices  $A_{\lambda}$  implies that of the matrix  $A_{\text{EWA}}$  and, therefore,  $\langle (I - \frac{\hat{A} + A_{\lambda}}{2})(A_{\text{EWA}} - A_{\lambda}) Y | \Sigma(A_{\text{EWA}} - A_{\lambda}) Y \rangle_n \leq \langle (A_{\text{EWA}} - A_{\lambda}) Y | \Sigma(A_{\text{EWA}} - A_{\lambda}) Y \rangle_n$ . This inequality, in conjunction with (34) implies that

$$\begin{aligned} \hat{r}_{\text{EWA}} &\leq \int_{\Lambda} \left( \hat{r}_{\lambda} - \|\hat{f}_{\lambda} - \hat{f}_{\text{EWA}}\|_n^2 + \frac{8}{\beta} \langle (A_{\text{EWA}} - A_{\lambda}) Y | \Sigma(A_{\text{EWA}} - A_{\lambda}) Y \rangle_n \right) \theta(\lambda) \pi(d\lambda) \\ &= \int_{\Lambda} \left( \hat{r}_{\lambda} - \|\hat{f}_{\lambda} - \hat{f}_{\text{EWA}}\|_n^2 + \frac{8}{\beta} \langle \hat{f}_{\text{EWA}} - \hat{f}_{\lambda} | \Sigma(\hat{f}_{\text{EWA}} - \hat{f}_{\lambda}) \rangle_n \right) \theta(\lambda) \pi(d\lambda) \\ &\leq \int_{\Lambda} \left( \hat{r}_{\lambda} - \left(1 - \frac{8 \max_i \sigma_i^2}{\beta}\right) \|\hat{f}_{\lambda} - \hat{f}_{\text{EWA}}\|_n^2 \right) \theta(\lambda) \pi(d\lambda). \end{aligned}$$

Taking into account the fact that  $\beta \geq 8 \max_i \sigma_i^2$ , we get  $\hat{r}_{\text{EWA}} \leq \int_{\Lambda} \hat{r}_{\lambda} \theta(\lambda) \pi(d\lambda) \leq \int_{\Lambda} \hat{r}_{\lambda} \hat{\pi}(d\lambda) + \frac{\beta}{n} \mathcal{K}(\hat{\pi}, \pi)$ . To conclude, it suffices to remark that  $\hat{\pi}$  is the probability measure minimizing the criterion  $\int_{\Lambda} \hat{r}_{\lambda} p(d\lambda) + \frac{\beta}{n} \mathcal{K}(p, \pi)$  among all  $p \in \mathcal{P}_{\Lambda}$  (see for instance Catoni (2004) p.160). Thus, for every  $p \in \mathcal{P}_{\Lambda}$ , it holds that

$$\hat{r}_{\text{EWA}} \leq \int_{\Lambda} \hat{r}_{\lambda} p(d\lambda) + \frac{\beta}{n} \mathcal{K}(p, \pi).$$

Taking the expectation of both sides, the desired result follows with Fatou's lemma.  $\blacksquare$

**Proof** [ when  $\mathbf{C}_1$  is satisfied] We can do the same calculation as when  $C_2$  is satisfied until (32). In view of Equations (2) and (4), as well as the assumptions  $A_\lambda^2 = A_\lambda^\top = A_\lambda$  and  $A_\lambda^\top \mathbf{b}_\lambda \equiv 0$ , we get

$$\nabla_{\mathbf{Y}} \hat{r}_\lambda = \frac{2}{n} (I_{n \times n} - A_\lambda)^\top (I_{n \times n} - A_\lambda) \mathbf{Y} - \frac{2}{n} (I_{n \times n} - A_\lambda)^\top \mathbf{b}_\lambda = \frac{2}{n} (I_{n \times n} - A_\lambda) \mathbf{Y} - \frac{2}{n} \mathbf{b}_\lambda. \quad (35)$$

Using the same shorthand  $I = I_{n \times n}$  with Eq. (35) we come up with

$$\hat{r}_{\text{EWA}} = \int_{\Lambda} \left( \hat{r}_\lambda - \|\hat{\mathbf{f}}_\lambda - \hat{\mathbf{f}}_{\text{EWA}}\|_n^2 - \frac{4}{\beta} \langle \mathbf{Y} - \hat{\mathbf{f}}_\lambda | \Sigma (\hat{\mathbf{f}}_\lambda - \hat{\mathbf{f}}_{\text{EWA}}) \rangle_n \right) \theta(\lambda) \pi(d\lambda). \quad (36)$$

Now, since  $\hat{\mathbf{f}}$  is the expectation of  $\hat{\mathbf{f}}_\lambda$  with respect to the measure  $\theta \cdot \pi$ , we have

$$\begin{aligned} \hat{r}_{\text{EWA}} &= \int_{\Lambda} \left( \hat{r}_\lambda - \|\hat{\mathbf{f}}_\lambda - \hat{\mathbf{f}}_{\text{EWA}}\|_n^2 + \frac{4}{\beta} \langle \mathbf{Y} - \hat{\mathbf{f}}_{\text{EWA}} + \hat{\mathbf{f}}_{\text{EWA}} - \hat{\mathbf{f}}_\lambda | \Sigma (\hat{\mathbf{f}}_\lambda - \hat{\mathbf{f}}_{\text{EWA}}) \rangle_n \right) \theta(\lambda) \pi(d\lambda) \\ &= \int_{\Lambda} \left( \hat{r}_\lambda - \|\hat{\mathbf{f}}_\lambda - \hat{\mathbf{f}}_{\text{EWA}}\|_n^2 + \frac{4}{\beta} \langle \hat{\mathbf{f}}_{\text{EWA}} - \hat{\mathbf{f}}_\lambda | \Sigma (\hat{\mathbf{f}}_{\text{EWA}} - \hat{\mathbf{f}}_\lambda) \rangle_n \right) \theta(\lambda) \pi(d\lambda) \\ &\leq \int_{\Lambda} \left( \hat{r}_\lambda - \left(1 - \frac{4 \max_i \sigma_i^2}{\beta}\right) \|\hat{\mathbf{f}}_\lambda - \hat{\mathbf{f}}_{\text{EWA}}\|_n^2 \right) \theta(\lambda) \pi(d\lambda). \end{aligned}$$

Taking into account the fact that  $\beta \geq 4 \max_i \sigma_i^2$ , we get the same results as with condition  $\mathbf{C}_2$ :  $\hat{r}_{\text{EWA}} \leq \int_{\Lambda} \hat{r}_\lambda \theta(\lambda) \pi(d\lambda) \leq \int_{\Lambda} \hat{r}_\lambda \hat{\pi}(d\lambda) + \frac{\beta}{n} \mathcal{K}(\hat{\pi}, \pi)$ . The end of the proof is unchanged and leads to the same general result as with condition  $\mathbf{C}_2$ , except for the choice of  $\alpha$ . ■

### Appendix C. Continuous oracle inequality

**Proposition 2** Let  $\Lambda \subset \mathbb{R}^M$  be an open and bounded set and let  $\pi$  be the uniform probability on  $\Lambda$ . Assume that the mapping  $\lambda \mapsto r_\lambda$  is Lipschitz continuous, i.e.,  $|r_{\lambda'} - r_\lambda| \leq L_r \|\lambda' - \lambda\|_2$ ,  $\forall \lambda, \lambda' \in \Lambda$ . Under the conditions  $\mathbf{C}_1$  or  $\mathbf{C}_1$  aggregate  $\hat{\mathbf{f}}_{\text{EWA}}$  satisfies the inequality

$$\mathbb{E} \|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2 \leq \inf_{\lambda \in \Lambda} \left\{ \mathbb{E} \|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|_n^2 + \frac{\beta M}{n} \log \left( \frac{\sqrt{M}}{2 \min(n^{-1}, d_2(\lambda, \Lambda))} \right) \right\} + \frac{L_r + \beta \log(\text{Leb}(\Lambda))}{n}.$$

for every  $\beta \geq \alpha \|\Sigma\|$  where  $\alpha = 4$  if  $\mathbf{C}_1$  holds true and  $\alpha = 8$  if  $\mathbf{C}_2$  holds true.

**Proof** It suffices to apply Theorem 1 and to bound from above the RHS of inequality (9)

$$\begin{aligned} \mathbb{E} (\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) &\leq \inf_{p \in \mathcal{P}_{\Lambda}} \left( \int_{\Lambda} r_\lambda p(d\lambda) + \frac{\beta}{n} \mathcal{K}(p, \pi) \right) \\ \mathbb{E} (\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) &\leq \inf_{p \in \mathcal{P}_{\Lambda}} \left( \int_{\Lambda} [|r_\lambda - r_{\lambda_0}| + r_{\lambda_0}] p(d\lambda) + \frac{\beta}{n} \mathcal{K}(p, \pi) \right) \end{aligned}$$

Then, the RHS of the last inequality can be bounded from above by the minimum over all measures having as density  $p_{\lambda_0, \tau_0}(\lambda) = \mathbb{1}_{B_{\lambda_0}(\tau_0)}(\lambda) / \text{Leb}(B_{\lambda_0}(\tau_0))$ , with  $\lambda_0 \in \Lambda$  and  $\tau_0 = \min(1/n, d_2(\lambda_0, \Lambda))$

(hence  $B_{\lambda_0}(\tau_0) \subset \Lambda$ ). Using the Lipschitz condition on  $r_\lambda$ , the bound on the risk becomes

$$\begin{aligned}\mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) &\leq \int_{\Lambda} [|r_\lambda - r_{\lambda_0}| + r_{\lambda_0}] p_{\lambda_0, \tau_0}(d\lambda) + \frac{\beta}{n} \mathcal{K}(p_{\lambda_0, \tau_0}, \pi) \\ \mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) &\leq r_{\lambda_0} + L_r \int_{\Lambda} \|\lambda - \lambda_0\|_2 p_{\lambda_0, \tau_0}(d\lambda) + \frac{\beta}{n} \mathcal{K}(p_{\lambda_0, \tau_0}, \pi) \\ \mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) &\leq r_{\lambda_0} + L_r \tau_0 + \frac{\beta}{n} \mathcal{K}(p_{\lambda_0, \tau_0}, \pi)\end{aligned}\tag{37}$$

Now, since  $\lambda_0$  is such that  $B_{\lambda_0}(\tau_0) \subset \Lambda$ , the measure  $p_{\lambda_0, \tau_0}(\lambda) d\lambda$  is absolutely continuous w.r.t.  $\pi$  and the Kullback-Leibler divergence between these measures equals  $\log\{\text{Leb}(\Lambda)/\text{Leb}(B_{\lambda_0}(\tau_0))\}$ . By the simple inequality  $\|x\|_2^2 \leq M\|x\|_\infty^2$  for any  $x \in \mathbb{R}^M$ , one can see that the Euclidean ball of radius  $\tau_0$  contains the hypercube of width  $\frac{2\tau_0}{\sqrt{M}}$ . So we have the following lower bound for the volume  $B_{\lambda_0}$ :  $\text{Leb}(B_{\lambda_0}(\tau_0)) \geq (2\tau_0/\sqrt{M})^M$ . By combining this with inequality (37) the results of Proposition 3 is straightforward. ■

## Appendix D. Sparsity oracle inequality

Let us choose a prior  $\pi$  that promotes the sparsity of  $\lambda$ . This can be done in the same vein as in Dalalyan and Tsybakov Dalalyan and Tsybakov (2007, 2008), by means of the heavy tailed prior (Student  $t(3)$  distribution):

$$\pi(d\lambda) \propto \prod_{j=1}^M \frac{1}{(1 + |\lambda_j/\tau|^2)^2} \mathbb{1}_\Lambda(\lambda),\tag{38}$$

where  $\tau > 0$  is a tuning parameter, that takes small values.

**Proposition 3** *Let  $\Lambda = \mathbb{R}^M$  and let  $\pi$  be defined by (12). Assume that the mapping  $\lambda \mapsto r_\lambda$  is continuously differentiable and, for some  $M \times M$  matrix  $\mathcal{M}$ , satisfies:*

$$r_\lambda - r_{\lambda'} - \nabla r_{\lambda'}^\top (\lambda - \lambda') \leq (\lambda - \lambda')^\top \mathcal{M}(\lambda - \lambda'), \quad \forall \lambda, \lambda' \in \Lambda.$$

*If either one of the conditions **C**<sub>1</sub> and **C**<sub>2</sub> (cf. Theorem 1) is fulfilled, then the aggregate  $\hat{\mathbf{f}}_{\text{EWA}}$  defined by Equations (7) and (4) satisfies the inequality*

$$\mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) \leq \inf_{\lambda \in \mathbb{R}^M} \left\{ \mathbb{E} \|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|_n^2 + \frac{4\beta}{n} \sum_{j=1}^M \log\left(1 + \frac{|\lambda_j|}{\tau}\right) \right\} + \text{Tr}(\mathcal{M})\tau^2$$

*provided that  $\beta \geq \alpha \max_{i=1, \dots, n} \sigma_i^2$ , where  $\alpha = 4$  if **C**<sub>1</sub> holds true and  $\alpha = 8$  if **C**<sub>2</sub> holds true.*

**Proof** The proof is a simplified version of proofs given in Dalalyan and Tsybakov (2007, 2008), since  $\Lambda$  is the whole space,  $\Lambda = \mathbb{R}^M$  instead of a bounded subset of  $\mathbb{R}^M$ .

We begin the proof as for the previous proposition, but pushing the development of the function  $\lambda \rightarrow r_\lambda$  up to second order. So, for any  $\lambda^* \in \mathbb{R}^M$ , we have

$$\begin{aligned}\mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) &\leq \inf_{\lambda^* \in \mathbb{R}^M} \left( r_{\lambda^*} + \int_{\Lambda} (\nabla r_{\lambda^*}^\top (\lambda - \lambda^*) + (\lambda - \lambda^*)^\top \mathcal{M}(\lambda - \lambda^*)) p_{\lambda^*}(d\lambda) + \frac{\beta}{n} \mathcal{K}(p_{\lambda^*}, \pi) \right)\end{aligned}$$

By choosing  $p_{\lambda^*}(\lambda) = \pi(\lambda - \lambda^*)$  for any  $\lambda \in \mathbb{R}^M$ , the second term in the last display vanishes since the distribution  $\pi$  is symmetric. The third term is computed thanks to the moment of order 2 of a scaled Student  $t(3)$  distribution. Recall that if  $T$  is drawn from the scaled Student  $t(3)$  distribution, its distribution function is  $u \rightarrow 2/[\pi(1+u^2)^2]$ , and that  $\mathbb{E}T^2 = 1$ . Thus, we have that  $\int_{\Lambda} \lambda_1^2 \pi(\lambda) d\lambda = \tau^2$ . We can then bound the risk of the EWA estimator by

$$\mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) \leq \inf_{\lambda^* \in \mathbb{R}^M} \left( r_{\lambda^*} + \text{Tr}(\mathcal{M})\tau^2 + \frac{\beta}{n} \mathcal{K}(p_{\lambda^*}, \pi) \right) \quad (39)$$

So far, the particular choice of heavy tailed prior has not been used. This choice is important to control the Kullback-Leibler divergence between two translated versions of the same distribution

$$\begin{aligned} \mathcal{K}(p_{\lambda^*}, \pi) &= \int_{\Lambda} \log \left[ \prod_{j=1}^M \frac{(\tau^2 + \lambda_j^2)^2}{(\tau^2 + (\lambda_j - \lambda_j^*)^2)^2} \right] p_{\lambda^*}(d\lambda) \\ \mathcal{K}(p_{\lambda^*}, \pi) &= 2 \sum_{j=1}^M \int_{\Lambda} \log \left[ \frac{\tau^2 + \lambda_j^2}{\tau^2 + (\lambda_j - \lambda_j^*)^2} \right] p_{\lambda^*}(d\lambda). \end{aligned}$$

We bound the quotient in the above equality by

$$\begin{aligned} \frac{\tau^2 + \lambda_j^2}{\tau^2 + (\lambda_j - \lambda_j^*)^2} &= 1 + \frac{2\tau(\lambda_j - \lambda_j^*)}{\tau^2 + (\lambda_j - \lambda_j^*)^2} \frac{\lambda_j^*}{\tau} + \frac{(\lambda_j^*)^2}{\tau^2 + (\lambda_j - \lambda_j^*)^2} \\ \frac{\tau^2 + \lambda_j^2}{\tau^2 + (\lambda_j - \lambda_j^*)^2} &\leq 1 + \left| \frac{\lambda_j^*}{\tau} \right| + \left( \frac{\lambda_j^*}{\tau} \right)^2 \leq \left( 1 + \left| \frac{\lambda_j^*}{\tau} \right| \right)^2. \end{aligned}$$

Since the last inequality is independent of  $\lambda$ , the integral disappears ( $p_{\lambda^*}$  is a probability measure) in the previous bound on the Kullback-Leibler divergence, so we eventually get

$$\mathcal{K}(p_{\lambda^*}, \pi) \leq 4 \sum_{j=1}^M \log \left( 1 + \left| \frac{\lambda_j^*}{\tau} \right| \right),$$

and combine with Inequality (39), this ends the proof of the proposition. ■

## Appendix E. Application to minimax estimation

Let us denote by  $\theta_k(f) = \langle f | \varphi_k \rangle_n$  the coefficients of the (orthogonal) discrete sine transform of  $f$  and the Sobolev ellipsoids  $\mathcal{F}(\alpha, R) = \{f \in \mathbb{R}^n : \sum_{k=1}^n k^{2\alpha} \theta_k(f)^2 \leq R\}$ . Assume in this section that the noise is homoscedastic ( $\Sigma = \sigma^2 I_{n \times n}$ ) and  $A_{\alpha, w} = \mathcal{D}^\top \text{diag}((1 - k^\alpha / w)_+ ; k = 1, \dots, n) \mathcal{D}$  is the Pinsker filter in the discrete sine basis.

**Proposition 5** *Let  $\lambda = (\alpha, w) \in \Lambda = \mathbb{R}_+^2$  and consider the prior*

$$\pi(d\lambda) = \frac{2n_\sigma^{-\alpha/(2\alpha+1)}}{(1 + n_\sigma^{-\alpha/(2\alpha+1)} w)^3} e^{-\alpha} d\alpha dw,$$

where  $n_\sigma = n/\sigma^2$ . Then, in model (1) with homoscedastic errors, the aggregate  $\hat{\mathbf{f}}_{\text{EWA}}$  based on the temperature  $\beta = 8\sigma^2$  and the constituent estimators  $\hat{\mathbf{f}}_{\alpha,w} = A_{\alpha,w}\mathbf{Y}$  (with  $A_{\alpha,w}$  being the Pinsker filter) is adaptive in the exact minimax sense on the family of classes  $\{\mathcal{F}(\alpha, R) : \alpha > 0, R > 0\}$ .

**Proof** We assume, without loss of generality, that the matrix  $n^{1/2}\mathcal{D}$  coincides with the identity matrix. First, let us fix  $\alpha_0 > 0$  and  $R_0 > 0$ , such that  $n^{-1/2}\mathbf{f} \in \mathcal{F}(\alpha_0, R_0)$  and define  $\lambda_0 = (\alpha_0, w_0) \in \Lambda$  with  $w_0$  chosen such that the Pinsker estimator  $\hat{\mathbf{f}}_{\alpha_0, w_0}$  is minimax over the ellipsoid  $\mathcal{F}(\alpha_0, R_0)$ .

In what follows, we set  $n_\sigma = n/\sigma^2$  and denote by  $p_\pi$  the density of  $\pi$  w.r.t. the Lebesgue measure on  $\mathbb{R}_+^2$ :  $p_\pi(\alpha, w) = e^{-\alpha} n_\sigma^{-\alpha/(2\alpha+1)} p_w(w n_\sigma^{-\alpha/(2\alpha+1)})$ , where  $p_w$  is a probability density function supported by  $(0, \infty)$  such that  $\int u p_w(u) du = 1$ . One easily checks that

$$\int_{\mathbb{R}^2} \alpha p_\pi(\alpha, w) d\alpha dw = 1, \quad \int_{\mathbb{R}^2} w p_\pi(\alpha, w) d\alpha dw \leq n_\sigma^{1/2}. \quad (40)$$

Let  $\tau$  be a positive number such that  $\tau \leq \min(1, \alpha_0/(2\log w_0))$  and choose  $p_{\lambda_0, \tau}$  as a translation/dilatation of  $\pi$ , concentrating on  $\lambda_0$  when  $\tau \rightarrow 0$ :

$$p_{\lambda_0, \tau}(d\lambda) = p_\pi\left(\frac{\lambda - \lambda_0}{\tau}\right) \frac{d\lambda}{\tau^2}.$$

In view of Theorem 1,

$$\mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) \leq r_{\lambda_0} + \int_{\mathbb{R}^2} |r_{\alpha, w} - r_{\alpha_0, w_0}| p_{\lambda_0, \tau}(d\lambda) + \frac{\beta}{n} \mathcal{K}(p_{\lambda_0, \tau}, \pi). \quad (41)$$

Let us decompose the term  $r_{\alpha, w} - r_{\alpha_0, w_0}$  into two pieces:  $r_{\alpha, w} - r_{\alpha_0, w_0} = \{r_{\alpha, w} - r_{\alpha, w_0}\} + \{r_{\alpha, w_0} - r_{\alpha_0, w_0}\}$  and find upper bounds for the resulting terms. With our choice of estimator, the difference between the risk functions at  $(\alpha, w)$  and  $(\alpha, w_0)$  is:

$$\begin{aligned} n(r_{\alpha, w} - r_{\alpha, w_0}) &= \sum_{k=1}^n [((1 - k^\alpha/w)_+ - 1)^2 - ((1 - k^\alpha/w_0)_+ - 1)^2] f_k^2 \\ &\quad + \sum_{k=1}^n [((1 - k^\alpha/w)_+)^2 - ((1 - k^\alpha/w_0)_+)^2] \sigma^2 \end{aligned}$$

Since the weights of the Pinsker estimators are in  $[0, 1]$ , we have

$$n|r_{\alpha, w} - r_{\alpha, w_0}| \leq 2 \sum_{k=1}^n (f_k^2 + \sigma^2) |(1 - k^\alpha/w)_+ - (1 - k^\alpha/w_0)_+|. \quad (42)$$

For any  $x, y \in \mathbb{R}$ , the inequality  $|x_+ - y_+| \leq |x - y|$  is obvious. Combined with  $\alpha_0 \leq \alpha$  and  $w_0 \leq w$ , we have that

$$\left| \left(1 - \frac{k^\alpha}{w}\right)_+ - \left(1 - \frac{k^\alpha}{w_0}\right)_+ \right| \leq \left| \frac{k^\alpha}{w} - \frac{k^\alpha}{w_0} \right| \mathbf{1}_{\{k^\alpha \leq w\}} \leq \frac{w - w_0}{w_0}. \quad (43)$$

By virtue of Inequalities (42) and (43) we get

$$|r_{\alpha, w} - r_{\alpha, w_0}| \leq 2n^{-1} \sum_{k=1}^n (f_k^2 + \sigma^2) \frac{(w - w_0)}{w_0} \leq 2(R_0 + \sigma^2) \frac{w - w_0}{w_0}. \quad (44)$$

Similar calculations lead to a bound for the other absolute difference between risk functions:

$$\begin{aligned} |r_{\alpha, w_0} - r_{\alpha_0, w_0}| &\leq 2n^{-1} \sum_{k=1}^n (f_k^2 + \sigma^2) \frac{k^\alpha - k^{\alpha_0}}{w_0} \mathbf{1}_{\{k^{\alpha_0} \leq w_0\}} \\ &\leq 2(R_0 + \sigma^2) (w_0^{\frac{\alpha-\alpha_0}{\alpha_0}} - 1). \end{aligned} \quad (45)$$

Recall that we aim to bound the second term in the RHS of (41). To this end, we need an accurate upper bound on the integrals of the RHSs of (44) and (45) w.r.t. the probability measure  $p_{\lambda_0, \tau}$ . For the first one, we get

$$\begin{aligned} \int |r_{\alpha, w} - r_{\alpha, w_0}| p_{\lambda_0, \tau}(d\lambda) &\leq 2(R_0 + \sigma^2) w_0^{-1} \int_{\mathbb{R}^2} (w - w_0) p_{\lambda_0, \tau}(d\lambda) \\ &\leq 4n_\sigma^{1/2} w_0^{-1} \tau (R_0 + \sigma^2). \end{aligned} \quad (46)$$

Similar arguments apply to bound the integral of the second difference between risk functions:

$$\begin{aligned} \int_{\mathbb{R}^2} |r_{\alpha, w_0} - r_{\alpha_0, w_0}| p_{\lambda_0, \tau}(d\lambda) &\leq 2(R_0 + \sigma^2) \int_{\mathbb{R}^2} (w_0^{\frac{\alpha-\alpha_0}{\alpha_0}} - 1) p_{\lambda_0, \tau}(d\lambda) \\ &= \frac{2\tau(R_0 + \sigma^2) \log w_0}{\alpha_0 - \tau \log w_0} \\ &\leq 4\tau(R_0 + \sigma^2) \alpha_0^{-1} \log w_0, \end{aligned} \quad (47)$$

where we used the inequality  $\tau \leq \alpha_0 / (2 \log w_0)$ .

The last term to bound in inequality (41) requires the evaluation of the Kullback-Leibler divergence between  $p_{\lambda_0, \tau}$  and  $\pi$ . It can be done as follows:

$$\begin{aligned} \mathcal{K}(p_{\lambda_0, \tau}, \pi) &= \int_{\mathbb{R}^2} \log \left( \frac{e^{-\frac{\alpha-\alpha_0}{\tau}} p_w \left( \frac{w-w_0}{n_\sigma^{\alpha/(2\alpha+1)} \tau} \right)}{e^{-\alpha} p_w \left( \frac{w}{n_\sigma^{\alpha/(2\alpha+1)}} \right)} \frac{1}{\tau^2} \right) p_{\lambda_0, \tau}(d\lambda) \\ &= \int_{\mathbb{R}^2} \left\{ \alpha - \frac{\alpha - \alpha_0}{\tau} + \log \frac{p_w \left( \frac{w-w_0}{n_\sigma^{\alpha/(2\alpha+1)} \tau} \right)}{p_w \left( \frac{w}{n_\sigma^{\alpha/(2\alpha+1)}} \right)} \right\} p_{\lambda_0, \tau}(d\lambda) - 2 \log(\tau) \\ &\leq \alpha_0 + (\tau - 1) + \int_{\mathbb{R}_+^2} \log \left( 1 + \frac{w}{n_\sigma^{\alpha/(2\alpha+1)}} \right)^3 p_{\lambda_0, \tau}(d\lambda) - 2 \log(\tau). \end{aligned}$$

where the third equality is derived thanks to Eq. (40) and the obvious relation  $\|p_w\|_\infty = 2$ . Now, making the change of variable  $w = w_0 + \tau n_\sigma^{\alpha/(2\alpha+1)} u$  and using the fact that  $w_0 + \tau n_\sigma^{\alpha/(2\alpha+1)} u \leq n_\sigma^{\alpha/(2\alpha+1)} (w_0 + u)$ , we get

$$\begin{aligned} \int_{\mathbb{R}_+^2} \log \left( 1 + \frac{w}{n_\sigma^{\alpha/(2\alpha+1)}} \right)^3 p_{\lambda_0, \tau}(d\lambda) &\leq 3 \int_{\mathbb{R}_+} \log(1 + w_0 + u) p_w(u) du \\ &\leq 3 \log(1 + w_0 + \int_{\mathbb{R}_+} u p_w(u) du) \\ &= 3 \log(2 + w_0). \end{aligned}$$

Eventually, we can reformulate our bound on the risk of the EWA given in (41), leading to

$$\mathbb{E}(\|\hat{\mathbf{f}}_{\text{EWA}} - \mathbf{f}\|_n^2) \leq r_{\lambda_0} + 4\tau(R_0 + \sigma^2) \left( \frac{n_\sigma^{1/2}}{w_0} + \frac{\log w_0}{\alpha_0} \right) + \frac{8\sigma^2(\alpha_0 + 3\log(\frac{2+w_0}{\tau}))}{n}. \quad (48)$$

To conclude the proof of the proposition, we set

$$\tau = \frac{\alpha_0}{n_\sigma^2 + \alpha_0 + 2\log w_0}, \quad w_0 = \left( \frac{R_0(\alpha_0 + 1)(2\alpha_0 + 1)}{\alpha_0} \right)^{\frac{\alpha_0}{2\alpha_0 + 1}} n_\sigma^{\frac{\alpha_0}{2\alpha_0 + 1}}.$$

According to Pinsker's theorem (see, for instance, Tsybakov (2009), Theorem 3.2)

$$\max_{f \in \mathcal{F}(\alpha_0, R_0)} r_{\lambda_0} = (1 + o_n(1)) \min_{\hat{f}} \max_{f \in \mathcal{F}(\alpha_0, R_0)} \mathbb{E}(\|\hat{f} - f\|_n^2).$$

In view of this result, taking the max over  $f \in \mathcal{F}(\alpha_0, R_0)$  in (48), we get

$$\max_{f \in \mathcal{F}(\alpha_0, R)} \mathbb{E}(\|\hat{f}_{\text{EWA}} - f\|_n^2) \leq (1 + o_n(1)) \min_{\hat{f}} \max_{f \in \mathcal{F}(\alpha_0, R)} \mathbb{E}(\|\hat{f} - f\|_n^2) + O\left(\frac{\log n}{n}\right).$$

This leads to the desired result in view of the relation

$$\liminf_{n \rightarrow \infty} \min_{\hat{f}} \max_{f \in \mathcal{F}(\alpha_0, R)} n^{\frac{2\alpha_0}{2\alpha_0 + 1}} \mathbb{E}(\|\hat{f} - f\|_n^2) > 0,$$

which follows from (Tsybakov, 2009, Theorem 3.1). ■

## Collaborative Filtering with the Trace Norm: Learning, Bounding, and Transducing

**Ohad Shamir**

*One Memorial Drive, Cambridge MA 02142 USA*

OHADSH@MICROSOFT.COM

**Shai Shalev-Shwartz**

*Givat Ram, Jerusalem 91904, Israel*

SHAIS@CS.HUJI.AC.IL

**Editor:** Sham Kakade, Ulrike von Luxburg

### Abstract

Trace-norm regularization is a widely-used and successful approach for collaborative filtering and matrix completion. However, its theoretical understanding is surprisingly weak, and despite previous attempts, there are no distribution-free, non-trivial learning guarantees currently known. In this paper, we bridge this gap by providing such guarantees, under mild assumptions which correspond to collaborative filtering as performed in practice. In fact, we claim that previous difficulties partially stemmed from a mismatch between the standard learning-theoretic modeling of collaborative filtering, and its practical application. Our results also shed some light on the issue of collaborative filtering with bounded models, which enforce predictions to lie within a certain range. In particular, we provide experimental and theoretical evidence that such models lead to a modest yet significant improvement.

**Keywords:** Collaborative Filtering, Trace-Norm Regularization, Transductive Learning, Sample Complexity

### 1. Introduction

We consider the problem of matrix-based collaborative filtering, where the goal is to predict entries of an unknown matrix based on a subset of its observed entries. An increasingly popular approach to achieve this is via trace-norm regularization, where one seeks a matrix which agrees well with the observed entries, while constraining its complexity in terms of the trace-norm. The trace-norm is well-known to be a convex surrogate to the matrix rank, and has repeatedly shown good performance in practice (see for instance Srebro et al. (2004); Salakhutdinov and Mnih (2007); Bach (2008); Candès and Tao (2009)).

However, in terms of distribution-free guarantees, the current learning-theoretic understanding of trace-norm regularization is surprisingly weak. Most non-trivial guarantees currently known (e.g., Srebro and Shraibman (2005); Candès and Tao (2009); Candès and Recht (2009)) assume that the observed entries are sampled uniformly at random. In collaborative filtering, this is an extremely unrealistic assumption. For example, in the Netflix challenge dataset, where the matrix contains the ratings of users (rows) for movies (columns), the number and distribution of ratings differ drastically between users. Modeling such data as a uniform sample is not a reasonable assumption. Recently, Negahban and Wainwright (2010) studied the problem of matrix completion under a non-uniform distribution. However, the

analysis is still not distribution-free, and requires strong assumptions on the underlying matrix. Moreover, the results do not apply to standard trace-norm regularization, but rather to a carefully re-weighted version of trace-norm regularization.

In practice, we know that standard trace-norm regularization works well even for data which is very non-uniform. Moreover, we know that in other learning problems, one is able to derive distribution-free guarantees, and there is no a-priori reason why this should not be possible here. Nevertheless, obtaining a non-trivial guarantee for trace-norm regularization has remained elusive. As a result, some works suggested to use other complexity measures for collaborative filtering, such as the max-norm (Srebro et al. (2004); Lee et al. (2010)) and weighted trace-norm (Salakhutdinov and Srebro (2010)).

In this paper, we bridge this gap between our theoretical understanding and practical performance of trace-norm regularization. We show that by adding very mild assumptions, which correspond to collaborative filtering as performed *in practice*, it is possible to obtain non-trivial, distribution-free guarantees on learning with the trace norm. In fact, we claim that the difficulties in providing such guarantees partially stemmed from a mismatch between the standard theoretical modeling of collaborative filtering, and its practical application.

First, we show that one can obtain such guarantees, if one takes into account that the values to be predicted are usually bounded in practice. For example, in predicting movie ratings, it is known in advance that the ratings are on a scale of (say) 1 to 5, and practitioners usually clip their predictions to be inside this range. While this seems like an almost trivial operation, we show that taking it into account has far-reaching implications in terms of the theoretical guarantees. The proof relies on a decomposition technique which might also be useful for regularizers other than the trace-norm.

Second, we argue that the standard inductive model of learning, where the training data is assumed to be sampled i.i.d. from some distribution, may not be the best way to analyze collaborative filtering. Instead, we look at the transductive model, where sampling of the data is done without replacement. In the context of collaborative filtering, we show this makes a large difference in terms of the attainable guarantees.

Our results show that a transductive model, and boundedness assumptions, play an important role in obtaining distribution-free guarantees. This relates to a line of recent works, which suggest to incorporate prior knowledge on the range of predicted values into the learning process, by explicitly bounding the predictions. We provide an empirical study, which indicates that this indeed provides a modest, yet significant, improvement in performance, and corroborates our theoretical findings.

## 2. Setting

Our goal is to predict entries of an unknown  $m \times n$  matrix  $X$ , based on a random subset of observed entries of  $X$ . A common way to achieve this, following standard learning approaches, is to find an  $m \times n$  matrix  $W$  from a constrained class of matrices  $\mathcal{W}$ , which minimizes the discrepancy from  $X$  on the observed entries. More precisely, if we let  $S = \{i_\alpha, j_\alpha\}$  denote the set of (row,column) observed entries, and  $\ell$  is a loss function measuring the discrepancy between the predicted and actual value, then we solve the optimization

problem

$$\min_{W \in \mathcal{W}} \frac{1}{|S|} \sum_{\alpha=1}^{|S|} \ell(W_{i_\alpha, j_\alpha}, X_{i_\alpha, j_\alpha}), \quad (1)$$

An important and widely used class of matrices  $\mathcal{W}$  are those with bounded *trace-norm*. Given a matrix  $W$ , its trace-norm  $\|W\|_{tr}$  is defined as the sum of the singular values. The class of matrices with bounded trace-norm has several useful properties, such as it being a convex approximation of the class of rank-bounded matrices (see e.g. Srebro and Shraibman (2005)). The trace-norm of  $m \times n$  matrices with bounded entries is typically on the order of  $\sqrt{mn}$ , and thus we will focus on classes with trace norm bounded by  $t = \Theta(\sqrt{mn})$ .

For now, we will consider the inductive model of learning, which parallels the standard agnostic-PAC learnability framework. The model is defined as follows: We assume there exists an unknown distribution  $\mathcal{D}$  over  $\{1, \dots, m\} \times \{1, \dots, n\}$ . Each instantiation  $(i, j)$  provides the value  $X_{i,j}$  of an entry at a randomly picked row  $i$  and column  $j$ . An i.i.d. sample  $S = \{i_\alpha, j_\alpha\}$  of indices is chosen, and the corresponding entries  $\{X_{i_\alpha, j_\alpha}\}$  are revealed. Our goal is to find a matrix  $W \in \mathcal{W}$  such that its risk (or generalization error),  $\mathbb{E}_{(i,j) \sim \mathcal{D}} [\ell(W_{i,j}, X_{i,j})]$ , is as close as possible to the smallest possible risk over all  $W \in \mathcal{W}$ . It is well-known that this can be achieved by solving the optimization problem in Eq. (1), if we can provide a non-trivial uniform sample complexity bound, namely a bound on

$$\sup_{W \in \mathcal{W}} \left( \mathbb{E}_{i,j} [\ell(W_{i,j}, X_{i,j})] - \frac{1}{|S|} \sum_{\alpha=1}^{|S|} \ell(W_{i_\alpha, j_\alpha}, X_{i_\alpha, j_\alpha}) \right). \quad (2)$$

A major focus of this paper is studying the difficulties and possibilities of obtaining such bounds.

### 3. Sample Complexity Bounds for the Trace-Norm

Consider the class of trace-norm constrained matrices,  $\mathcal{W} = \{W : \|W\|_{tr} \leq t\}$ , where we assume  $t = \Theta(\sqrt{mn})$ . Although learning with respect to this class is widely used in collaborative filtering, understanding its generalization and sample-complexity properties has proven quite elusive. In Srebro and Shraibman (2005), sample complexity bounds of the form  $O(\sqrt{(m+n)/|S|})$  (ignoring logarithmic factors) were obtained under the strong assumption of a uniform distribution over the matrix entries. However, this assumption does not correspond to real-world collaborative filtering datasets, where the distribution of the revealed entries appears to be highly non-uniform. Other works, which focused on exact matrix completion, such as Candès and Tao (2009); Candès and Recht (2009), also assume a uniform sampling distribution.

While the bounds in Srebro and Shraibman (2005) can be adapted to a non-uniform distribution, they lead to bounds which are no better than  $O((m+n)/\sqrt{|S|})$ . This implies a required sample size comparable or larger than the total number of entries in the matrix. It is a trivial bound, since the entire goal of collaborative filtering is prediction based on observing just a small subset of the matrix entries.

The analysis mentioned above used Rademacher complexity to quantify the richness of the hypothesis class  $\mathcal{W}$ , and will be utilized in our analysis as well. Formally, we define

the Rademacher complexity of a hypothesis class  $\mathcal{W}$  combined with a loss function  $\ell$ , with respect to a sample  $S$ , as

$$R_S(\ell \circ \mathcal{W}) = \mathbb{E}_{\sigma} \left[ \sup_{W \in \mathcal{W}} \frac{1}{|S|} \sum_{\alpha=1}^{|S|} \sigma_{\alpha} \ell(W_{i_{\alpha}, j_{\alpha}}, X_{i_{\alpha}, j_{\alpha}}) \right], \quad (3)$$

where  $\sigma_1, \dots, \sigma_{|S|}$  are i.i.d. random variables taking the values  $-1$  and  $+1$  with equal probability.

Rademacher complexities play a key role in obtaining sample complexity bounds, either in expectation or in high probability. The following is a typical example (based on (Boucheron and Lugosi, 2005, Theorem 3.2)):

**Theorem 1** *The expected value of Eq. (2) is at most  $2R_S(\ell \circ \mathcal{W})$ . Moreover, if there is a constant  $b_{\ell}$  such that  $\sup_{i,j, W \in \mathcal{W}} |\ell(W_{i,j}, X_{i,j})| \leq b_{\ell}$ , then for any  $\delta \in (0, 1)$ , Eq. (2) is bounded with probability at least  $1 - \delta$  by  $2R_S(\ell \circ \mathcal{W}) + b_{\ell} \sqrt{2 \log(2/\delta)/|S|}$ .*

Thus, to get non-trivial learning guarantees, one can focus on effectively bounding the Rademacher complexity  $R_S(\ell \circ \mathcal{W})$ .

Unfortunately, for the class  $\mathcal{W} = \{W : \|W\|_{tr} \leq t\}$  and general distributions and losses, the analysis of  $R_S(\ell \circ \mathcal{W})$  performed in Srebro and Shraibman (2005) appears to be tight, yet still leads to vacuous results. The main drive of our paper is that by modifying the setting in some very simple ways, which often correspond to collaborative filtering as done in practice, one can obtain non-trivial learning guarantees.

#### 4. Results for the Inductive Model

In this section, we show that by introducing *boundedness* conditions into the learning problem, one can obtain non-trivial bounds on the Rademacher complexity, and hence on the sample complexity of learning with trace-norm constraints.

We will start with the case where we actually learn with respect to the hypothesis class of trace-norm-constrained matrices,  $\mathcal{W} = \{W : \|W\|_{tr} \leq t\}$ , and the only boundedness is in terms of the loss function:

**Theorem 2** *Consider the hypothesis class  $\mathcal{W} = \{W : \|W\|_{tr} \leq t\}$ . Suppose that for all  $i, j$  the loss function  $\ell(\cdot, X_{i,j})$  is both  $b_{\ell}$ -bounded and  $l_{\ell}$ -Lipschitz in its first argument: Namely, that  $|\ell(W_{i,j}, X_{i,j})| \leq b_{\ell}$  for any  $W, i, j$ , and that  $\frac{|\ell(W_{i,j}, X_{i,j}) - \ell(W'_{i,j}, X_{i,j})|}{|W_{i,j} - W'_{i,j}|} \leq l_{\ell}$  for any  $W, W', i, j$ . Then*

$$R_S(\ell \circ \mathcal{W}) \leq \sqrt{9C l_{\ell} b_{\ell} \frac{t(\sqrt{m} + \sqrt{n})}{|S|}},$$

where  $C$  is the universal constant appearing in Thm. 7.

Assuming  $t = \Theta(\sqrt{mn})$ , the theorem implies that a sample of size  $O(n\sqrt{m} + m\sqrt{n})$  is sufficient to obtain good generalization performance. We note that the boundedness assumption is non-trivial, since the trace-norm constraint does not imply entries of constant magnitude (the entries can be as large as  $t$  for a matrix whose trace norm is  $t$ ). On the other

hand, as discussed earlier, the obtainable bound on the Rademacher complexity without a boundedness assumption is no better than  $O((m + n)/\sqrt{|S|})$ , which leads to a trivial required sample size of  $O((m + n)^2)$ . Moreover, we emphasize that the result makes no assumptions on the underlying distribution from which the data was sampled. The proof is presented in Subsection 7.1. We note that it relies on a decomposition technique which might also be useful for regularizers other than the trace-norm.

An alternative way to introduce boundedness, and get a non-trivial guarantee, is by composing the entries of a matrix  $W$  with a bounded transfer function. In particular, rather than just learning a matrix  $W$  with bounded trace-norm, we can learn a model  $\phi \circ W$ , where  $W$  has bounded trace-norm, and  $\phi : \mathbb{R} \mapsto I$  is a fixed mapping of each entry of  $W$  into some bounded interval  $I \subseteq \mathbb{R}$ . This model is used in practice, and is useful in the common situation where the entries of  $X$  are known to be in a certain bounded interval. In Sec. 6, we return to this model in greater depth. In terms of the theoretical guarantee, one can provide a result substantially similar to Thm. 2, without assuming boundedness of the loss function. The proof is provided in the appendix - it uses similar techniques to the one of Thm. 2, but applies them somewhat differently.

**Theorem 3** *Consider the hypothesis class  $\mathcal{W} = \{\phi \circ W : \|W\|_{tr} \leq t\}$ . Let  $\phi : \mathbb{R} \mapsto [-b_\phi, b_\phi]$  be a bounded  $l_\phi$ -Lipschitz function, and suppose that for all  $i, j$ ,  $\ell(\cdot, X_{i,j})$  is  $l_\ell$ -Lipschitz on the domain  $[-b_\phi, b_\phi]$ . Then*

$$R_S(\ell \circ \mathcal{W}) \leq l_\ell \sqrt{9Cl_\phi b_\phi \frac{t(\sqrt{m} + \sqrt{n})}{|S|}},$$

where  $C$  is the universal constant appearing in Thm. 7.

The bound in this theorem scales similar to Thm. 2, in terms of its dependence on  $m, n$ . Another possible variant is directly learning a matrix  $W$  with both a constraint on the trace-norm, as well as an  $\infty$ -norm constraint (i.e.  $\max_{i,j} |W_{i,j}|$ ) which forces the matrix entries to be constant. This model has some potential benefits which shall be further discussed in Sec. 8.

**Theorem 4** *Consider the hypothesis class  $\mathcal{W} = \{W : \|W\|_{tr} \leq t, \|W\|_\infty \leq b\}$ , where  $\|W\|_\infty = \max_{i,j} |W_{i,j}|$ . Suppose that for all  $i, j$ ,  $\ell(\cdot, X_{i,j})$  is  $l_\ell$ -Lipschitz on the domain  $[-b, b]$ . Then*

$$R_S(\ell \circ \mathcal{W}) \leq l_\ell \sqrt{9Cb \frac{t(\sqrt{n} + \sqrt{m})}{|S|}},$$

where  $C$  is the universal constant appearing in Thm. 7.

Assuming  $b$  is a constant (which is the reasonable assumption here), we get a similar bound as before.

So, we see that by inserting mild boundedness assumptions on the loss function or the matrix entries, it is possible to derive non-trivial guarantees for learning with trace norm constraints. These were all obtained under the standard inductive model, where we assume that our data is an i.i.d. sample from an underlying distribution. In the next section, we will

discuss a different learning model, which we argue to more closely resemble collaborative filtering as done in practice, and leads to better bounds on the Rademacher complexity, without making boundedness assumptions.

## 5. Improved Results for the Transductive Model

In the inductive model we have considered so far, the goal is to predict well with respect to an unknown distribution over matrix entries, given an i.i.d. sample from that distribution. The *transductive* learning model (see for instance Vapnik (1998)) is different, in that our goal is to predict well with respect to a *specific* subset of entries, whose location is known in advance. More formally, we fix an arbitrary subset of  $S$  entries, and then split it uniformly at random into two subsets  $S_{train} \cup S_{test}$ . We are then given the values of the entries in  $S_{train}$ , and our goal is to predict the values of the entries in  $S_{test}$ . For simplicity, we will assume that  $|S_{train}| = |S_{test}| = |S|/2$ , but our results can be easily generalized to more general partitions.

We note that this procedure is *exactly* the one often performed in experiments reported in the literature: Given a dataset of entries, one randomly splits it into a training set and a test set, learns a matrix on the training set, and measures its performance on the test set (e.g., Toh and Yun (2009); Jaggi and Sulovský (2010)). Even for other train-test split methods, such as holding out a certain portion of entries from each row, the transductive model seems closer to reality than the inductive model. Moreover, the transductive model captures another important feature of real-world collaborative filtering: the fact that no entry is repeated in the training set. In contrast, in the inductive model the training set is collected i.i.d., so the same entry might be sampled several time over. In fact, this is virtually certain to happen whenever the sample size is at least on the order of  $\sqrt{mn}$ , due to the birthday paradox. This does not appear to be a mere technicality, since the proofs of our theorems in the inductive model have to rely on a careful separation of the entries according to the number of times they were sampled. However, in reality each entry appears in the dataset only once, matching the transductive learning setting.

To analyze the transductive model, we require analogues of the tools we have for the inductive model, such as the Rademacher complexity. Fortunately, such analogues were already obtained in El-Yaniv and Pechyoni (2009), and we will rely on their results. In particular, based on Theorem 1 in that paper, we can use our notion of Rademacher complexity, as defined in Eq. (3), to provide sample complexity bounds in the transductive model<sup>1</sup>:

**Theorem 5** *Fix a hypothesis class  $\mathcal{W}$ , and suppose that  $\sup_{i,j,W \in \mathcal{W}} |\ell(W_{i,j}, X_{i,j})| \leq b_\ell$ . Let a set  $S$  of  $\geq 2$  distinct indices be fixed, and suppose it is uniformly and randomly split to two equal subsets  $S_{train}, S_{test}$ . Then with probability at least  $1 - \delta$  over the random split, it*

---

1. In El-Yaniv and Pechyoni (2009), a more general notion of transductive Rademacher complexity was defined, where the  $\sigma_\alpha$  random variables could also take 0 values. However, when  $|S_{train}| = |S_{test}|$ , that complexity can always be upper bounded by the standard definition of Rademacher complexity - see Lemma 1 in their paper.

holds for any  $W \in \mathcal{W}$  that

$$\begin{aligned} & \frac{\sum_{(i,j) \in S_{test}} \ell(W_{i,j}, X_{i,j})}{|S_{test}|} - \frac{\sum_{(i,j) \in S_{train}} \ell(W_{i,j}, X_{i,j})}{|S_{train}|} \\ & \leq 4R_S(\ell \circ \mathcal{W}) + \frac{b_\ell \left( 11 + 4\sqrt{\log(1/\delta)} \right)}{\sqrt{|S_{train}|}}. \end{aligned}$$

We now present our main result for the transductive model, which implies non-trivial bounds on the Rademacher complexity of matrices with constrained trace-norm. Unlike the inductive model, here we make no additional boundedness assumptions, yet the bound is superior. The proof appears in Subsection 7.2.

**Theorem 6** Consider the hypothesis class  $\mathcal{W} = \{W : \|W\|_{tr} \leq t\}$ . Then in the transductive model, if we let  $N_i = \max_i |\{j : (i, j) \in S\}|$  and  $N_j = \max_j |\{i : (i, j) \in S\}|$ , then

$$R_S(\ell \circ \mathcal{W}) \leq Cl_\ell \frac{3t(\sqrt{m} + \sqrt{n})}{2|S|},$$

where  $C$  is the universal constant appearing in Thm. 7. Alternatively, we also have

$$R_S(\ell \circ \mathcal{W}) \leq Cl_\ell \frac{t \max \{ \max_i \sqrt{N_i}, \max_j \sqrt{N_j} \}}{|S|} \sqrt[4]{\log(\min\{m, n\})},$$

where  $C$  is the universal constant appearing in Thm. 8.

We note that the second bound, while containing an additional logarithmic term, depends on the distribution of the entries, and can be considerably tighter than the worst-case. To see this, suppose as usual that  $t = \Theta(\sqrt{mn})$ , and (for simplicity)  $m = n$ . Then in the worst-case, the bound becomes meaningful when  $|S| = \Omega(n^{3/2})$ . However, if the entries in  $S$  are (approximately) uniformly distributed throughout the matrix, then the maximal number of entries in each row and column is  $O(|S|/n)$ . In that case, we obtain the bound

$$R_S(\ell \circ \mathcal{W}) \leq \tilde{O} \left( \sqrt{\frac{n}{|S|}} \right)$$

(ignoring logarithmic factors), which is already meaningful when  $|S| = \tilde{\Omega}(n)$ . Interestingly, this bound is similar (up to logarithmic factors) to previous bounds in the inductive setting, such as in Srebro and Shraibman (2005), which relied on a uniform distribution assumption. However, our Rademacher complexity bound in Thm. 6 also applies to non-uniform distributions, and is meaningful for any distribution.

Compared to the results in Sec. 4, the result here is also superior in that the Rademacher complexity does not depend on the loss magnitude bound  $b_\ell$ . Although this factor does appear in a different term in the overall sample complexity bound (Thm. 5), we do not know if this dependence is essential. Indeed, in the inductive setting, such terms appear only in high-probability bounds, and even then it is possible to prove oracle inequalities which depend merely on  $\sqrt{b_\ell/|S|}$  rather than  $b_\ell/\sqrt{|S|}$  (see for instance (Srebro et al., 2010,

Theorem 1)). If such bounds can be proven in the transductive case as well, it will lead to non-trivial sample complexity results without any boundedness assumptions whatsoever<sup>2</sup>.

Another interesting feature of Thm. 6 is that the Rademacher complexity falls off at the rate of  $O(1/|S|)$  rather than  $O(1/\sqrt{|S|})$ . While such a “fast rate” is unusual in the inductive setting, here it is a natural outcome of the different modeling of the training data. This does not lead to a  $O(1/|S|)$  sample complexity bound, because the bound in Thm. 5 contains an additional low rate term  $O(1/\sqrt{|S|})$ . However, it still leads to a better bound because the low rate term is not explicitly multiplied by functions of  $m, n$  or  $t$ .

## 6. Should We Enforce Boundedness?

As mentioned previously, we often know the range of entries to be predicted (e.g. 1 to 5 for movie rating prediction). The results of Sec. 4 suggest that in the inductive model, some sort of boundedness seems essential to get non-trivial results. In the transductive model, boundedness also plays a somewhat smaller role, by appearing in the final sample-complexity bound (Thm. 5), although not in the Rademacher complexity bound (Thm. 6). These results suggest the natural idea of incorporating into the learning algorithm the prior knowledge we have on the range of entries. Indeed, several recent papers have considered the possibility of directly learning a model  $\phi \circ W$ , where  $\phi$  is usually a sigmoid function (Salakhutdinov and Mnih (2007); Ma et al. (2008); Piotte and Chabbert (2009); Kozma et al. (2009)). Another common practice (not just with trace-norm regularization) is to clip the learned matrix entries to the known range. Since our theoretical results are just upper bounds, the effect of boundedness is not sufficiently clear. Thus, it is of interest to understand how clipping or enforcing boundedness in the learning model helps in practice. We note that while bounded models have been tested experimentally, we could not find in the literature a clear empirical study of their effect, in the context of trace-norm regularization.

We conducted experiments on two standard collaborative filtering datasets, movielens100K and movielens1M<sup>3</sup>. movielens100K contains  $10^5$  ratings of 943 users for 1770 movies, while movielens1M contains  $10^6$  ratings of 6040 users for 3706 movies. All ratings are in the range [1, 5]. For each dataset, we performed a random 80% – 20% of the data to obtain a training set and a test set. We considered two hypothesis classes: trace-norm constrained matrices  $\{W : \|W\|_{tr} \leq t\}$ , and bounded trace-norm constrained matrices  $\{\phi \circ W : \|W\|_{tr} \leq t\}$ , where  $\phi$  is a sigmoid function interpolating between 1 and 5. For each hypothesis class, we trained a trace-norm regularized algorithm using the squared loss. Specifically, we used the common approach of stochastic gradient descent on a factorized representation  $W = U^\top V$ . This approach is based on the well-known fact, that the trace norm can also be defined as  $\|W\|_{tr} = \min_{W=U^\top V} \frac{1}{2} (\|U\|_F^2 + \|V\|_F^2)$ . Thus, finding the best  $W$  by optimizing a soft trace-norm regularized problem,  $\sum_{(i,j) \in S} (X_{i,j} - W_{i,j})^2 + \lambda \|W\|_{tr}$ , can be reduced to finding  $U, V$  which minimize

$$\sum_{(i,j) \in S} (X_{i,j} - U_i^\top V_j)^2 + \frac{\lambda}{2} (\|U\|_F^2 + \|V\|_F^2). \quad (4)$$

---

2. To see why, suppose that the trace norm  $t = \Theta(\sqrt{mn})$ , and the loss function is Lipschitz. Then  $b_\ell = \Theta(\sqrt{mn})$  as well, so a  $\sqrt{b_\ell}/|S|$  term will be on the order of  $\sqrt{\sqrt{mn}/|S|}$ .

3. [www.grouplens.org/node/73](http://www.grouplens.org/node/73)

Similarly, for learning bounded models, we can find  $U, V$  which minimize

$$\sum_{(i,j) \in S} \left( X_{i,j} - \phi(U_i^\top V_j) \right)^2 + \frac{\lambda}{2} (\|U\|_F^2 + \|V\|_F^2). \quad (5)$$

We note that both problems are non-convex, although for the formulation in Eq. (4), it is possibly to show there are any local minimum is also a global one.

Tuning of  $\lambda$  was performed with a validation set. Note that in practice, for computational reasons, one often constrains  $U$  and  $V$  to have a bounded number of rows. However, this forces  $W$  to have low rank, which is an additional complexity control. Since our goal is to study the performance of trace-norm constrained matrices, and not matrices which are also low-rank, we did not constrain  $U, V$  in this manner. The downside of this is that we were unable to perform experiments on very large-scale datasets, such as Netflix, and that is why we focused on the more modest-sized movielens100K and movielens1M datasets.

To estimate the performance of the learned matrix  $W$  on the test set, we used two measures which are standard in the literature: the root-mean-squared-error (RMSE),

$$\sqrt{\sum_{(i,j) \in S_{test}} \frac{(W_{i,j} - X_{i,j})^2}{|S_{test}|}},$$

and the normalized-mean-absolute-error (NMAE),

$$\sum_{i,j \in S_{test}} \frac{|W_{i,j} - X_{i,j}|}{r|S_{test}|},$$

where  $r$  is the range of possible values in  $X$  ( $5 - 1 = 4$  for our datasets).

The experiments were repeated 5 times over random train-test splits of the data, and the results are summarized in Table 1. From the table, we see that in almost all cases, clipping and bounding lead to a statistically significant improvement. However, note that in absolute terms, the improvement is rather modest, especially with the NMAE measure which is less sensitive to large mispredictions. This accords with our theoretical results: boundedness seems to be an important and useful property, but in the transductive model (corresponding to our experiments) it plays only a modest role.

Empirically, one would have expected the use of bounded models to help a lot (in absolute terms), if learning just trace-norm constrained matrices (without clipping/bounding) leads to many predictions being outside the interval  $[1, 5]$ , in which we know the ratings lie. But indeed, this does not seem to be the case. Table 2 shows the prediction with largest magnitude, over all entries in the test set, as well as the percentage of predictions which fall outside the  $[1, 5]$  interval. It is clearly evident that such out-of-interval predictions are relatively rare, and this explains why the bounding and clipping only leads to modest improvements.

We emphasize that our results should only be interpreted in the context of pure trace-norm regularization. There are many other approaches to collaborative filtering, and it is quite possible that using bounded models has more or less impact in the context of other approaches or for other application domains.

	100K (NMAE)	100K (RMSE)	1M (NMAE)	1M (RMSE)
unclipped	$0.1882 \pm 0.0005$	$0.9543 \pm 0.0019$	$0.1709 \pm 0.0003$	$0.8670 \pm 0.0016$
clipped	$0.1874 \pm 0.0005$	$0.9486 \pm 0.0018$	$0.1706 \pm 0.0002$	$0.8666 \pm 0.0016$
bounded	$0.1871 \pm 0.0004$	$0.9434 \pm 0.0023$	$0.1698 \pm 0.0002$	$0.8618 \pm 0.0017$
$\Delta$ Clipping ( $\ast 10^{-3}$ )	$0.77 \pm 0.07$	$5.7 \pm 0.6$	$0.33 \pm 0.01$	$0.48 \pm 0.04$
$\Delta$ Bounding ( $\ast 10^{-3}$ )	$0.3 \pm 0.4$	$5.2 \pm 1.5$	$0.79 \pm 0.02$	$4.8 \pm 0.1$

Table 1: Error on test set (mean and standard deviation over 5 repeats of the experiment).

The columns refer to the dataset (movielens100K or movielens1M) and the performance measure used (NMAE or RMSE). The first two rows refer to the results using the ‘unbounded’ model as in Eq. (4), with the output used as-is or clipped to the range [1 – 5]. The third row refers to the results using the ‘bounded’ model as in Eq. (5). The fourth row is the improvement in test error by clipping the predictions after learning (i.e. the difference between the first and second row). The fifth row is the additional improvement achieved by using a bounded model (i.e., the difference between the second and third row).

	100K	1M
largest value	$5.95 \pm 0.35$	$6.13 \pm 0.16$
% outside interval	$0.69 \pm 0.05$	$0.79 \pm 0.01$

Table 2: Out-of-Interval Values

## 7. Proofs

Our proofs utilize the following two theorems, which bounds the expected spectral norm of random matrices.

**Theorem 7 (Latała (2005))** *Let  $Z$  be a matrix composed of independent zero-mean entries. Then for some fixed constant  $C$ ,  $\mathbb{E}[\|Z\|_{sp}]$  is at most*

$$C \left( \max_i \sqrt{\sum_j \mathbb{E}[Z_{i,j}^2]} + \max_j \sqrt{\sum_i \mathbb{E}[Z_{i,j}^2]} + \sqrt[4]{\sum_{i,j} \mathbb{E}[Z_{i,j}^4]} \right).$$

**Theorem 8 (Seginer (2000))** *Let  $A$  be an arbitrary  $m \times n$  matrix, such that  $m, n > 1$ . Let  $Z$  denote a matrix composed of independent zero-mean entries, such that  $Z_{i,j} = A_{i,j}$  with probability 1/2 and  $Z_{i,j} = -A_{i,j}$  with probability 1/2. Then for some fixed constant  $C$ ,  $\mathbb{E}[\|A\|_{sp}]$  is at most*

$$C \sqrt[4]{\log(\min\{m, n\})} \max \left\{ \max_i \sqrt{\sum_j A_{i,j}^2}, \max_j \sqrt{\sum_i A_{i,j}^2} \right\}$$

### 7.1. Proof of Thm. 2

We write  $R_S(\ell \circ \mathcal{W})$  as

$$\frac{1}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W \in \mathcal{W}} \sum_{i,j} \Sigma_{i,j} \ell(W_{i,j}, X_{i,j}) \right],$$

where  $\Sigma$  is a matrix whose  $(i, j)$ -th entry is defined as  $\sum_{\alpha: i_\alpha = i, j_\alpha = j} \sigma_\alpha$ . A standard Rademacher analysis will usually proceed to reduce this to  $\frac{1}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W \in \mathcal{W}} \sum_{i,j} \Sigma_{i,j} W_{i,j} \right]$ . However, this will lead to a trivial bound. Instead, we will do something more refined. Given  $i, j$ , let  $h_{i,j}$  be the number of times the sample  $S$  hits entry  $i, j$ , or more precisely  $h_{i,j} = |\{\alpha : i_\alpha = i, j_\alpha = j\}|$ . Let  $p > 0$  be an arbitrary parameter to be specified later, and define

$$Y_{i,j} = \begin{cases} \Sigma_{i,j} & h_{i,j} > p \\ 0 & h_{i,j} \leq p \end{cases} \quad Z_{i,j} = \begin{cases} 0 & h_{i,j} > p \\ \Sigma_{i,j} & h_{i,j} \leq p. \end{cases} \quad (6)$$

Clearly, we have  $\Sigma = Y + Z$ . Thus, we can upper bound the Rademacher complexity by

$$\frac{1}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W \in \mathcal{W}} \sum_{i,j} Y_{i,j} \ell(W_{i,j}, X_{i,j}) \right] + \frac{1}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W \in \mathcal{W}} \sum_{i,j} Z_{i,j} \ell(W_{i,j}, X_{i,j}) \right]. \quad (7)$$

Since  $|\ell(W_{i,j}, X_{i,j})| \leq b_\ell$ , the first term can be upper bounded by

$$\frac{1}{|S|} \mathbb{E}_{\sigma} \left[ b_\ell \sum_{i,j} |Y_{i,j}| \right] = \frac{b_\ell}{|S|} \mathbb{E}_{\sigma} [\|Y\|_1]. \quad (8)$$

Using the Rademacher contraction principle<sup>4</sup>, the second term in Eq. (7) can be upper bounded by

$$\frac{b_\ell}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W \in \mathcal{W}} \sum_{i,j} Z_{i,j} W_{i,j} \right].$$

Applying Hölder's inequality, and using the fact that the spectral norm  $\|\cdot\|_{sp}$  is the dual to the trace norm  $\|\cdot\|_{tr}$ , we can upper bound the above by

$$\frac{b_\ell}{|S|} \mathbb{E}_{\sigma} \sup_{W \in \mathcal{W}} [\|Z\|_{sp} \|W\|_{tr}] = \frac{b_\ell t}{|S|} \mathbb{E}_{\sigma} [\|Z\|_{sp}]. \quad (9)$$

Combining this with Eq. (8) and substituting into Eq. (7), we get an upper bound of the form

$$\frac{b_\ell}{|S|} \mathbb{E}_{\sigma} [\|Y\|_1] + \frac{b_\ell t}{|S|} \mathbb{E}_{\sigma} [\|Z\|_{sp}].$$

Using Lemma 9 and Lemma 10, which are given below, we can upper bound this by

$$\frac{b_\ell}{\sqrt{p}} + \frac{2.2C\ell t \sqrt{p}(\sqrt{m} + \sqrt{n})}{|S|},$$

---

4. Strictly speaking, we use a slight generalization of it, where the loss function is allowed to differ w.r.t. every  $W_{i,j}$  - see (Meir and Zhang, 2003, Lemma 5)

where  $p$  is the parameter used to define  $Y$  and  $Z$  in Eq. (6). Choosing  $p = \frac{|S|b_\ell}{2.2Cl_\ell t(\sqrt{m} + \sqrt{n})}$ , we get the bound in the theorem.

**Lemma 9** *Let  $Y$  be a random matrix defined as in Eq. (6). Then*

$$\mathbb{E}[\|Y\|_1] \leq \mathbb{E} \left[ \sum_{i,j:h_{i,j}>p} \sqrt{h_{i,j}} \right] \leq \frac{|S|}{\sqrt{p}}$$

**Proof**  $\mathbb{E}[\|Y\|_1]$  equals

$$\mathbb{E} \left[ \sum_{i,j:h_{i,j}>p} |\Sigma_{i,j}| \right] = \mathbb{E} \left[ \mathbb{E} \left[ \sum_{i,j:h_{i,j}>p} \left( \left| \sum_{\alpha:(i_\alpha,j_\alpha)=(i,j)} \sigma_\alpha \right| \right) \mid \{h_{i,j}\} \right] \right]$$

The expression inside the absolute value is the sum of  $h_{i,j}$  i.i.d. random variables, and it is easily seen that its expected absolute value is at most  $\sqrt{h_{i,j}}$ . Therefore, we can upper bound the above by  $\mathbb{E}[\sum_{i,j:h_{i,j}>p} \sqrt{h_{i,j}}]$ . We can further upper bound it, in a manner which does not depend on the values of  $h_{i,j}$ , by

$$\max_{c \in \{1, \dots, mn\}} \max_{h_1, \dots, h_c \in \mathbb{R}: \forall i \ h_i > p, \sum_{i=1}^c h_i = |S|} \sum_{i=1}^c \sqrt{h_i}.$$

Note that the constraints imply that

$$|S| = \sum_{i=1}^c h_i \geq \sqrt{p} \sum_{i=1}^c \sqrt{h_i},$$

so  $\sum_{i=1}^c \sqrt{h_i}$  can be at most  $|S|/\sqrt{p}$  as required. ■

**Lemma 10** *Let  $Z$  be a random matrix defined as in Eq. (6). Then  $\mathbb{E}_\sigma[\|Z\|_{sp}]$  is at most*

$$C \left( \max_i \sqrt{\sum_{j:h_{i,j} \leq p} h_{i,j}} + \max_j \sqrt{\sum_{i:h_{i,j} \leq p} h_{i,j}} + \sqrt[4]{3 \sum_{i,j:h_{i,j} \leq p} h_{i,j}^2} \right),$$

where  $C$  is the universal constant which appears in the main theorem of Latała (2005). Moreover, this quantity can be upper bounded by  $2.2C\sqrt{p}(\sqrt{m} + \sqrt{n})$ .

**Proof** With  $h_{i,j}$  held fixed,  $Z$  is a random matrix composed of independent entries. By using Thm. 7, we only need to analyze  $\mathbb{E}[Z_{i,j}^2]$  and  $\mathbb{E}[Z_{i,j}^4]$ . For any  $i, j$ , if  $h_{i,j} \leq p$  then  $Z_{i,j}$  is a sum of  $h_{i,j}$  i.i.d. variables taking values in  $\{-1, +1\}$ . Therefore,  $\mathbb{E}[Z_{i,j}^2] = h_{i,j}$  and  $\mathbb{E}[Z_{i,j}^4] \leq 3h_{i,j}^2$ . Plugging into Thm. 7 yields the first part of the lemma. To get the second part, we can upper bound the right-hand side of the first part by

$$C\sqrt{p} \left( \sqrt{m} + \sqrt{n} + \sqrt[4]{3mn} \right) \leq C\sqrt{p} \left( \sqrt{m} + \sqrt{n} + \sqrt[4]{3/2}(\sqrt{m} + \sqrt{n}) \right) \leq 2.2C\sqrt{p}(\sqrt{m} + \sqrt{n}).$$
■

## 7.2. Proof of Thm. 6

We write  $R_S(\ell \circ \mathcal{W})$  as

$$\frac{1}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W \in \mathcal{W}} \sum_{i,j} \Sigma_{i,j} \ell(W_{i,j}, X_{i,j}) \right],$$

where  $\Sigma$  is a matrix with  $\sigma_{i,j}$  in its  $(i,j)$ -th entry, if  $(i,j) \in S$ , and 0 otherwise. By the Rademacher contraction property<sup>5</sup>, we can upper bound this by

$$\frac{l_\ell}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W \in \mathcal{W}} \sum_{i,j} \Sigma_{i,j} W_{i,j} \right].$$

By Hölder's inequality, this is at most

$$\frac{l_\ell}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W \in \mathcal{W}} \|\Sigma\|_{sp} \|W\|_{tr} \right] = \frac{l_\ell t}{|S|} \mathbb{E}_{\sigma} [\|\Sigma\|_{sp}]. \quad (10)$$

The setting so far is rather similar to the one we had in the inductive setting (see the proof of any of the theorems in Sec. 4). But now, we need to bound just the expected spectral norm of  $\Sigma$ , which is guaranteed to have only a single Rademacher variable in each entry. By applying Thm. 7 on Eq. (10), we get

$$R_S(\ell \circ \mathcal{W}) \leq C l_\ell \frac{t \left( \sqrt{N_i} + \sqrt{N_j} + \sqrt[4]{|S|} \right)}{|S|}.$$

Since  $S$  can contain at most  $m$  and  $n$  indices for any single row and column respectively, and  $\sqrt[4]{|S|} \leq \sqrt[4]{mn} \leq \frac{1}{2} (\sqrt{m} + \sqrt{n})$ , we can upper bound the above by  $3C l_\ell t (\sqrt{m} + \sqrt{n}) / (2|S|)$ .

To get the other bound in the theorem, we apply Thm. 8 instead of Thm. 8 on Eq. (10).

## 8. Discussion

In this paper, we analyzed the sample complexity of collaborative filtering with trace-norm regularization, obtaining the first non-trivial, distribution-free guarantees. Our results were based on either mild boundedness assumptions, or a switch from the standard inductive learning model to the transductive learning model. Moreover, we argue that such a transductive model may be a better way to model collaborative filtering as performed in practice, as it seems more natural and leads to a substantial difference in terms of obtainable results. Finally, we discussed the issue of learning with bounded models, and provided an empirical study which indicates that these lead to a modest yet significant improvement in performance, corroborating our theoretical findings.

On the theoretical side, one obvious open question is the tightness of our bounds. In a nutshell, if we assume that we learn  $n \times n$  matrices with a  $\Theta(n)$  trace-norm, then our bounds imply a required sample complexity of  $O(n^{3/2})$ . On the other hand, Salakhutdinov and Srebro (2010) presented an example requiring a sample complexity of at least  $\Omega(n^{4/3})$ ,

---

5. As in the inductive case, we use in fact a slight generalization where the loss function is allowed to differ w.r.t. every  $W_{i,j}$ , as in (Meir and Zhang, 2003, Lemma 5).

which applies to all the settings we have discussed earlier. Currently, we do not know how to bridge this gap. Another issue is understanding the implications of our analysis to other types of matrix regularization, such as weighted trace-norm (Salakhutdinov and Srebro (2010)).

As to the use of bounded models, we note that although they seem beneficial in our experiments, they can also lead to non-convex optimization problems. While this did not seem to hurt performance in our experiments, it might be more harmful in other datasets of applications. One possible way to enforce bounded predictions while maintaining convexity is using  $\infty$ -norm constraints, as in Thm. 4. Minimizing the average loss with respect to such constraints is a convex optimization problem, and can be done with a generic SDP solver. However, a generic solver won't scale to large datasets. Thus, designing an *efficient* convex optimization algorithm, which combines trace-norm and  $\infty$ -norm constraints, is a potentially useful, yet non-trivial challenge.

### Acknowledgements

We thank Nati Srebro and Ruslan Salakhutdinov for helpful discussions, as well as the anonymous reviewers for valuable comments.

### References

- F. Bach. Consistency of trace-norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, 2008.
- O. Bousquet Boucheron, S. and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323 – 375, 2005.
- E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9, 2009.
- E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2009.
- Ran El-Yaniv and Dmitry Pechyoni. Transductive rademacher complexity and its applications. *Journal of AI Research*, 35:193–234, 2009.
- M. Jaggi and M. Sulovský. A simple algorithm for nuclear norm regularized problems. In *ICML*, 2010.
- L. Kozma, A. Ilin, and T. Raiko. Binary principal component analysis in the netflix collaborative filtering task. In *IEEE MLSP Workshop*, 2009.
- R. Lata  . Some estimates of norms of random matrices. *Proceedings of the AMS*, 133(5): 1273–1282, 2005.
- J. Lee, B. Recht, R. Salakhutdinov, N. Srebro, and J. Tropp. Practical large-scale optimization for max-norm regularization. In *NIPS*, 2010.

- H. Ma, H. Yang, M. Lyu, and I. King. Sorec: social recommendation using probabilistic matrix factorization. In *CIKM*, 2008.
- R. Meir and T. Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.
- S. Negahban and M. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. arXiv:1009.2118, 2010.
- M. Piotte and M. Chabbert. The pragmatic theory solution to the netflix grand prize. Available at [http://www.netflixprize.com/assets/GrandPrize2009\\_BPC\\_PragmaticTheory.pdf](http://www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf), 2009.
- R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS*, 2007.
- R. Salakhutdinov and N. Srebro. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *NIPS*, 2010.
- Yoav Seginer. The expected norm of random matrices. *Combinatorics, Probability & Computing*, 9(2):149–166, 2000.
- N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *COLT*, 2005.
- N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. In *NIPS*, 2004.
- N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low-noise and fast rates. In *NIPS*, 2010.
- K. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Optimization Online*, 2009.
- V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

## Appendix A. Proof of Thm. 3

We can rewrite the definition of  $R_S(\ell \circ \mathcal{W})$  (see Eq. (3)) as

$$\frac{1}{|S|} E_{\sigma} \left[ \sup_{W \in \mathcal{W}} \sum_{i,j} \Sigma_{i,j} \ell(W_{i,j}, X_{i,j}) \right],$$

where  $\Sigma$  is a matrix defined as  $\Sigma_{i,j} = \sum_{\alpha: i_\alpha = i, j_\alpha = j} \sigma_\alpha$ . Using the Rademacher contraction principle (as in (Meir and Zhang, 2003, Lemma 5)), this is at most

$$\frac{l_\ell}{|S|} E_{\sigma} \left[ \sup_{W \in \mathcal{W}} \sum_{i,j} \Sigma_{i,j} W_{i,j} \right]. \quad (11)$$

Decomposing  $\Sigma = Y + Z$  as in Eq. (6) according to a parameter  $p$ , we can upper bound the Rademacher complexity by

$$\frac{l_\ell}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W \in \mathcal{W}} \sum_{i,j} Y_{i,j} W_{i,j} \right] + \frac{l_\ell}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W \in \mathcal{W}} \sum_{i,j} Z_{i,j} W_{i,j} \right]. \quad (12)$$

By definition of  $\mathcal{W}$ ,  $|W_{i,j}| \leq b_\phi$ , so the first term can be upper bounded by

$$\frac{l_\ell}{|S|} \mathbb{E}_{\sigma} \left[ b_\phi \sum_{i,j} |Y_{i,j}| \right] = \frac{l_\ell b_\phi}{|S|} \mathbb{E}_{\sigma} [\|Y\|_1]. \quad (13)$$

The second term in Eq. (12) equals

$$\frac{l_\ell}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W: \|W\|_{tr} \leq t} \sum_{i,j} Z_{i,j} \phi(W_{i,j}) \right] \leq \frac{l_\ell l_\phi}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W: \|W\|_{tr} \leq t} \sum_{i,j} Z_{i,j} W_{i,j} \right],$$

again by the Rademacher contraction principle. Applying Hölder's inequality, and using the fact that the spectral norm  $\|\cdot\|_{sp}$  is the dual to the trace norm  $\|\cdot\|_{tr}$ , we can upper bound the above by

$$\frac{l_\ell l_\phi}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W: \|W\|_{tr} \leq t} \|Z\|_{sp} \|W\|_{tr} \right] = \frac{l_\ell l_\phi t}{|S|} \mathbb{E}_{\sigma} [\|Z\|_{sp}].$$

Combining this with Eq. (13) and substituting into Eq. (12), we get an upper bound of the form

$$\frac{l_\ell b_\phi}{|S|} \mathbb{E}_{\sigma} [\|Y\|_1] + \frac{l_\ell l_\phi t}{|S|} \mathbb{E}_{\sigma} [\|Z\|_{sp}].$$

Using Lemma 9 and Lemma 10, we can upper bound this by

$$\frac{l_\ell b_\phi}{\sqrt{p}} + \frac{2.2Cl_\ell l_\phi t \sqrt{p}(\sqrt{m} + \sqrt{n})}{|S|},$$

where  $p$  is the parameter used to define  $Y$  and  $Z$  in Eq. (6). Choosing  $p = \frac{|S|b_\phi}{2.2Cl_\phi t(\sqrt{m} + \sqrt{n})}$ , we get the bound in the theorem.

## Appendix B. Proof of Thm. 4

Before we begin, we will need the following technical result:

**Lemma 11** *The dual of the norm  $\|W\| = \max\{\|W\|_{tr}/t, \|W\|_\infty/b\}$  equals*

$$\|\Sigma\|_* = \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp},$$

**Proof** By definition of a dual norm, we have

$$\|\Sigma\|_* = \sup_{W: \|W\| \leq 1} \langle \Sigma, W \rangle,$$

and our goal is to show that

$$\sup_{W: \|W\| \leq 1} \langle \Sigma, W \rangle = \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}.$$

First, we note that the dual norm of  $\|W\|_{tr}$  is well-known to be the spectral norm  $\|W\|_{sp}$  (namely, its largest singular value), and the dual of  $\|W\|_\infty$  is the 1-norm  $\|W\|_1 = \sum_{i,j} |W_{i,j}|$ . Now, for any  $Y, Z$  such that  $Y + Z = \Sigma$ , we have by Hölder's inequality that

$$\sup_{W: \|W\| \leq 1} \langle \Sigma, W \rangle = \sup_{W: \|W\| \leq 1} \langle Y, W \rangle + \langle Z, W \rangle \leq \sup_{W: \|W\| \leq 1} \|Y\|_1 \|W\|_\infty + \|Z\|_{sp} \|W\|_{tr} \leq b\|Y\|_1 + t\|Z\|_{sp}.$$

This holds for any  $Y, Z$ , and in particular

$$\sup_{W: \|W\| \leq 1} \langle \Sigma, W \rangle \leq \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}. \quad (14)$$

It remains to show the opposite direction, namely

$$\sup_{W: \|W\| \leq 1} \langle \Sigma, W \rangle \geq \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}.$$

To show this, let  $W^*$  be the matrix which maximizes the inner product above. We know that  $\|W^*\| \leq 1$ , which means that either  $\|W^*\|_\infty \leq b$ , or  $\|W^*\|_{tr} \leq t$ . If  $\|W^*\|_\infty \leq b$ , it follows that

$$\sup_{W: \|W\| \leq 1} \langle \Sigma, W \rangle = \sup_{W: \|W\|_\infty \leq b} \langle \Sigma, W \rangle = b\|\Sigma\|_1 \geq \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}.$$

In the other case, if  $\|W^*\|_{tr} \leq t$ , it follows that

$$\sup_{W: \|W\| \leq 1} \langle \Sigma, W \rangle = \sup_{W: \|W\|_{tr} \leq t} \langle \Sigma, W \rangle = t\|\Sigma\|_{sp} \geq \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}.$$

So in either case,

$$\sup_{W: \|W\| \leq 1} \langle \Sigma, W \rangle \geq \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}.$$

Combining this with Eq. (14), the result follows. ■

We now turn to the proof of Thm. 4 itself. Since  $\ell(W_{i,j}, X_{i,j})$  is assumed to be  $l_\ell$ -Lipschitz, we can use the Rademacher contraction principle to upper bound  $R_S(\ell \circ \mathcal{W})$  by

$$l_\ell \mathbb{E}_{\boldsymbol{\sigma}} \left[ \sup_{W \in \mathcal{W}} \frac{1}{|S|} \sum_{\alpha=1}^{|S|} \sigma_\alpha W_{i_\alpha, j_\alpha} \right] = \frac{l_\ell}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[ \sup_{W \in \mathcal{W}} \sum_{i,j} \Sigma_{i,j} W_{i,j} \right],$$

where  $\Sigma$  is a matrix defined as  $\Sigma_{i,j} = \sum_{\alpha: i_\alpha=i, j_\alpha=j} \sigma_\alpha$ .

Thinking of  $\Sigma, W$  as vectors, the equation above is the expected supremum of an inner product between  $\Sigma$  and  $W$ . By Hölder's inequality, we can upper bound this by

$$\frac{l_\ell}{|S|} \mathbb{E}_{\sigma} \left[ \sup_{W \in \mathcal{W}} \|\Sigma\|_* \|W\| \right] \quad (15)$$

for any norm  $\|\cdot\|$  and its dual norm  $\|\cdot\|_*$ . In particular, we will choose the norm  $\|W\| = \max\{\|W\|_{tr}/t, \|W\|_\infty/b\}$ . Note that by definition of  $W$ ,  $\sup_{W \in \mathcal{W}} \|W\| \leq 1$ . Also, by Lemma 11,

$$\|\Sigma\|_* = \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp},$$

where  $\|Y\|_1 = \sum_{i,j} |Y_{i,j}|$ , and  $\|Z\|_{sp}$  is the spectral norm of  $Z$ . Thus, we can upper bound Eq. (15) by

$$\frac{l_\ell}{|S|} \mathbb{E}_\Sigma \left[ \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp} \right]. \quad (16)$$

Recall that  $\Sigma$  is random matrix, where each entry is the sum of Rademacher variables. Let  $h_{i,j}$  denote the number of variables 'hitting' entry  $(i, j)$  - formally,  $h_{i,j} = |\{\alpha : (i_\alpha = i, j_\alpha = j)\}|$ . We can upper bound Eq. (16) by replacing the optimal decomposition of  $\Sigma$  into  $Y, Z$  by any fixed decomposition rule. In particular, for an arbitrary parameter  $p$ , we can decompose  $\Sigma$  into  $Y, Z$  as in Eq. (6), and get an upper bound on Eq. (16) of the form

$$\frac{l_\ell}{|S|} (b\mathbb{E}_\Sigma[\|Y\|_1] + t\mathbb{E}_\Sigma[\|Z\|_{sp}]). \quad (17)$$

Bounds for the two expectations are provided in Lemma 9 and Lemma 10. Plugging them in, we get

$$\frac{bl_\ell}{\sqrt{p}} + \frac{2.2l_\ell Ct\sqrt{p}(\sqrt{m} + \sqrt{n})}{|S|}.$$

Choosing  $p = \frac{b|S|}{2.2Ct(\sqrt{m} + \sqrt{n})}$  and simplifying, we get the bound in the theorem.

# Contextual Bandits with Similarity Information\*

Aleksandrs Slivkins

SLIVKINS@MICROSOFT.COM

*Microsoft Research Silicon Valley, Mountain View, CA 94043, USA.*

## Abstract

In a multi-armed bandit (MAB) problem, an online algorithm makes a sequence of choices. In each round it chooses from a time-invariant set of alternatives and receives the payoff associated with this alternative. While the case of small strategy sets is by now well-understood, a lot of recent work has focused on MAB problems with exponentially or infinitely large strategy sets, where one needs to assume extra structure in order to make the problem tractable. In particular, recent literature considered information on similarity between arms.

We consider similarity information in the setting of *contextual bandits*, a natural extension of the basic MAB problem where before each round an algorithm is given the *context* – a hint about the payoffs in this round. Contextual bandits are directly motivated by placing advertisements on webpages, one of the crucial problems in sponsored search. A particularly simple way to represent similarity information in the contextual bandit setting is via a *similarity distance* between the context-arm pairs which bounds from above the difference between the respective expected payoffs.

Prior work on contextual bandits with similarity uses “uniform” partitions of the similarity space, so that each context-arm pair is approximated by the closest pair in the partition. Algorithms based on “uniform” partitions disregard the structure of the payoffs and the context arrivals, which is potentially wasteful. We present algorithms that are based on *adaptive* partitions, and take advantage of “benign” payoffs and context arrivals without sacrificing the worst-case performance. The central idea is to maintain a finer partition in high-payoff regions of the similarity space and in popular regions of the context space. Our results apply to several other settings, e.g. MAB with constrained temporal change (Slivkins and Upfal, 2008) and sleeping bandits (Kleinberg et al., 2008a).

**ACM Categories and subject descriptors:** F.2.2 [Analysis of Algorithms and Problem Complexity]: Nonnumerical Algorithms and Problems; F.1.2 [Computation by Abstract Devices]: Modes of Computation—*Online computation* **General Terms:** Algorithms, Theory. **Keywords:** online learning, multi-armed bandits, contextual bandits, regret minimization, metric spaces.

---

\* The full version of this paper is available from [arxiv.org](http://arxiv.org). A preliminary version (which does not include the results in Section 6) was posted to [arxiv.org](http://arxiv.org) in July 2009. The results in Section 7 have been obtained while the author was a postdoc at Brown University, and a preliminary write-up has been circulated in 2007.

## 1. Introduction

In a multi-armed bandit problem (henceforth, “multi-armed bandit” will be abbreviated as MAB), an algorithm is presented with a sequence of trials. In each round, the algorithm chooses one alternative from a set of alternatives (*arms*) based on the past history, and receives the payoff associated with this alternative. The goal is to maximize the total payoff of the chosen arms. The MAB setting has been introduced in 1952 in Robbins (1952) and studied intensively since then in Operations Research, Economics and Computer Science. This setting is a clean model for the exploration-exploitation trade-off, a crucial issue in sequential decision-making under uncertainty.

One standard way to evaluate the performance of a bandit algorithm is *regret*, defined as the difference between the expected payoff of an optimal arm and that of the algorithm. By now the MAB problem with a small finite set of arms is quite well understood, e.g. see Lai and Robbins (1985); Auer et al. (2002b,a). However, if the arms set is exponentially or infinitely large, the problem becomes intractable unless we make further assumptions about the problem instance. Essentially, a bandit algorithm needs to find a needle in a haystack; for each algorithm there are inputs on which it performs as badly as random guessing.

Bandit problems with large sets of arms have been an active area of investigation in the past decade (see Section 2 for a discussion of related literature). A common theme in these works is to assume a certain *structure* on payoff functions. Assumptions of this type are natural in many applications, and often lead to efficient learning algorithms (Kleinberg, 2005). In particular, a line of work started in Agrawal (1995) assumes that some information on similarity between arms is available.

In this paper we consider similarity information in the setting of *contextual bandits* (Woodroffe, 1979; Auer, 2002; Wang et al., 2005; Pandey et al., 2007; Langford and Zhang, 2007), a natural extension of the basic MAB problem where before each round an algorithm is given the *context* – a hint about the payoffs in this round. Contextual bandits are directly motivated by the problem of placing advertisements on webpages, one of the crucial problems in sponsored search. One can cast it as a bandit problem so that arms correspond to the possible ads, and payoffs correspond to the user clicks. Then the context consists of information about the page, and perhaps the user this page is served to. Furthermore, we assume that similarity information is available on both the context and the arms. Following the work in Agrawal (1995); Kleinberg (2004); Auer et al. (2007); Kleinberg et al. (2008b) on the (non-contextual) bandits, a particularly simple way to represent similarity information in the contextual bandit setting is via a *similarity distance* between the context-arm pairs, which gives an upper bound on the difference between the corresponding payoffs.

### Our model: contextual bandits with similarity information.

The contextual bandits framework is defined as follows. Let  $X$  be the *context set* and  $Y$  be the *arms set*, and let  $\mathcal{P} \subset X \times Y$  be the set of feasible context-arms pairs. In each round  $t$ , the following events happen in succession:

1. a context  $x_t \in X$  is revealed to the algorithm,
2. the algorithm chooses an arm  $y_t \in Y$  such that  $(x_t, y_t) \in \mathcal{P}$ ,
3. payoff (reward)  $\pi_t \in [0, 1]$  is revealed.

The sequence of context arrivals  $(x_t)_{t \in \mathbb{N}}$  is fixed before the first round, and does not depend on the subsequent choices of the algorithm. With *stochastic payoffs*, for each pair  $(x, y) \in \mathcal{P}$

there is a distribution  $\Pi(x, y)$  with expectation  $\mu(x, y)$ , so that  $\pi_t$  is an independent sample from  $\Pi(x_t, y_t)$ . With *adversarial payoffs*, this distribution can change from round to round. For simplicity, we present the subsequent definitions for the stochastic setting only, whereas the adversarial setting is fleshed out later in the paper (Section 7).

In general, the goal of a bandit algorithm is to maximize the total payoff  $\sum_{t=1}^T \pi_t$ , where  $T$  is the *time horizon*. In the contextual MAB setting, we benchmark the algorithm's performance in terms of the context-specific "best arm". Specifically, the goal is to minimize the *contextual regret*:

$$R(T) \triangleq \sum_{t=1}^T \mu(x_t, y_t) - \mu^*(x_t), \quad \text{where } \mu^*(x) \triangleq \sup_{y \in Y: (x, y) \in \mathcal{P}} \mu(x, y).$$

The context-specific best arm is a more demanding benchmark than the best arm used in the "standard" (context-free) definition of regret.

The similarity information is given to an algorithm as a metric space  $(\mathcal{P}, \mathcal{D})$  which we call the *similarity space*, such that the following Lipschitz condition<sup>1</sup> holds:

$$|\mu(x, y) - \mu(x', y')| \leq \mathcal{D}((x, y), (x', y')). \quad (1)$$

Without loss of generality,  $\mathcal{D} \leq 1$ . The absence of similarity information is modeled as  $\mathcal{D} = 1$ .

An instructive special case is the *product similarity space*  $(\mathcal{P}, \mathcal{D}) = (X \times Y, \mathcal{D}_X + \mathcal{D}_Y)$ , where  $(X, \mathcal{D}_X)$  is a metric space on contexts (*context space*), and  $(Y, \mathcal{D}_Y)$  is a metric space on arms (*arms space*), and

$$\mathcal{D}((x, y), (x', y')) = \min(1, \mathcal{D}_X(x, x') + \mathcal{D}_Y(y, y')). \quad (2)$$

**Prior work: uniform partitions.** Hazan and Megiddo (2007) consider contextual MAB with similarity information on contexts. They suggest an algorithm that chooses a "uniform" partition  $S_X$  of the context space and approximates  $x_t$  by the closest point in  $S_X$ , call it  $x'_t$ . Specifically, the algorithm creates an instance  $\mathcal{A}(x)$  of some bandit algorithm  $\mathcal{A}$  for each point  $x \in S_X$ , and invokes  $\mathcal{A}(x'_t)$  in each round  $t$ . The granularity of the partition is adjusted to the time horizon, the context space, and the black-box regret guarantee for  $\mathcal{A}$ . Furthermore, Kleinberg (2004) provides a bandit algorithm  $\mathcal{A}$  for the adversarial MAB problem on a metric space that has a similar flavor: pick a "uniform" partition  $S_Y$  of the arms space, and run a  $k$ -arm bandit algorithm such as EXP3 Auer et al. (2002b) on the points in  $S_Y$ . Again, the granularity of the partition is adjusted to the time horizon, the arms space, and the black-box regret guarantee for EXP3.

Applying these two ideas to our setting (with the product similarity space) gives a simple algorithm which we call the *uniform algorithm*. Its contextual regret, even for adversarial payoffs, is

$$R(T) \leq O(T^{1-1/(2+d_X+d_Y)})(\log T), \quad (3)$$

where  $d_X$  is the covering dimension of the context space and  $d_Y$  is that of the arms space.

---

1. In other words,  $\mu$  is a Lipschitz-continuous function on  $(X, \mathcal{P})$ , with Lipschitz constant  $K_{\text{Lip}} = 1$ . Assuming  $K_{\text{Lip}} = 1$  is without loss of generality (as long as  $K_{\text{Lip}}$  is known to the algorithm), since we can re-define  $\mathcal{D} \leftarrow K_{\text{Lip}} \mathcal{D}$ .

**Our contributions.** Using “uniform” partitions disregards the potentially benign structure of expected payoffs and context arrivals. The central topic in this paper is *adaptive partitions* of the similarity space which are adjusted to frequently occurring contexts and high-paying arms, so that the algorithms can take advantage of the problem instances in which the expected payoffs or the context arrivals are “benign” (“low-dimensional”), in a sense that we make precise later.

We present two main results, one for stochastic payoffs and one for adversarial payoffs. For stochastic payoffs, we provide an algorithm called *contextual zooming* which “zooms in” on the regions of the context space that correspond to frequently occurring contexts, and the regions of the arms space that correspond to high-paying arms. Unlike the algorithms in prior work, this algorithm considers the context space and the arms space *jointly* – it maintains a partition of the similarity space, rather than one partition for contexts and another for arms. We develop provable guarantees that capture the “benign-ness” of the context arrivals and the expected payoffs. In the worst case, we match the guarantee (3) for the uniform algorithm. We obtain nearly matching lower bounds using the KL-divergence techniques from (Auer et al., 2002b; Kleinberg, 2004; Kleinberg et al., 2008b). The lower bound is very general as it holds for every given (product) similarity space *and* for every fixed value of the upper bound.

Our stochastic contextual MAB setting, and specifically the contextual zooming algorithm, can be fruitfully applied beyond the ad placement scenario described above and beyond MAB with similarity information per se. First, writing  $x_t = t$  one can incorporate “temporal constraints” (across time, for each arm), and combine them with “spatial constraints” (across arms, for each time). The analysis of contextual zooming yields concrete, meaningful bounds this scenario. In particular, we recover one of the main results in Slivkins and Upfal (2008). Second, our setting subsumes the stochastic *sleeping bandits* problem Kleinberg et al. (2008a), where in each round some arms are “asleep”, i.e. not available in this round. Here contexts correspond to subsets of arms that are “awake”. Contextual zooming recovers and generalizes the corresponding result in Kleinberg et al. (2008a). Third, following the publication of a preliminary version of this paper, contextual zooming has been applied to bandit learning-to-rank in Slivkins et al. (2010).

For the adversarial setting, we provide an algorithm which maintains an adaptive partition of the context space and thus takes advantage of “benign” context arrivals. We develop provable guarantees that capture this “benign-ness”. In the worst case, the contextual regret is bounded in terms of the covering dimension of the context space, matching (3). Our algorithm is in fact a *meta-algorithm*: given an adversarial bandit algorithm **Bandit**, we present a contextual bandit algorithm which calls **Bandit** as a subroutine. Our setup is flexible: depending on what additional constraints are known about the adversarial payoffs, one can plug in a bandit algorithm from the prior work on the corresponding version of adversarial MAB, so that the regret bound for **Bandit** plugs into the overall regret bound.

**Discussion.** Adaptive partitions (of the arms space) for context-free MAB with similarity information have been introduced in (Kleinberg et al., 2008b; Bubeck et al., 2008). This paper further explores the potential of the zooming technique in (Kleinberg et al., 2008b). Specifically, contextual zooming extends this technique to adaptive partitions of the entire similarity space, which necessitates a technically different algorithm and a more delicate analysis (see Discussion 1). We obtain a clean algorithm for contextual MAB with

improved (and nearly optimal) bounds. Moreover, this algorithm applies to several other, seemingly unrelated problems and unifies some results from prior work.

One alternative approach is to maintain a partition of the context space, and run a separate instance of the zooming algorithm from Kleinberg et al. (2008b) on each set in this partition. Fleshing out this idea leads to the meta-algorithm that we present for adversarial payoffs (with **Bandit** being the zooming algorithm). This meta-algorithm is parameterized (and constrained) by a specific a priori regret bound for **Bandit**. Unfortunately, any a priori regret bound for zooming algorithm would be a pessimistic one, which negates its main strength – the ability to adapt to “benign” expected payoffs.

**Map of the paper.** Section 2 is related work, and Section 3 is Preliminaries. Contextual zooming is presented in Section 4. Lower bounds are in Section 5. Some applications of contextual zooming are discussed in Section 6. The adversarial setting is treated in Section 7. All omitted proofs appear in the full version.

## 2. Related work

A proper discussion of the literature on bandit problems is beyond the scope of this paper. A reader is encouraged to refer to Cesa-Bianchi and Lugosi (2006) for background.

Most relevant to this paper is the work on bandits with large sets of arms, specifically bandits with similarity information (Agrawal, 1995; Kleinberg, 2004; Auer et al., 2007; Pandey et al., 2007; Kočsík and Szepesvári, 2006; Munos and Coquelin, 2007; Kleinberg et al., 2008b; Bubeck et al., 2008; Kleinberg and Slivkins, 2010; Maillard and Munos, 2010). Another commonly assumed structure is linear or convex payoffs, e.g. (Awerbuch and Kleinberg, 2008; Flaxman et al., 2005; Dani et al., 2007; Abernethy et al., 2008; Hazan and Kale, 2009). Linear/convex payoffs is a much stronger assumption than similarity, essentially because it allows to make strong inferences about far-away arms. Other assumptions have been considered, e.g. (Wang et al., 2008; Bubeck and Munos, 2010). The distinction between stochastic and adversarial payoffs is orthogonal to the structural assumption (such as Lipschitz-continuity or linearity). Papers on MAB with linear/convex payoffs typically allow adversarial payoffs, whereas papers on MAB with similarity information focus on stochastic payoffs, with notable exceptions of Kleinberg (2004) and Maillard and Munos (2010).<sup>2</sup>

The notion of structured adversarial payoffs in this paper is less restrictive than the one in Maillard and Munos (2010) (which in turn specializes the notion from linear/convex payoffs), in the sense that the Lipschitz condition is assumed on the expected payoffs rather than on realized payoffs. This is a non-trivial distinction, essentially because our notion generalizes stochastic payoffs whereas the other one does not. In particular, Maillard and Munos (2010) achieve regret  $\tilde{O}(\sqrt{dT})$  for  $d$ -dimensional real space, whereas (even) for stochastic payoffs there is a lower bound  $\Omega(T^{1-1/(d+2)})$  (Kleinberg, 2004; Bubeck et al., 2008).

**Contextual MAB.** In (Auer, 2002) and (Chu et al., 2011)<sup>2</sup> payoffs are linear in context, which is a feature vector. (Woodroffe, 1979; Wang et al., 2005) and (Rigollet and Zeevi, 2010)<sup>2</sup> study contextual MAB with stochastic payoffs, under the name *bandits with*

---

2. This paper is concurrent and independent work w.r.t. the preliminary publication of this paper on [arxiv.org](http://arxiv.org).

*covariates*: the context is a random variable correlated with the payoffs; they consider the case of two arms, and make some additional assumptions. Lazaric and Munos (2009)<sup>2</sup> consider an online labeling problem with stochastic inputs and adversarially chosen labels; inputs and hypotheses (mappings from inputs to labels) can be thought of as “contexts” and “arms” respectively. All these papers are not directly applicable to the present setting.

Experimental work on contextual MAB includes (Pandey et al., 2007) and (Li et al., 2010, 2011).<sup>2</sup>

Lu et al. (2010)<sup>2</sup> consider the setting in this paper for a product similarity space and, essentially, recover the uniform algorithm and a lower bound that matches (3). The same guarantee (3) can also be obtained as follows. The “uniform partition” described above can be used to define “experts” for a bandit-with-expert-advice algorithm such as EXP4 (Auer et al., 2002b): for each set of the partition there is an expert whose advise is simply an arbitrary arm in this set. Then the regret bound for EXP4 yields (3). Instead of EXP4 one could use an algorithm in McMahan and Streeter (2009)<sup>2</sup> which improves over EXP4 if the experts are not “too distinct”; however, it is not clear if it translates into concrete improvements over (3).

If the context  $x_t$  is time-invariant, our setting reduces to the Lipschitz MAB problem as defined in (Kleinberg et al., 2008b), which in turn reduces to continuum-armed bandits (Agrawal, 1995; Kleinberg, 2004; Auer et al., 2007) if the metric space is a real line, and to MAB with stochastic payoffs (Auer et al., 2002a) if the similarity information is absent.

### 3. Preliminaries

We will use the notation from Introduction. In particular,  $x_t$  will denote the  $t$ -th *context arrival*, i.e. the context that arrives in round  $t$ , and  $y_t$  will denote the arm chosen by the algorithm in that round. We will use  $x_{(1..T)}$  to denote the sequence of the first  $T$  context arrivals ( $x_1, \dots, x_T$ ). The *badness* of a point  $(x, y) \in \mathcal{P}$  is defined as  $\Delta(x, y) \triangleq \mu^*(x) - \mu(x, y)$ . The context-specific best arm is

$$y^*(x) \in \operatorname{argmax}_{y \in Y: (x, y) \in \mathcal{P}} \mu(x, y), \quad (4)$$

where ties are broken in an arbitrary but fixed way. To ensure that the max in (4) is attained by some  $y \in Y$ , we will assume that the similarity space  $(\mathcal{P}, \mathcal{D})$  is compact.

**Metric spaces.** Covering dimension and related notions are crucial throughout this paper. Let  $\mathcal{P}$  be a set of points in a metric space, and fix  $r > 0$ . An *r-covering* of  $\mathcal{P}$  is a collection of subsets of  $\mathcal{P}$ , each of diameter strictly less than  $r$ , that cover  $\mathcal{P}$ . The minimal number of subsets in an *r-covering* is called the *r-covering number* of  $\mathcal{P}$  and denoted  $N_r(\mathcal{P})$ .<sup>3</sup> The *covering dimension* of  $\mathcal{P}$  (with multiplier  $c$ ) is the smallest  $d$  such that  $N_r(\mathcal{P}) \leq cr^{-d}$  for each  $r > 0$ . In particular, if  $S$  is a subset of Euclidean space then its covering dimension is at most the linear dimension of  $S$ , but can be (much) smaller.

Covering is closely related to packing. A subset  $S \subset \mathcal{P}$  is an *r-packing* of  $\mathcal{P}$  if the distance between any two points in  $S$  is at least  $r$ . The maximal number of points in an *r-packing* is

---

3. The covering number can be defined via radius- $r$  balls rather than diameter- $r$  sets. This alternative definition lacks the appealing “robustness” property:  $N_r(\mathcal{P}') \leq N_r(\mathcal{P})$  for any  $\mathcal{P}' \subset \mathcal{P}$ , but (other than that) is equivalent for this paper.

called the *r-packing number* and denoted  $N_r^{\text{pack}}(\mathcal{P})$ . It is well-known that *r*-packing numbers are essentially the same as *r*-covering numbers, namely  $N_{2r}(\mathcal{P}) \leq N_r^{\text{pack}}(\mathcal{P}) \leq N_r(\mathcal{P})$ .

The *doubling constant*  $c_{\text{DBL}}(\mathcal{P})$  of  $\mathcal{P}$  is the smallest  $k$  such that any ball can be covered by  $k$  balls of half the radius. The doubling constant has been a standard notion in theoretical computer science since Gupta et al. (2003). It is known that  $c_{\text{DBL}}(\mathcal{P}) \geq c2^d$  if  $d$  is the covering dimension of  $\mathcal{P}$  with multiplier  $c$ , and that  $c_{\text{DBL}}(\mathcal{P}) \leq 2^d$  if  $\mathcal{P}$  is a subset of  $d$ -dimensional Euclidean space. A useful observation is that if distance between any two points in  $S$  is  $> r$ , then any ball of radius  $r$  contains at most  $c_{\text{DBL}}^2$  points of  $S$ .

A ball with center  $x$  and radius  $r$  is denoted  $B(x, r)$ . Formally, we will treat a ball as a (center, radius) pair rather than a set of points. A function  $f : \mathcal{P} \rightarrow \mathbb{R}$  if a Lipschitz function on a metric space  $(\mathcal{P}, \mathcal{D})$ , with Lipschitz constant  $K_{\text{Lip}}$ , if the *Lipschitz condition* holds:  $|f(x) - f(x')| \leq K_{\text{Lip}} \mathcal{D}(x, x')$  for each  $x, x' \in \mathcal{P}$ .

**Accessing the similarity space.** We assume full and computationally unrestricted access to the similarity information. While the issues of efficient representation thereof are important in practice, we believe that a proper treatment of these issues would be specific to the particular application and the particular similarity metric used, and would obscure the present paper. One clean formal way to address this issue is to assume *oracle access*: an algorithm accesses the similarity space via a few specific types of queries, and invokes an “oracle” that answers such queries.

**Time horizon.** We assume that the time horizon is fixed and known in advance. This assumption is without loss of generality in our setting. This is due to the well-known *doubling trick* which converts a bandit algorithm with a fixed time horizon into one that runs indefinitely and achieves essentially the same regret bound. Suppose for any fixed time horizon  $T$  there is an algorithm  $\text{ALG}_T$  whose regret is at most  $R(T)$ . The new algorithm proceeds in phases  $i = 1, 2, 3, \dots$  of duration  $2^i$  rounds each, so that in each phase  $i$  a fresh instance of  $\text{ALG}_{2^i}$  is run. This algorithm has regret  $O(\log T)R(T)$  for each round  $T$ , and  $O(R(T))$  in the typical case when  $R(T) \geq T^\gamma$  for some constant  $\gamma > 0$ .

#### 4. The contextual zooming algorithm

In this section we consider the contextual MAB problem with stochastic payoffs. We present an algorithm for this problem, called *contextual zooming*, which takes advantage of both the “benign” context arrivals and the “benign” expected payoffs. The algorithm adaptively maintains a partition of the similarity space, “zooming in” on both the “popular” regions on the context space and the high-payoff regions of the arms space.

**Discussion 1** *Contextual zooming extends the (context-free) zooming technique in (Kleinberg et al., 2008b), which necessitates a somewhat more complicated algorithm. In particular, selection and activation rules are defined differently, there is a new notion of “domains” and the distinction between “pre-index” and “index”. The analysis is more delicate, both the high-probability argument in Claim 4 and the subsequent argument that bounds the number of samples from suboptimal arms. Also, the key step of setting up the regret bounds is very different, esp. in Section 4.4.*

#### 4.1. Provable guarantees

Let us define the notions that express the performance of contextual zooming. These notions rely on the packing number  $N_r(\cdot)$  in the similarity space  $(\mathcal{P}, \mathcal{D})$ , and the more refined versions thereof that take into account “benign” expected payoffs and “benign” context arrivals.

Our guarantees have the following form, for some integer numbers  $\{N_r\}_{r \in (0,1)}$ :

$$R(T) \leq C_0 \inf_{r_0 \in (0,1)} \left( r_0 T + \sum_{r=2^{-i}: i \in \mathbb{N}, r_0 \leq r \leq 1} \frac{1}{r} N_r \log T \right). \quad (5)$$

Here and thereafter,  $C_0 = O(1)$  unless specified otherwise. In the pessimistic version,  $N_r = N_r(\mathcal{P})$  is the  $r$ -packing number of  $\mathcal{P}$ .<sup>4</sup> The main contribution is refined bounds in which  $N_r$  is smaller.

For every guarantee of the form (5), call it  $N_r$ -type guarantee, prior work (e.g., Kleinberg (2004); Kleinberg et al. (2008b); Bubeck et al. (2008)) suggests a more tractable dimension-type guarantee. This guarantee is in terms of the *covering-type dimension* induced by  $N_r$ , defined as follows:<sup>5</sup>

$$d_c \triangleq \inf\{d > 0 : N_r \leq c r^{-d} \quad \forall r \in (0, 1)\}. \quad (6)$$

Using (5) with  $r_0 = T^{-1/(d_c+2)}$ , we obtain

$$R(T) \leq O(C_0) (c T^{1-1/(2+d_c)} \log T) \quad (\forall c > 0). \quad (7)$$

For the pessimistic version ( $N_r = N_r(\mathcal{P})$ ), the corresponding covering-type dimension  $d_c$  is the covering dimension of the similarity space. The resulting guarantee (7) subsumes the bound (3) from prior work (because the covering dimension of a product similarity space is  $d_X + d_Y$ ), and extends this bound from product similarity spaces (2) to arbitrary similarity spaces.

To account for “benign” expected payoffs, instead of  $r$ -packing number of the entire set  $\mathcal{P}$  we consider the  $r$ -packing number of a subset of  $\mathcal{P}$  which only includes points with near-optimal expected payoffs:

$$\mathcal{P}_{\mu,r} \triangleq \{(x, y) \in \mathcal{P} : \mu^*(x) - \mu(x, y) \leq 12r\}. \quad (8)$$

We define the  $r$ -zooming number as  $N_r(\mathcal{P}_{\mu,r})$ , the  $r$ -packing number of  $\mathcal{P}_{\mu,r}$ . The corresponding covering-type dimension (6) is called the *contextual zooming dimension*.

The  $r$ -zooming number can be seen as an optimistic version of  $N_r(\mathcal{P})$ : while equal to  $N_r(\mathcal{P})$  in the worst case, it can be much smaller if the set of near-optimal context-arm pairs is “small” in terms of the packing number. Likewise, the contextual zooming dimension is an optimistic version of the covering dimension.

- 
- 4. Then (5) can be simplified to  $R(T) \leq \inf_{r \in (0,1)} O(rT + \frac{1}{r} N_r(\mathcal{P}) \log T)$  since  $N_r(\mathcal{P})$  is non-increasing in  $r$ .
  - 5. One standard definition of the covering dimension is (6) for  $N_r = N_r(\mathcal{P})$  and  $c = 1$ . Following Kleinberg et al. (2008b), we include an explicit dependence on  $c$  in (6) to obtain a more efficient regret bound (which holds for any  $c$ ).

**Theorem 2** Consider the contextual MAB problem with stochastic payoffs. The contextual regret  $R(T)$  of the contextual zooming algorithm satisfies (5), where  $N_r = N_r(\mathcal{P}_{\mu,r})$  is the  $r$ -zooming number. Consequently,  $R(T)$  satisfies the dimension-type guarantee (7), where  $d_c$  is the contextual zooming dimension.

In Theorem 2, the same algorithm enjoys the bound (7) for each  $c > 0$ . This is a useful trade-off since different values of  $c$  may result in drastically different values of the dimension  $d_c$ . On the contrary, the “uniform algorithm” from prior work essentially needs to take the  $c$  as input.

Further refinements to take into account “benign” context arrivals are deferred to Section 4.4.

#### 4.2. Description of the algorithm

The algorithm is parameterized by the time horizon  $T$ . In each round  $t$ , it maintains a finite collection  $\mathcal{A}_t$  of balls in  $(\mathcal{P}, \mathcal{D})$  (called *active balls*) which collectively cover the similarity space. Adding active balls is called *activating*; balls stay active once they are activated. Initially there is only one active ball which has radius 1 and therefore contains the entire similarity space.

On a high level, each round  $t$  proceeds as follows. Context  $x_t$  arrives. Then the algorithm selects an active ball  $B$  and an arm  $y_t$  such that  $(x_t, y_t) \in B$ , according to the “selection rule”. Arm  $y_t$  is played. Then one ball may be activated, according to the “activation rule”.

Let us define the two rules. First we need several definitions. Fix an active ball  $B$  and round  $t$ . Let  $r(B)$  be the radius of  $B$ . The *confidence radius* of  $B$  at time  $t$  is

$$\text{rad}_t(B) \triangleq \text{rad}(n_t(B)) \triangleq 4 \sqrt{\frac{\log T}{1+n_t(B)}}, \quad (9)$$

where  $n_t(B)$  is the number of times  $B$  has been selected by the algorithm before round  $t$ . The *domain* of ball  $B$  in round  $t$ , denoted  $\text{dom}(B, \mathcal{A}_t)$ , is a subset of  $B$  that excludes all balls  $B' \in \mathcal{A}_t$  of strictly smaller radius:

$$\text{dom}(B, \mathcal{A}_t) \triangleq B \setminus \left( \bigcup_{B' \in \mathcal{A}_t: r(B') < r(B)} B' \right). \quad (10)$$

$B$  is called *relevant* in round  $t$  if  $(x_t, y) \in \text{dom}(B, \mathcal{A}_t)$  for some arm  $y$ . In each round, the algorithm chooses among relevant balls  $B$  according to a numerical score  $I_t(B)$  called *index*. (The definition of index is deferred to the end of this subsection.) Now we are ready to state the two rules:

- **selection rule.** In round  $t$ , select a relevant ball  $B$  with the maximal index (break ties arbitrarily). Select an arbitrary arm  $y$  such that  $(x_t, y) \in \text{dom}(B, \mathcal{A}_t)$ .
- **activation rule.** If in round  $t$  the selection rule selects  $(B, y)$  such that  $\text{rad}(n_t(B) + 1) \leq r(B)$ , then a ball with center  $(x_t, y)$  and radius  $\frac{1}{2}r(B)$  is activated. ( $B$  is then called the *parent* of this ball.)

See Algorithm 1 for the pseudocode.

It remains to define the index  $I_t(B)$ . Let  $\text{rew}_t(B)$  be the total payoff from all rounds up to  $t - 1$  in which ball  $B$  has been selected by the algorithm. Then the average payoff

**Algorithm 1** Contextual zooming algorithm.

---

1: **Input:** Similarity space  $(\mathcal{P}, \mathcal{D})$  of diameter  $\leq 1$ ,  $\mathcal{P} \subset X \times Y$ . Time horizon  $T$ .  
 2: **Data:** collection  $\mathcal{A}$  of “active balls” in  $(\mathcal{P}, \mathcal{D})$ ; counters  $n(B)$ ,  $\text{rew}(B)$  for each  $B \in \mathcal{A}$ .

3: **Init:**  $B \leftarrow B(p, 1)$ ;  $\mathcal{A} \leftarrow \{B\}$ ;  $n(B) = \text{rew}(B) = 0$  //  $p \in \mathcal{P}$  is arbitrary  
 4: **Main loop:** for each round  $t$  // use definitions (9-12)  
 5: Input context  $x_t$ .  
 6:  $\text{relevant} \leftarrow \{B \in \mathcal{A} : (x_t, y) \in \text{dom}(B, \mathcal{A}) \text{ for some arm } y\}$ .  
 7:  $B \leftarrow \underset{B \in \text{relevant}}{\text{argmax}} I_t(B)$ . // ball  $B$  is selected  
 8:  $y \leftarrow$  any arm  $y$  such that  $(x_t, y) \in \text{dom}(B, \mathcal{A})$ .  
 9: **if**  $\text{rad}(n_t(B) + 1) \leq \text{radius}(B)$  **then**  
 10:    $B' \leftarrow B((x_t, y), \frac{1}{2} \text{radius}(B))$  // new ball to be activated  
 11:    $\mathcal{A} \leftarrow \mathcal{A} \cup \{B'\}$ ;  $n(B') = \text{rew}(B') = 0$ .  
 12: Play arm  $y$ , observe payoff  $\pi$ .  
 13: Update counters:  $n(B) \leftarrow n(B) + 1$ ,  $\text{rew}(B) \leftarrow \text{rew}(B) + \pi$ .

---

from  $B$  is  $\nu_t(B) \triangleq \frac{\text{rew}_t(B)}{\max(1, n_t(B))}$ . The *pre-index* of  $B$  is defined as the average  $\nu_t(B)$  plus an “uncertainty term”:

$$I_t^{\text{pre}}(B) \triangleq \nu_t(B) + 2r(B) + \text{rad}_t(B). \quad (11)$$

The “uncertainty term” in (11) reflects both uncertainty due to a location in the metric space and uncertainty due to an insufficient number of samples. The index of  $B$  is obtained by taking a minimum over all active balls  $B'$  of radius at least  $r(B)$  (letting  $\mathcal{D}(B, B')$  is the distance between the centers of the two balls).

$$I_t(B) \triangleq \min_{B' \in \mathcal{A}_t: r(B') \geq r(B)} I_t^{\text{pre}}(B') + \mathcal{D}(B, B'). \quad (12)$$

### 4.3. Analysis of the algorithm: proof of Theorem 2

We start by observing that the activation rule ensures several important invariants.

**Claim 3** *The following invariants are maintained:*

- (centering) if  $B$  is activated in round  $t$  with parent  $B^{\text{par}}$ , then the center of  $B$  is  $(x_t, y_t) \in \text{dom}(B^{\text{par}}, \mathcal{A})$ .
- (confidence)  $\text{rad}_t(B) > r(B)$  for all active balls  $B$  and all rounds  $t$ .
- (covering) in each round  $t$ , the domains of active balls cover the similarity space.
- (separation) for any two active balls of radius  $r$ , their centers are at distance  $\geq r$ .

**Proof** The first two invariants are immediate. For the covering invariant, note that  $\cup_{B \in \mathcal{A}} \text{dom}(B, \mathcal{A}) = \cup_{B \in \mathcal{A}} B$  for any finite collection  $\mathcal{A}$  of balls in the similarity space. (For each  $v \in \cup_{B \in \mathcal{A}} B$ , consider a smallest radius ball in  $\mathcal{A}$  that contains  $B$ . Then  $v \in \text{dom}(B, \mathcal{A})$ .) The covering invariant then follows since  $\mathcal{A}_t$  contains a ball that covers the entire similarity space.

To show the separation invariant, let  $B$  and  $B'$  be two balls of radius  $r$  such that  $B$  is activated at time  $t$ , with parent  $B^{\text{par}}$ , and  $B'$  is activated before time  $t$ . The center of  $B$

is some point  $(x_t, y_t) \in \text{dom}(B^{\text{par}}, \mathcal{A}_t)$ . Since  $r(B^{\text{par}}) > r(B')$ , it follows that  $(x_t, y_t) \notin B'$ . ■

Throughout the analysis we will use the following notation. For a ball  $B$  with center  $(x, y) \in \mathcal{P}$ , define the expected payoff of  $B$  as  $\mu(B) \triangleq \mu(x, y)$ . Let  $B_t^{\text{sel}}$  be the active ball selected by the algorithm in round  $t$ . Recall that the *badness* of  $(x, y) \in \mathcal{P}$  is defined as  $\Delta(x, y) \triangleq \mu^*(x) - \mu(x, y)$ .

**Claim 4** *If ball  $B$  is active in round  $t$ , then with probability at least  $1 - T^{-2}$  we have that*

$$|\nu_t(B) - \mu(B)| \leq r(B) + \text{rad}_t(B). \quad (13)$$

**Proof** Fix ball  $V$  with center  $(x, y)$ . Let  $S$  be the set of rounds  $s \leq t$  when ball  $B$  was selected by the algorithm, and let  $n = |S|$  be the number of such rounds. Then  $\nu_t(B) = \frac{1}{n} \sum_{s \in S} \pi_s(x_s, y_s)$ .

Define  $Z_k = \sum (\pi_s(x_s, y_s) - \mu(x_s, y_s))$ , where the sum is taken over the  $k$  smallest elements  $s \in S$ . Then  $\{Z_{k \wedge n}\}_{k \in \mathbb{N}}$  is a martingale with bounded increments. (Note that  $n$  here is a random variable.) So by the Azuma-Hoeffding inequality with probability at least  $1 - T^{-3}$  it holds that  $\frac{1}{k} |Z_{k \wedge n}| \leq \text{rad}_t(B)$ , for each  $k \leq T$ . Taking the Union Bound, it follows that  $\frac{1}{n} |Z_n| \leq \text{rad}_t(B)$ . Note that  $|\mu(x_s, y_s) - \mu(B)| \leq r(B)$  for each  $s \in S$ , so  $|\nu_t(B) - \mu(B)| \leq r(B) + \frac{1}{n} |Z_n|$ , which completes the proof. ■

Call a run of the algorithm *clean* if (13) holds for each round. From now on we will focus on a clean run, and argue deterministically using (13). The heart of the analysis is the following lemma.

**Lemma 5** *Consider a clean run of the algorithm. Then  $\Delta(x_t, y_t) \leq 15 r(B_t^{\text{sel}})$  in each round  $t$ .*

**Proof** Fix round  $t$ . By the covering invariant,  $(x_t, y^*(x_t)) \in B$  for some active ball  $B$ . Recall from (12) that  $I_t(B) = I^{\text{pre}}(B') + \mathcal{D}(B, B')$  for some active ball  $B'$  of radius  $r(B') \geq r(B)$ . Therefore

$$\begin{aligned} I_t(B_t^{\text{sel}}) &\geq I_t(B) = I^{\text{pre}}(B') + \mathcal{D}(B, B') && \text{(selection rule, defn of index (12))} \\ &= \nu_t(B') + 2r(B') + \text{rad}_t(B') + \mathcal{D}(B, B') && \text{(defn of preindex (11))} \\ &\geq \mu(B') + r(B) + \mathcal{D}(B, B') && \text{(Claim 4 and } r(B') \geq r(B)\text{)} \\ &\geq \mu(B) + r(B) \geq \mu(x_t, y^*(x_t)) = \mu^*(x_t). && \text{(Lipschitz property (1), twice)} \end{aligned} \quad (14)$$

On the other hand, letting  $B^{\text{par}}$  be the parent of  $B_t^{\text{sel}}$  and noting that by the selection rule

$$\text{rad}_t(B^{\text{par}}) \leq r(B^{\text{par}}) = 2r(B_t^{\text{sel}}), \quad (15)$$

we can upper-bound  $I_t(B_t^{\text{sel}})$  as follows:

$$\begin{aligned}
I_t(B_t^{\text{sel}}) &\leq I^{\text{pre}}(B^{\text{par}}) + r(B^{\text{par}}) && (\text{defn of index (12)}) \\
&= \nu_t(B^{\text{par}}) + 3r(B^{\text{par}}) + \text{rad}_t(B^{\text{par}}) && (\text{defn of preindex (11)}) \\
&\leq \mu(B^{\text{par}}) + 4r(B^{\text{par}}) + 2\text{rad}_t(B^{\text{par}}) && (\text{Claim 4}) \\
&\leq \mu(B^{\text{par}}) + 12r(B_t^{\text{sel}}) && ("parenthood" (15)) \\
&\leq \mu(x_t, y_t) + 15r(B_t^{\text{sel}}) && (\text{Lipschitz property (1)}). \quad (16)
\end{aligned}$$

In the last inequality we used the fact that  $(x_t, y_t)$  is within distance  $3r(B_t^{\text{sel}})$  from the center of  $B^{\text{par}}$ . Putting the pieces together,  $\mu^*(x_t) \leq I_t(B_t^{\text{sel}}) \leq \mu(x_t, y_t) + 15r(B_t^{\text{sel}})$ . ■

**Corollary 6** *In a clean run, if ball  $B$  is activated in round  $t$  then  $\Delta(x_t, y_t) \leq 12r(B)$ .*

**Proof** By the activation rule,  $B_t^{\text{sel}}$  is the parent of  $B$ . Thus by Lemma 5 we immediately have  $\Delta(x_t, y_t) \leq 15r(B_t^{\text{sel}}) = 30r(B)$ . To obtain the constant of 12 that is claimed here, it suffices to prove a more efficient special case of Lemma 5: if  $\text{rad}_t(B_t^{\text{sel}}) \leq r(B_t^{\text{sel}})$  then  $\Delta(x_t, y_t) \leq 6r(B_t^{\text{sel}})$ . To prove this, we simply replace (16) in the proof of Lemma 5 by similar inequality in terms of  $I^{\text{pre}}(B_t^{\text{sel}})$  rather than  $I^{\text{pre}}(B^{\text{par}})$ :

$$\begin{aligned}
I_t(B_t^{\text{sel}}) &\leq I^{\text{pre}}(B_t^{\text{sel}}) = \nu_t(B_t^{\text{sel}}) + 2r(B_t^{\text{sel}}) + \text{rad}_t(B_t^{\text{sel}}) && (\text{defns (11-12)}) \\
&\leq \mu(B_t^{\text{sel}}) + 3r(B_t^{\text{sel}}) + 2\text{rad}_t(B_t^{\text{sel}}) && (\text{Claim 4}) \\
&\leq \mu(x_t, y_t) + 6r(B_t^{\text{sel}})
\end{aligned}$$

■

Now we are ready for the final regret computation. For a given  $r = 2^{-i}$ ,  $i \in \mathbb{N}$ , let  $\mathcal{F}_r$  be the collection of all balls of radius  $r$  that have been activated throughout the execution of the algorithm. A ball  $B \in \mathcal{F}_r$  is called *full* in round  $t$  if  $\text{rad}_t(B) \leq r$ . Note that in each round, if a full ball is selected then some other ball is activated. Thus, we will partition the rounds among active balls as follows: for each ball  $B \in \mathcal{F}_r$ , let  $S_B$  be the set of rounds which consists of the round when  $B$  was activated and all rounds  $t$  when  $B$  was selected and not full. It is easy to see that  $|S_B| \leq O(r^{-2} \log T)$ . Moreover, by Lemma 5 and Corollary 6 we have  $\Delta(x_t, y_t) \leq 15r$  in each round  $t \in S_B$ .

If ball  $B \in \mathcal{F}_r$  is activated in round  $t$ , then by the activation rule its center is  $(x_t, y_t)$ , and Corollary 6 asserts that  $(x_t, y_t) \in \mathcal{P}_{\mu, r}$ , as defined in (8). By the separation invariant, the centers of balls in  $\mathcal{F}_r$  are within distance at least  $r$  from one another. It follows that  $|\mathcal{F}_r| \leq N_r$ , where  $N_r$  is the  $r$ -zooming number.

Fixing some  $r_0 \in (0, 1)$ , note that in each rounds  $t$  when a ball of radius  $< r_0$  was selected, regret is  $\Delta(x_t, y_t) \leq O(r_0)$ , so the total regret from all such rounds is at most

$O(r_0 T)$ . Therefore, contextual regret can be written as follows:

$$\begin{aligned} R(T) &= \sum_{t=1}^T \Delta(x_t, y_t) \\ &= O(r_0 T) + \sum_{r=2^{-i}: r_0 \leq r \leq 1} \sum_{B \in \mathcal{F}_r} \sum_{t \in S_B} \Delta(x_t, y_t) \\ &\leq O(r_0 T) + \sum_{r=2^{-i}: r_0 \leq r \leq 1} \sum_{B \in \mathcal{F}_r} |S_B| O(r) \\ &\leq O\left(r_0 T + \sum_{r=2^{-i}: r_0 \leq r \leq 1} \frac{1}{r} N_r \log(T)\right). \end{aligned}$$

The  $N_r$ -type regret guarantee in Theorem 2 follows by taking inf on all  $r_0 \in (0, 1)$ .

#### 4.4. Improved regret bounds

Let us provide regret bounds that take into account “benign” context arrivals. The main difficulty here is to develop the corresponding definitions; the analysis then carries over without much modification. The added value is two-fold: first, we establish the intuition that benign context arrivals matter, and then the specific regret bound is used in Section 6.2 to match the result in Slivkins and Upfal (2008).

A crucial step in the proof of Theorem 2 is to bound the number of active radius- $r$  balls by  $N_r(\mathcal{P}_{\mu,r})$ , which is accomplished by observing that their centers form an  $r$ -packing  $S$  of  $\mathcal{P}_{\mu,r}$ . We make this step more efficient, as follows. An active radius- $r$  ball is called *full* if  $\text{rad}_t(B) \leq r$  for some round  $t$ . Note that each active ball is either full or a child of some other ball that is full. The number of children of a given ball is bounded by the doubling constant of the similarity space. Thus, it suffices to consider the number of active radius- $r$  balls that are full, which is at most  $N_r(\mathcal{P}_{\mu,r})$ , and potentially much smaller.

Consider active radius- $r$  active balls that are full. Their centers form an  $r$ -packing  $S$  of  $\mathcal{P}_{\mu,r}$  with an additional property: each context arrival  $x_t$  can be assigned to exactly one point  $p \in S$  so that  $(x_t, y) \in B(p, r)$  for some arm  $y$ , and each point in  $S$  is assigned at least  $1/r^2$  arrivals. A set  $S \subset \mathcal{P}$  with this property is called  *$r$ -consistent* (with context arrivals). The *adjusted  $r$ -packing number* of a set  $\mathcal{P}' \subset \mathcal{P}$ , denoted  $N_r^{\text{adj}}(\mathcal{P}')$ , is the maximal size of an  $r$ -consistent  $r$ -packing of  $\mathcal{P}'$ . It can be much smaller than the  $r$ -packing number of  $\mathcal{P}'$  if most context arrivals fall into a small region of the similarity space.

We make one further optimization. A point  $(x, y) \in \mathcal{P}$  is called an  *$r$ -winner* if for each  $(x', y') \in B((x, y), 2r)$  it holds that  $\mu(x', y') = \mu^*(x')$ . Let  $\mathcal{W}_{\mu,r}$  be the set of all  $r$ -winners. It is easy to see that if  $B$  is a radius- $r$  ball centered at an  $r$ -winner, and  $B$  or its child is selected in a given round, then this round does not contribute to contextual regret. Therefore, it suffices to consider ( $r$ -consistent)  $r$ -packings of  $\mathcal{P}_{\mu,r} \setminus \mathcal{W}_{\mu,r}$ . This can be a significant saving if for most context arrivals  $x_t$  expected payoff  $\mu(x_t, y)$  is either optimal or very suboptimal (see Section 6.2 for an example).

Our final guarantee is in terms of  $N_r^{\text{adj}}(\mathcal{P}_{\mu,r} \setminus \mathcal{W}_{\mu,r})$ , which we term the *adjusted  $r$ -zooming number*.

**Theorem 7** Consider the contextual MAB problem with stochastic payoffs. The contextual regret  $R(T)$  of the contextual zooming algorithm satisfies (5), where  $N_r$  is the adjusted  $r$ -zooming number and  $C_0 = O(c_{\text{DBL}})$ . Here  $c_{\text{DBL}}$  is the doubling constant of the similarity space. Consequently,  $R(T)$  satisfies the dimension-type guarantee (7), where  $d_c$  is the corresponding covering-type dimension.

## 5. Lower bounds

We match the upper bound in Theorem 2 up to  $O(\log T)$  factors. Our lower bound is very general: it applies to an arbitrary product similarity space, and moreover for a given similarity space it matches, up to  $O(\log T)$  factors, any fixed value of the upper bound (as explained below).

We construct a distribution over problem instances on a given metric space, so that the lower bound is for a problem instance drawn from this distribution. A single problem instance would not suffice to establish a lower bound because a trivial algorithm that picks arm  $y^*(x)$  for each context  $x$  will achieve regret 0.

To formulate our result, let  $R_\mu^{\text{UB}}(T)$  denote the upper bound in Theorem 2, i.e. is the right-hand side of (5) where  $N_r = N_r(\mathcal{P}_{\mu,r})$  is the  $r$ -zooming number. Let  $R^{\text{UB}}(T)$  denote the pessimistic version of this bound, namely right-hand side of (5) where  $N_r = N_r(\mathcal{P})$  is the packing number of  $\mathcal{P}$ .

**Theorem 8** *Consider the contextual MAB problem with stochastic payoffs, Let  $(\mathcal{P}, \mathcal{D})$  be a product similarity space. Fix an arbitrary time horizon  $T$  and a positive number  $R \leq R^{\text{UB}}(T)$ . Then there exists a distribution  $\mathcal{I}$  over problem instances on  $(\mathcal{P}, \mathcal{D})$  with the following two properties:*

- (a)  $R_\mu^{\text{UB}}(T) \leq O(R)$  for each problem instance in  $\text{support}(\mathcal{I})$ .
- (b) for any contextual bandit algorithm it holds that  $\mathbb{E}_{\mathcal{I}}[R(T)] \geq \Omega(R/\log T)$ ,

To prove this theorem, we build on the lower-bounding technique from Auer et al. (2002b), and its extension to (context-free) bandits in metric spaces in Kleinberg (2004). In particular, we use the basic *needle-in-the-haystack* example from Auer et al. (2002b), where the “haystack” consists of several arms with expected payoff  $\frac{1}{2}$ , and the “needle” is an arm whose expected payoff is slightly higher. Roughly, for suitably chosen parameter  $r \in (0, 1)$  such that  $N = Tr^2 \leq N_r(\mathcal{P})$  we pick an  $r$ -net  $S_X$  in the context space and an  $r$ -net  $S_Y$  in the arms space so that  $|S_X| \times |S_Y| = N$ . For each  $x \in S_X$  we construct a needle-in-the-haystack example on the set  $S_Y$ : we pick some  $y^*(x) \in S_Y$  to be the “needle” (independently and uniformly at random), and define  $\mu(x, y^*(x)) = \frac{1}{2} + \frac{r}{2}$ , and  $\mu(x, y) = \frac{1}{2} + \frac{r}{4}$  for other  $y \in S_Y$ . We smoothen the expected payoffs so that far from  $S_X \times S_Y$  expected payoffs are  $\frac{1}{2}$  and the Lipschitz condition holds. The sequence  $x_{(1..T)}$  of context arrivals is defined in an arbitrary round-robin fashion over the points in  $S_X$ . We show that in  $T$  rounds each context in  $S_X$  contributes  $\Omega(|S_Y|/r)$  to contextual regret resulting in total contextual regret  $\Omega(N/r)$ , and  $R_\mu^{\text{UB}}(T) \leq O(N/r)(\log T)$  for each problem instance in our construction. See the full version for details.

## 6. Applications of contextual zooming

We describe several applications of contextual zooming: to MAB with slow adversarial change (Section 6.1), to MAB with stochastically evolving payoffs (Section 6.2), and to the “sleeping bandits” problem (Section 6.3). In particular, we recover some of the main results in Slivkins and Upfal (2008) and Kleinberg et al. (2008a). Also, in Section 6.3 we discuss a recent application of contextual zooming to bandit learning-to-rank, which has been published in Slivkins et al. (2010). Most of the proofs are deferred to the full version.

### 6.1. MAB with slow adversarial change

Consider the (context-free) adversarial MAB problem in which expected payoffs of each arm change over time *gradually*. Specifically, we assume that expected payoff of each arm  $y$  changes by at most  $\sigma_y$  in each round, for some a-priori known *volatilities*  $\sigma_y$ . The algorithm's goal here is to adapt to the changing environment. Thus, we define *dynamic regret*: regret with respect to a benchmark which in each round plays the best arm for this round. We are primarily interested in the long-term performance quantified by average dynamic regret  $\hat{R}(T) \triangleq R(T)/T$ . We call this setting the ***drifting MAB problem***.

We restate this setting as a contextual MAB problem with stochastic payoffs in which the  $t$ -th context arrival is simply  $x_t = t$ . Then  $\mu(t, y)$  is the expected payoff of arm  $y$  at time  $t$ , and dynamic regret coincides with contextual regret specialized to the case  $x_t = t$ . Each arm  $y$  satisfies a “temporal constraint”:

$$|\mu(t, y) - \mu(t', y)| \leq \sigma_y |t - t'| \quad (17)$$

for some constant  $\sigma_y$ . To set up the corresponding similarity space  $(\mathcal{P}, \mathcal{D})$ , let  $\mathcal{P} = [T] \times Y$ , and

$$\mathcal{D}((t, y), (t', y')) = \min(1, \sigma_y |t - t'| + \mathbf{1}_{\{y \neq y'\}}). \quad (18)$$

Our solution for the drifting MAB problem is the contextual zooming algorithm parameterized by the similarity space  $(\mathcal{P}, \mathcal{D})$ . To obtain guarantees for the long-term performance, we run contextual zooming with a suitably chosen time horizon  $T_0$ , and restart it every  $T_0$  rounds; we call this version *contextual zooming with period  $T_0$* . The general guarantees are provided by Theorem 2 and Theorem 7. Below we work out some specific, tractable corollaries.

**Corollary 9** Consider the drifting MAB problem with  $k$  arms and volatilities  $\sigma_y \equiv \sigma$ . Contextual zooming with period  $T_0$  has average dynamic regret  $\hat{R}(T) = O(k\sigma \log T_0)^{1/3}$ , whenever  $T \geq T_0 \geq (\frac{k}{\sigma^2})^{1/3} \log \frac{k}{\sigma}$ .

**Proof** Since  $\hat{R}(T) \leq 2\hat{R}(T_0)$  for any  $T \geq T_0$ , it suffices to bound  $\hat{R}(T_0)$ . Therefore, from here on we can focus on analyzing contextual zooming itself (rather than contextual zooming with period).

The main step is to derive the regret bound (5) with a specific upper bound on  $N_r$ . We will show that

$$\text{dynamic regret } R(\cdot) \text{ satisfies (5) with } N_r \leq k \lceil \frac{T\sigma}{r} \rceil. \quad (19)$$

Plugging  $N_r \leq k(1 + \frac{T\sigma}{r})$  into (5) and taking  $r_0 = (k\sigma \log T)^{1/3}$  we obtain<sup>6</sup>

$$R(T) \leq O(T)(k\sigma \log T)^{1/3} + O(\frac{k^2}{\sigma})^{1/3}(\log T) \quad \forall T \geq 1.$$

Therefore, for any  $T \geq (\frac{k}{\sigma^2})^{1/3} \log \frac{k}{\sigma}$  we have  $\hat{R}(T) = O(k\sigma \log T)^{1/3}$ .

---

6. This choice of  $r_0$  minimizes the inf expression in (5) up to constant factors by equating the two summands.

It remains to prove (19). We use a pessimistic version of Theorem 2: (5) with  $N_r = N_r(\mathcal{P})$ , the  $r$ -packing number of  $\mathcal{P}$ . Fix  $r \in (0, 1]$ . For any  $r$ -packing  $S$  of  $\mathcal{P}$  and each arm  $y$ , each time interval  $I$  of duration  $\Delta_r \triangleq r/\sigma$  provides at most one point for  $S$ : there exists at most one time  $t \in I$  such that  $(t, y) \in S$ . Since there are at most  $\lceil T/\Delta_r \rceil$  such intervals  $I$ , it follows that  $N_r(\mathcal{P}) \leq k \lceil T/\Delta_r \rceil \leq k(1 + T\frac{\sigma}{r})$ . ■

The restriction  $\sigma_y \equiv \sigma$  is non-essential: it is not hard to obtain the same bound with  $\sigma = \frac{1}{k} \sum_y \sigma_y$ . Modifying the construction in Section 5 (details omitted from this version) one can show that Corollary 9 is optimal up to  $O(\log T)$  factors.

**Drifting MAB with spatial constraints.** The temporal version ( $x_t = t$ ) of our contextual MAB setting with stochastic payoffs subsumes the drifting MAB problem and furthermore allows to combine the temporal constraints (17) described above (for each arm, across time) with “spatial constraints” (for each time, across arms). To the best of our knowledge, such MAB models are quite rare in the literature.<sup>7</sup> A clean example is

$$\mathcal{D}((t, y), (t', y')) = \min(1, \sigma |t - t'| + \mathcal{D}_Y(y, y')), \quad (20)$$

where  $(Y, \mathcal{D}_Y)$  is the arms space. For this example, we can obtain an analog of Corollary 9, where the regret bound depends on the covering dimension of the arms space  $(Y, \mathcal{D}_Y)$ .

## 6.2. Bandits with stochastically evolving payoffs

We consider a special case of drifting MAB problem in which expected payoffs of each arm evolve over time according to a stochastic process with a uniform stationary distribution. We obtain improved regret bounds for contextual zooming, taking advantage of the full power of our analysis in Section 4.

In particular, we address a version in which the stochastic process is a random walk with step  $\pm\sigma$ . This version has been previously studied in Slivkins and Upfal (2008) under the name “Dynamic MAB”. For the main case ( $\sigma_i \equiv \sigma$ ), our regret bound for Dynamic MAB matches that in Slivkins and Upfal (2008).

**Uniform marginals.** First we address the general version that we call *drifting MAB with uniform marginals*. Formally, we assume that expected payoffs  $\mu(y, \cdot)$  of each arm  $y$  evolve over time according to some stochastic process  $\Gamma_y$  that satisfies (17). We assume that the processes  $\Gamma_y$ ,  $y \in Y$  are mutually independent, and moreover that the marginal distributions  $\mu(y, t)$  are uniform on  $[0, 1]$ , for each time  $t$  and each arm  $y$ .<sup>8</sup> We are interested in  $\mathbb{E}_{\Gamma}[\hat{R}(T)]$ , average dynamic regret in expectation over the processes  $\Gamma_y$ .

We obtain a stronger version of (19) via Theorem 7. To use this theorem, we need to bound the adjusted  $r$ -zooming number, call it  $N_r$ . We show that

$$\mathbb{E}_{\Gamma}[N_r] = O(kr)\lceil \frac{T\sigma}{r} \rceil \text{ and } \left(r < \sigma^{1/3} \Rightarrow N_r = 0\right). \quad (21)$$

Then we obtain a different bound on dynamic regret, which is stronger than Corollary 9 for  $k < \sigma^{-1/2}$ .

---

7. The only other MAB model with this flavor that we are aware of, found in Hazan and Kale (2009), combines linear payoffs and bounded “total variation” (aggregate temporal change) of the cost functions.

8. E.g. this assumption is satisfied by any Markov Chain on  $[0, 1]$  with stationary initial distribution.

**Corollary 10** Consider drifting MAB with uniform marginals, with  $k$  arms and volatilities  $\sigma_y \equiv \sigma$ . Contextual zooming with period  $T_0$  satisfies  $\mathbb{E}_\Gamma[\hat{R}(T)] = O(k \sigma^{2/3} \log T_0)$ , whenever  $T \geq T_0 \geq \sigma^{-2/3} \log \frac{1}{\sigma}$ .

The crux of the proof is to show (21). Interestingly, it involves using all three optimizations in Theorem 7:  $N_r(\mathcal{P}_{\mu,r})$ ,  $N_r(\mathcal{P}_{\mu,r} \setminus \mathcal{W}_{\mu,r})$  and  $N_r^{\text{adj}}(\cdot)$ , whereas any two of them do not seem to suffice. The rest is a straightforward computation similar to the one in Corollary 9.

**Dynamic MAB.** Let us consider the Dynamic MAB problem from Slivkins and Upfal (2008). Here for each arm  $y$  the stochastic process  $\Gamma_y$  is a random walk with step  $\pm \sigma_y$ . To ensure that the random walk stays within the interval  $[0, 1]$ , we assume reflecting boundaries. Formally, we assume that  $1/\sigma_y \in \mathbb{N}$ , and once a boundary is reached, the next step is deterministically in the opposite direction.<sup>9</sup>

According to a well-known fact about random walks that

$$\Pr \left[ |\mu(t, y) - \mu(t', y)| \leq O(\sigma_y |t - t'|^{1/2} \log T_0) \right] \geq 1 - T_0^{-3} \quad \text{if } |t - t'| \leq T_0. \quad (22)$$

We use contextual zooming with period  $T_0$ , but we parameterize it by a different similarity space  $(\mathcal{P}, \mathcal{D}_{T_0})$  that we define according to (22). Namely, we set

$$\mathcal{D}_{T_0}((t, y), (t', y')) = \min(1, \sigma_y |t - t'|^{1/2} \log T_0 + \mathbf{1}_{\{y \neq y'\}}). \quad (23)$$

The following corollary is proved using the same technique as Corollary 10:

**Corollary 11** Consider the Dynamic MAB problem with  $k$  arms and volatilities  $\sigma_y \equiv \sigma$ . Let  $\text{ALG}_{T_0}$  denote the contextual zooming algorithm with period  $T_0$  which is parameterized by the similarity space  $(\mathcal{P}, \mathcal{D}_{T_0})$ . Then  $\text{ALG}_{T_0}$  satisfies  $\mathbb{E}_\Gamma[\hat{R}(T)] = O(k \sigma \log^2 T_0)$ , whenever  $T \geq T_0 \geq \frac{1}{\sigma} \log \frac{1}{\sigma}$ .

### 6.3. Other applications

**Sleeping bandits.** The *sleeping bandits* problem Kleinberg et al. (2008a) is an extension of MAB where in each round some arms can be “asleep”, i.e. not available in this round. One of the main results in Kleinberg et al. (2008a) is on sleeping bandits with stochastic payoffs. We recover this result using contextual zooming.

We model sleeping bandits as contextual MAB problem where each context arrival  $x_t$  corresponds to the set of arms that are “awake” in this round. More precisely, for every subset  $S \subset Y$  of arms there is a distinct context  $x_S$ , and  $\mathcal{P} = \{(x_S, y) : y \in S \subset Y\}$  is the set of feasible context-arm pairs. The similarity distance is simply  $\mathcal{D}((x, y), (x', y')) = \mathbf{1}_{\{y \neq y'\}}$ . Note that the Lipschitz condition (1) is satisfied.

For this setting, contextual zooming essentially reduces to the “highest awake index” algorithm in Kleinberg et al. (2008a). In fact, we can re-derive the result Kleinberg et al. (2008a) on sleeping MAB with stochastic payoffs as an easy corollary of Theorem 2.

Moreover, the contextual MAB problem extends the sleeping bandits setting by incorporating similarity information on arms. The contextual zooming algorithm (and its analysis) applies, and is geared to exploit this additional similarity information.

---

9. Slivkins and Upfal (2008) has a slightly more general setup which does not require  $1/\sigma_y \in \mathbb{N}$ .

**Bandit learning-to-rank.** Following a preliminary publication of this paper on [arxiv.org](https://arxiv.org), contextual zooming has been applied in Slivkins et al. (2010) to bandit learning-to-rank. Interestingly, the “contexts” studied in Slivkins et al. (2010) are very different from what we considered so far.

The basic setting, motivated by web search, was introduced in Radlinski et al. (2008). In each round a new user arrives. The algorithm selects a ranked list of  $k$  documents and presents it to the user who clicks on at most one document, namely on the first document that (s)he finds relevant. A user is specified by a binary vector over documents. The goal is to minimize *abandonment*: the number of rounds with no clicks.

Slivkins et al. (2010) study an extension in which metric similarity information is available. They consider a version with *stochastic payoffs*: in each round, the user vector is an independent sample from a fixed distribution, and assume a Lipschitz-style condition that connects expected clicks with the metric space. They run a separate bandit algorithm (e.g., contextual zooming) for each of the  $k$  “slots” in the ranking. Without loss of generality, in each round the documents are selected sequentially, in the top-down order. Since a document in slot  $i$  is clicked in a given round only if all higher ranked documents are not relevant, they treat the set of documents in the higher slots as a *context* for the  $i$ -th algorithm. The Lipschitz-style condition on expected clicks suffices to guarantee the corresponding Lipschitz-style condition on contexts.

## 7. Contextual bandits with adversarial payoffs

In this section we consider the adversarial setting. We provide an algorithm which maintains an adaptive partition of the context space and thus takes advantage of “benign” context arrivals. It is in fact a *meta-algorithm*: given a bandit algorithm *Bandit*, we present a contextual bandit algorithm, called *ContextualBandit*, which calls *Bandit* as a subroutine.

**Our setting.** Recall that in each round  $t$ , the context  $x_t \in X$  is revealed, then the algorithm picks an arm  $y_t \in Y$  and observes the payoff  $\pi_t \in [0, 1]$ . Here  $X$  is the context set, and  $Y$  is the arms set. In this section, all context-arms pairs are feasible:  $\mathcal{P} = X \times Y$ .

Adversarial payoffs are defined as follows. For each round  $t$ , there is a payoff function  $\hat{\pi}_t : X \times Y \rightarrow [0, 1]$  such that  $\pi_t = \hat{\pi}_t(x_t, y_t)$ . The payoff function  $\hat{\pi}_t$  is sampled independently from a time-specific distribution  $\Pi_t$  over payoff functions. Distributions  $\Pi_t$  are fixed by the adversary in advance, before the first round, and not revealed to the algorithm. Denote  $\mu_t(x, y) \triangleq \mathbb{E}[\Pi_t(x, y)]$ .

Following Hazan and Megiddo (2007), we generalize the notion of regret for context-free adversarial MAB to contextual MAB. The context-specific best arm is

$$y^*(x) \in \operatorname{argmax}_{y \in Y} \sum_{t=1}^T \mu_t(x, y), \quad (24)$$

where the ties are broken in an arbitrary but fixed way. We define *adversarial contextual regret* as

$$R(T) \triangleq \sum_{t=1}^T \mu_t(x_t, y_t) - \mu_t^*(x_t), \quad \text{where } \mu_t^*(x) \triangleq \mu_t(x, y^*(x)). \quad (25)$$

Similarity information is given to an algorithm as a pair of metric spaces: a metric space  $(X, \mathcal{D}_X)$  on contexts (the *context space*) and a metric space  $(Y, \mathcal{D}_Y)$  on arms (the

*arms space*), which form the product similarity space  $(X \times Y, \mathcal{D}_X + \mathcal{D}_Y)$ . We assume that for each round  $t$  functions  $\mu_t$  and  $\mu_t^*$  are Lipschitz on  $(X \times Y, \mathcal{D}_X + \mathcal{D}_Y)$  and  $(X, \mathcal{D}_X)$ , respectively, both with Lipschitz constant 1 (see Footnote 1). We assume that the context space is compact, in order to ensure that the max in (24) is attained by some  $y \in Y$ . Without loss of generality,  $\text{diameter}(X, \mathcal{D}_X) \leq 1$ .

Formally, a problem instance consists of metric spaces  $(X, \mathcal{D}_X)$  and  $(Y, \mathcal{D}_Y)$ , the sequence of context arrivals (denoted  $x_{(1..T)}$ ), and a sequence of distributions  $(\Pi_t)_{t \leq T}$ . Note that for a fixed distribution  $\Pi_t = \Pi$ , this setting reduces to the stochastic setting, as defined in Introduction. For the fixed context case ( $x_t = x$  for all  $t$ ) this setting reduces to the (context-free) MAB problem with a randomized oblivious adversary.

**Our results.** Our algorithm is parameterized by a regret guarantee for **Bandit** for the fixed context case, namely an upper bound on the convergence time.<sup>10</sup> For a more concrete theorem statement we will assume that the convergence time of **Bandit** is at most  $T_0(r) \triangleq c_Y r^{-(2+d_Y)} \log(\frac{1}{r})$  for some constants  $c_Y$  and  $d_Y$  that are known to the algorithm. In particular, an algorithm in Kleinberg (2004) achieves this guarantee if  $d_Y$  is the  $c$ -covering dimension of the arms space and  $c_Y = O(c^{2+d_Y})$ .

This is a flexible formulation that can leverage prior work on adversarial bandits. For instance, if  $Y \subset \mathbb{R}^d$  and for each fixed context  $x \in X$  distributions  $\Pi_t$  randomize over linear functions  $\hat{\pi}_t(x, \cdot) : Y \rightarrow \mathbb{R}$ , then one could take **Bandit** from the line of work on adversarial bandits with linear payoffs. In particular, there exist algorithms with  $d_Y = 0$  and  $c_Y = \text{poly}(d)$  (Dani et al., 2007; Abernethy et al., 2008). Likewise, for convex payoffs there exist algorithms with  $d_Y = 2$  and  $c_Y = O(d)$  (Flaxman et al., 2005). For a bounded number of arms, algorithm EXP3 (Auer et al., 2002b) achieves  $d_Y = 0$  and  $c_Y = O(\sqrt{|Y|})$ .

From here on, the context space  $(X, \mathcal{D}_X)$  will be only metric space considered; balls and other notions will refer to the context space only.

To quantify the “goodness” of context arrivals, our guarantees are in terms of the covering dimension of  $x_{(1..T)}$  rather than that of the entire context space. (This is the improvement over the guarantee (3) for the uniform algorithm.) In fact, in the full version we use a more refined notion which allows to disregard a limited number of “outliers” in  $x_{(1..T)}$ . Our result is stated as follows:

**Theorem 12** *Consider the contextual MAB problem with adversarial payoffs, and let **Bandit** be a bandit algorithm. Assume that the problem instance belongs to some class of problem instances such that for the fixed-context case, convergence time of **Bandit** is at most  $T_0(r) \triangleq c_Y r^{-(2+d_Y)} \log(\frac{1}{r})$  for some constants  $c_Y$  and  $d_Y$  that are known to the algorithm. Then **ContextualBandit** achieves adversarial contextual regret  $R(\cdot)$  such that for any time  $T$  and any constant  $c_X > 0$  it holds that*

$$R(T) \leq O(c_{\text{DBL}}^2 (c_X c_Y)^{1/(2+d_X+d_Y)} T^{1-1/(2+d_X+d_Y)} (\log T)), \quad (26)$$

where  $d_X$  is the covering dimension of  $x_{(1..T)}$  with multiplier  $c_X$ , and  $c_{\text{DBL}}$  is the doubling constant of  $x_{(1..T)}$ .

---

10. The  $r$ -convergence time  $T_0(r)$  is the smallest  $T_0$  such that regret is  $R(T) \leq rT$  for each  $T \geq T_0$ .

**Our algorithm.** The contextual bandit algorithm **ContextualBandit** is parameterized by a (context-free) bandit algorithm **Bandit**, which it uses as a subroutine, and a function  $T_0(\cdot) : (0, 1) \rightarrow \mathbb{N}$ .

The algorithm maintains a finite collection  $\mathcal{A}$  of balls, called *active balls*. Initially there is one active ball of radius 1. Ball  $B$  stays active once it is *activated*. Then a fresh instance  $\text{ALG}_B$  of **Bandit** is created, whose set of “arms” is  $Y$ .  $\text{ALG}_B$  can be parameterized by the time horizon  $T_0(r)$ , where  $r$  is the radius of  $B$ .

The algorithm proceeds as follows. In each round  $t$  the algorithm selects one active ball  $B \in \mathcal{A}$  such that  $x_t \in B$ , calls  $\text{ALG}_B$  to select an arm  $y \in Y$  to be played, and reports the payoff  $\pi_t$  back to  $\text{ALG}_B$ . A given ball can be selected at most  $T_0(r)$  times, after which it is called *full*.  $B$  is called *relevant* in round  $t$  if it contains  $x_t$  and is not full. The algorithm selects a relevant ball (breaking ties arbitrarily) if such ball exists. Otherwise, a new ball  $B'$  is activated and selected. Specifically, let  $B$  be the smallest-radius active ball containing  $x_t$ . Then  $B' = B(x_t, \frac{r}{2})$ , where  $r$  is the radius of  $B$ .  $B$  is then called the *parent* of  $B'$ . See Algorithm 2 for the pseudocode.

---

**Algorithm 2** Algorithm **ContextualBandit**.

---

```

1: Input:
2:   Context space  $(X, \mathcal{D}_X)$  of diameter  $\leq 1$ , set  $Y$  of arms.
3:   Bandit algorithm Bandit and a function  $T_0(\cdot) : (0, 1) \rightarrow \mathbb{N}$ .
4: Data structures:
5:   A collection  $\mathcal{A}$  of “active balls” in  $(X, \mathcal{D}_X)$ .
6:    $\forall B \in \mathcal{A}$ : counter  $n_B$ , instance  $\text{ALG}_B$  of Bandit on arms  $Y$ .
7: Initialization:
8:    $B \leftarrow B(x, 1)$ ;  $\mathcal{A} \leftarrow \{B\}$ ;  $n_B \leftarrow 0$ ; initiate  $\text{ALG}_B$ .           // center  $x \in X$  is
   arbitrary
9:    $\mathcal{A}^* \leftarrow \mathcal{A}$            // active balls that are not full
10: Main loop: for each round  $t$ 
11:   Input context  $x_t$ .
12:    $\text{relevant} \leftarrow \{B \in \mathcal{A}^* : x_t \in B\}$ .
13:   if  $\text{relevant} \neq \emptyset$  then
14:      $B \leftarrow$  any  $B \in \text{relevant}$ .
15:   else           // activate a new ball:
16:      $r \leftarrow \min_{B \in \mathcal{A}: x_t \in B} r_B$ .
17:      $B \leftarrow B(x_t, r/2)$ .           // new ball to be added
18:      $\mathcal{A} \leftarrow \mathcal{A} \cup \{B\}$ ;  $\mathcal{A}^* \leftarrow \mathcal{A}^* \cup \{B\}$ ;  $n_B \leftarrow 0$ ; initiate  $\text{ALG}_B$ .
19:      $y \leftarrow$  next arm selected by  $\text{ALG}_B$ .
20:     Play arm  $y$ , observe payoff  $\pi$ , report  $\pi$  to  $\text{ALG}_B$ .
21:      $n_B \leftarrow n_B + 1$ .
22:   if  $n_B = T_0(\text{radius}(B))$  then  $\mathcal{A}^* \leftarrow \mathcal{A}^* \setminus \{B\}$ .           // ball  $B$  is full

```

---

The analysis of this algorithm (which proves Theorem 12) is deferred to the full version.

**Acknowledgements.** The author is grateful to Ittai Abraham, Bobby Kleinberg and Eli Upfal for many conversations about multi-armed bandits, and to Sébastien Bubeck for

help with the manuscript. Also, comments from anonymous COLT reviewers have been tremendously useful.

## References

- Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the Dark: An Efficient Algorithm for Bandit Linear Optimization. In *21th COLT*, pages 263–274, 2008.
- Rajeev Agrawal. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33(6):1926–1951, 1995.
- Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. of Machine Learning Research (JMLR)*, 3:397–422, 2002. Preliminary version in *41st IEEE FOCS*, 2000.
- Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a. Preliminary version in *15th ICML*, 1998.
- Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b. Preliminary version in *36th IEEE FOCS*, 1995.
- Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved Rates for the Stochastic Continuum-Armed Bandit Problem. In *20th COLT*, pages 454–468, 2007.
- Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *J. of Computer and System Sciences*, 74(1):97–114, February 2008. Preliminary version in *36th ACM STOC*, 2004.
- Sébastien Bubeck and Rémi Munos. Open Loop Optimistic Planning. In *23rd COLT*, pages 477–489, 2010.
- Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvari. Online Optimization in X-Armed Bandits. In *NIPS*, pages 201–208, 2008.
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge Univ. Press, 2006.
- Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual Bandits with Linear Payoff Functions. In *14th*, 2011.
- Varsha Dani, Thomas P. Hayes, and Sham Kakade. The Price of Bandit Information for Online Optimization. In *20th NIPS*, 2007.
- Abraham Flaxman, Adam Kalai, and H. Brendan McMahan. Online Convex Optimization in the Bandit Setting: Gradient Descent without a Gradient. In *16th ACM-SIAM SODA*, pages 385–394, 2005.
- Anupam Gupta, Robert Krauthgamer, and James R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *44th IEEE FOCS*, pages 534–543, 2003.
- Elad Hazan and Satyen Kale. Better algorithms for benign bandits. In *20th ACM-SIAM SODA*, pages 38–47, 2009.
- Elad Hazan and Nimrod Megiddo. Online Learning with Prior Information. In *20th COLT*, pages 499–513, 2007.

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *18th NIPS*, 2004.

Robert Kleinberg. *Online Decision Problems with Large Strategy Sets*. PhD thesis, MIT, 2005.

Robert Kleinberg and Aleksandrs Slivkins. Sharp Dichotomies for Regret Minimization in Metric Spaces. In *21st ACM-SIAM SODA*, 2010.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. In *21st COLT*, pages 425–436, 2008a.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-Armed Bandits in Metric Spaces. In *40th ACM STOC*, pages 681–690, 2008b.

Levente Kocsis and Csaba Szepesvari. Bandit Based Monte-Carlo Planning. In *17th ECML*, pages 282–293, 2006.

T.L. Lai and Herbert Robbins. Asymptotically efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

John Langford and Tong Zhang. The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits. In *21st NIPS*, 2007.

Alessandro Lazaric and Rémi Munos. Hybrid Stochastic-Adversarial On-line Learning. In *22nd COLT*, 2009.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *19th*, 2010.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *4th*, 2011.

Tyler Lu, Dávid Pál, and Martin Pál. Showing Relevant Ads via Lipschitz Context Multi-Armed Bandits. In *14th*, 2010.

Odalric-Ambrym Maillard and Rémi Munos. Online Learning in Adversarial Lipschitz Environments. In *ECML PKDD*, pages 305–320, 2010.

H. Brendan McMahan and Matthew Streeter. Tighter Bounds for Multi-Armed Bandits with Expert Advice. In *22nd COLT*, 2009.

Rémi Munos and Pierre-Arnaud Coquelin. Bandit algorithms for tree search. In *23rd UAI*, 2007.

Sandeep Pandey, Deepak Agarwal, Deepayan Chakrabarti, and Vanja Josifovski. Bandits for Taxonomies: A Model-based Approach. 2007.

Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *25th ICML*, pages 784–791, 2008.

Philippe Rigollet and Assaf Zeevi. Nonparametric Bandits with Covariates. In *23rd COLT*, pages 54–66, 2010.

Herbert Robbins. Some Aspects of the Sequential Design of Experiments. *Bull. Amer. Math. Soc.*, 58:527–535, 1952.

Aleksandrs Slivkins and Eli Upfal. Adapting to a Changing Environment: the Brownian Restless Bandits. In *21st COLT*, pages 343–354, 2008.

CONTEXTUAL BANDITS WITH SIMILARITY INFORMATION

Aleksandrs Slivkins, Filip Radlinski, and Sreenivas Gollapudi. Learning optimally diverse rankings over large document collections. In *27th ICML*, pages 983–990, 2010.

Chih-Chun Wang, Sanjeev R. Kulkarni, and H. Vincent Poor. Bandit problems with side observations. *IEEE Trans. on Automatic Control*, 50(3):338355, 2005.

Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for Infinitely Many-Armed Bandits. In *NIPS*, pages 1729–1736, 2008.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *J. Amer. Statist. Assoc.*, 74(368), 1979.

SLIVKINS

# Adaptive Density Level Set Clustering

**Ingo Steinwart**

INGO.STEINWART@MATHEMATIK.UNI-STUTTGART.DE

*Institute for Stochastics and Applications*

*Faculty 8: Mathematics and Physics*

*University of Stuttgart*

*D-70569 Stuttgart Germany*

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

Clusters are often defined to be the connected components of a density level set. Unfortunately, this definition depends on a level that needs to be user specified by some means. In this paper we present a simple algorithm that is able to asymptotically determine the optimal level, that is, the level at which there is the first split in the cluster tree of the data generating distribution. We further show that this algorithm asymptotically recovers the corresponding connected components. Unlike previous work, our analysis does not require strong assumptions on the density such as continuity or even smoothness.

**Keywords:** Clustering, Density Level Sets

## 1. Introduction

A central and widely studied task in statistical learning theory or machine learning is cluster analysis, where the goal is to find clusters in unlabeled data. Unlike in supervised learning tasks such as classification or regression, a key problem in cluster analysis is already the definition of a learning goal that describes a conceptionally and mathematically convincing definition of clusters. A widely, but by no means generally accepted, definition of clusters has its roots in a paper by Carmichael et al. (1968), who define clusters to be densely populated areas in the input space that are separated by less populated areas. The *non-parametric* mathematical translation of this idea, which goes back to Hartigan (1975), usually assumes that the data  $D = (x_1, \dots, x_n) \in X^n$  is generated by some unknown probability measure  $P$  on a topological space  $X$  that has a density  $h$  with respect to some known reference measure  $\mu$  on  $X$ . Given a threshold  $\rho \geq 0$ , the clusters are then defined to be the connected components of the density level set  $\{h \geq \rho\} := \{x \in X : h(x) \geq \rho\}$ . Here, one typically considers the case, where  $X \subset \mathbb{R}^d$  and  $\mu$  is the Lebesgue measure on  $X$ . In addition, it is often assumed that the density  $h$  is continuous, since this removes or hides various pathologies regarding the topological notion of connectedness that are caused by changes of  $h$  on  $\mu$ -zero sets.

Historically, two distinct questions have been investigated for this cluster definition. The first one is the so-called single level approach, which tries to estimate the connected components of  $\{h \geq \rho\}$  for a *single and fixed* level  $\rho \geq 0$ . The single level approach has been studied by several authors, see, e.g., Hartigan (1975); Cuevas and Fraiman (1997); Rigollet (2007); Maier et al. (2009); Rinaldo and Wasserman (2010) and the references therein, and

hence it seems fair to say that it already enjoys a reasonably good statistical understanding. Unfortunately, however, it suffers from a conceptional problem, namely that of determining a good value of  $\rho$ , and recently Rinaldo and Wasserman (2010) actually remark that research in this direction “would be very useful”.

The second approach tries to address this issue by considering the hierarchical structure of the connected components for different levels. To be more precise, if  $h$  is a fixed density, which, for the sake of simplicity, is assumed to have *closed* density level sets, and  $A$  is a connected component of  $\{h \geq \rho\}$ , then, for every  $\rho' \in [0, \rho]$ , there exists exactly one connected component  $B$  of  $\{h \geq \rho'\}$  with  $A \subset B$ . Under some additional assumptions on  $\mu$  and the density  $h$ , this then leads to a *finite* tree, in which each node  $B$  is a connected component of some level set  $\{h \geq \rho'\}$  and all children of a node  $B$  are the connected components of  $\{h \geq \rho\}$  for some  $\rho > \rho'$  that are contained in  $B$ . We refer to Hartigan (1975); Stuetzle (2003); Chaudhuri and Dasgupta (2010); Stuetzle and Nugent (2010) for definitions and methods for estimating the structure of this tree. In particular, Chaudhuri and Dasgupta (2010) show that in a weak sense of Hartigan (1981), a modified single linkage algorithm converges to this tree under some assumptions on the density  $h$ . To be more precise, let  $A$  and  $A'$  be two different connected components of some level set of  $h$ , and  $D \in X^n$  be a data set from which the tree estimate is constructed. Furthermore, let  $A_D$  and  $A'_D$  be the smallest clusters in this tree estimate that satisfy  $A \cap D \subset A_D$  and  $A' \cap D \subset A'_D$ , respectively. Then the result by Chaudhuri and Dasgupta (2010) shows that we have  $A_D \cap A'_D = \emptyset$  with probability  $P^n$  converging to 1 for  $n \rightarrow \infty$ . Roughly speaking, this means that all parent/child relations of the cluster tree are eventually contained in the tree estimate, and Chaudhuri and Dasgupta (2010) actually show the latter by finite sample results. Unfortunately, however, neither of these results tell us *a)* how to find  $A_D$  and  $A'_D$  without knowing  $h$ , and *b)* how well  $A_D$  and  $A'_D$  approximate  $A$  and  $A'$ , respectively. Consequently, it seems fair to say that this approach reveals more about the cluster structure and less about the actual clusters.

In contrast to the these papers on the cluster tree approach this work focusses more on estimating the actual, maximal clusters<sup>1</sup>. Namely, we present a simple algorithm that automatically approximates the smallest possible value of  $\rho$  for which the level set contains more than one component. In addition, the algorithm approximates the resulting components arbitrarily well for  $n \rightarrow \infty$  under minimal and somewhat natural conditions, which include discontinuous densities.

Unlike basically all other papers on density based clustering, with the exception of Rinaldo and Wasserman (2010), we do not assume that the density  $h$  is continuous, or even Hölder continuous. While this approach enlarges the class of distributions significantly, it also produces some serious technical difficulties as we no longer have a canonical representative for the only  $\mu$ -almost surely defined density  $h$ . To be more precise, in general the topological properties of the level set  $\{h \geq \rho\}$  do dramatically depend on the chosen representative for the density, and hence the entire density based clustering approach becomes ill-defined. To address this problem, we first provide a definition for density level sets that make them actually *independent* of the chosen representative. As a consequence, it becomes mathematically rigorous to consider the infimum  $\rho^*$  over all levels  $\rho$  for which the

---

1. This distinction is, however, to some extend artificial, since recursively applying methods that estimate the maximal clusters well, automatically yields a consistent estimate of the cluster tree

corresponding density level sets contain more than connected component. For simplicity, we then assume that there exists some  $\rho^{**} > \rho^*$  such that the level sets for all  $\rho \in (\rho^*, \rho^{**}]$  contain exactly *two* connected components. Note that the persistence of the cluster structure over a small range of levels  $\rho \in (\rho^*, \rho^{**}]$  is assumed either explicitly or implicitly in basically all density based clustering approaches. On the other hand, the restriction to *two* components seems to be quite restrictive at first glance. Surprisingly, however, the opposite is true. To illustrate this, assume for simplicity that  $X = [0, 1]$  and  $h : X \rightarrow (0, \infty)$  is a continuous density with exactly two distinct strict local minima at say  $x_1$  and  $x_2$ . Now, if, e.g.,  $h(x_1) < h(x_2)$ , then  $\rho^* = h(x_1)$  and  $\rho^{**}$  can be any value with  $h(x_1) < \rho^{**} < h(x_2)$ . Moreover, for  $\rho \in (\rho^*, \rho^{**}]$ , the density level actually contains exactly two connected components, while for  $\rho > \rho^{**}$  the level sets may or may not contain three connected components. In other words, our assumption of two connected components for a small range above  $\rho^*$  would only be violated if  $h(x_1) = h(x_2)$ . Compared to the case  $h(x_1) \neq h(x_2)$ , the latter seems to be rather unlikely. Moreover, it is needless to say that in higher dimensions similar arguments can be made. Finally, it seems fair to say at this point that one more significant assumption on the level sets need to be made, namely one that excludes bridges and cusps that are too thin and long. However, while this is certainly unpleasant, it seems to be rather necessary, since such an assumption occurs in one form or the other in most articles dealing with density based clustering. With the assumptions described so far, our main result, Theorem 26, then shows that a simple histogram based algorithm both approximates  $\rho^*$  and the resulting clusters arbitrarily well for sample sizes  $n \rightarrow \infty$ .

The rest of this paper is organized as follows. In Section 2 we introduce our topologically robust notion of density level sets and establish some simple properties of these sets. We further consider maps that relate connected components of different level sets. These maps will be our fundamental tool for investigating the cluster structure of the true density sets and their empirical estimates. We further make the above notion of clusters rigorous and establish some results about the stability of the cluster structure under simple operations related to the blurriness of empirical estimates. In Section 3 we then present our algorithm and our main result, Theorem 26, followed by a discussion in Section 4. Finally, all proofs, together with some auxiliary results and some background material can be found in the appendix.

## 2. Preliminaries: Density level sets, connectivity, and clusters

In this section we introduce all notions related to the definition and analysis of clusters. We further present various technical result needed throughout the paper.

### 2.1. Density-independent density level sets and their regularity

Unlike to the rest of the paper, where we focus on compact metric spaces, we assume throughout this subsection that  $(X, d)$  denotes a complete separable metric space. Recall that compact metric spaces are both complete and separable, and hence everything developed in this subsection can actually be used in the remainder of the paper, too. Now, let  $\mathcal{B}(X)$  be the Borel  $\sigma$ -algebra on  $X$ ,  $\mu$  be a *known*  $\sigma$ -finite measure on  $\mathcal{B}(X)$ , and  $P$  be an *unknown*  $\mu$ -absolutely continuous probability measure on  $\mathcal{B}(X)$ . Recall that by Radon-Nykodým's theorem,  $P$  has a  $\mu$ -density  $h : X \rightarrow [0, \infty)$ , but this density is only  $\mu$ -almost

surely determined and therefore, for  $\rho \in [0, \infty)$ , the density level set  $\{h \geq \rho\}$  is also only  $\mu$ -almost surely determined. In particular, if we consider a measurable set  $A \subset X$  with

$$\mu(A \Delta \{h \geq \rho\}) = 0,$$

then there exists another  $\mu$ -density  $h' : X \rightarrow [0, \infty)$  of  $P$  such that  $A = \{h' \geq \rho\}$ . Now observe that the topological properties such as closedness or connectivity of  $\{h' \geq \rho\}$  may be quite different from those of  $\{h \geq \rho\}$ , since in general these properties may be changed by  $\mu$ -zero sets. Unfortunately, however, these topological properties play a crucial role in the definition of clusters, and hence we need a notion of “density level sets” that is independent of the particular choice of the density. To achieve this, observe that, for every  $\rho \in \mathbb{R}$ ,

$$\mu_\rho(A) := \mu(A \cap \{h \geq \rho\}), \quad A \in \mathcal{B}(X),$$

defines a measure  $\mu_\rho$  on  $(X, \mathcal{B}(X))$  that is *independent* of the particular choice of the  $\mu$ -density  $h$  of  $P$ . Consequently, the sets

$$\begin{aligned} M_\rho &:= \text{supp } \mu_\rho, \\ V_\rho &:= X \setminus M_\rho, \end{aligned}$$

where  $\text{supp } \mu_\rho$  denotes the support of  $\mu_\rho$ , are independent of this choice, too. In the following, we call  $M_\rho$  the density level set to the level  $\rho$ . To justify this notation, recall that by definition, the support of a measure is the complement of the largest open zero set, and hence  $\text{supp } \mu_\rho$  is the smallest closed subset  $B$  of  $X$  that satisfies  $\mu_\rho(X \setminus B) = 0$ . Moreover, recall that, for every measure on a complete, separable metric space, the support actually exists. Consequently, for any given  $\mu$ -density  $h : X \rightarrow [0, \infty)$  of  $P$ , we have

$$\mu(\{h \geq \rho\} \setminus M_\rho) = \mu(\{h \geq \rho\} \cap (X \setminus M_\rho)) = \mu_\rho(X \setminus M_\rho) = 0, \quad (1)$$

that is, up to  $\mu$ -zero sets no density level set  $\{h \geq \rho\}$  is larger than  $M_\rho$ . Moreover,  $M_\rho$  is the smallest closed set satisfying this equation, and hence we further obtain

$$M_\rho \subset \overline{\{h \geq \rho\}}, \quad (2)$$

where  $\overline{A}$  denotes the closure of an  $A \subset X$ . In addition, it is easy to check that we have

$$M_\rho = \{x \in X : \mu_\rho(U) > 0 \text{ for all open neighborhoods } U \text{ of } x\}. \quad (3)$$

Note that if  $\text{supp } \mu = X$ , we actually have  $V_\rho = \emptyset$  and  $M_\rho = X$  for all  $\rho \leq 0$ , but typically we are, of course, interested in the case  $\rho > 0$ , only. To state our first result, which provides a lower bound for the set  $M_\rho$ , we need to recall that the interior  $\mathring{A}$  of a set  $A \subset X$  is the largest open subset of  $A$ .

**Lemma 1** *Let  $(X, d)$  be a complete separable metric space,  $\mu$  be a  $\sigma$ -finite measure on  $X$  with  $\text{supp } \mu = X$ , and  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$ . Then, for all  $\mu$ -densities  $h$  of  $P$  and all  $\rho \in \mathbb{R}$ , we have*

$$\overline{\{h \geq \rho\}} \subset M_\rho \subset \overline{\{h \geq \rho\}}.$$

Lemma 1 in particular shows that  $\{h \geq \rho\} \subset M_\rho \subset \overline{\{h \geq \rho\}}$ . Therefore, the difference between the sets  $M_\rho$  and  $\{h \geq \rho\}$  is contained in the boundary of  $\{h \geq \rho\}$ , that is

$$M_\rho \Delta \{h \geq \rho\} \subset \partial\{h \geq \rho\}, \quad (4)$$

where  $\partial A := \overline{A} \setminus \mathring{A}$  denotes the boundary of a set  $A$ . Moreover, if  $P$  has a continuous density, we obtain the following corollary.

**Corollary 2** *Let  $(X, d)$  be a complete separable metric space,  $\mu$  be a  $\sigma$ -finite measure on  $X$  with  $\text{supp } \mu = X$ , and  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$  that has a continuous  $\mu$ -density  $h : X \rightarrow [0, \infty)$ . Then, for all  $\rho \in \mathbb{R}$ , we have*

$$\begin{aligned} \overline{\{h > \rho\}} &\subset M_\rho \subset \{h \geq \rho\}, \\ \{h > \rho\} &\subset \mathring{M}_\rho \subset \{h \geq \rho\}. \end{aligned}$$

The next lemma shows that the sets  $M_\rho$  and  $V_\rho$  are ordered the way one would expect density level sets to be ordered.

**Lemma 3** *Let  $(X, d)$  be a complete separable metric space,  $\mu$  be a  $\sigma$ -finite measure on  $X$ , and  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$ . Then, for all  $\rho_1 \leq \rho_2$ , we have*

$$\begin{aligned} M_{\rho_2} &\subset M_{\rho_1}, \\ V_{\rho_1} &\subset V_{\rho_2}. \end{aligned}$$

In turns out that we will not only need the equality  $\mu(\{h \geq \rho\} \setminus M_\rho) = 0$  but also the “converse” equality  $\mu(M_\rho \setminus \{h \geq \rho\}) = 0$ . This is ensured by the following definition.

**Definition 4** *Let  $(X, d)$  be a complete separable metric space,  $\mu$  be a  $\sigma$ -finite measure on  $X$ , and  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$ . We say that  $P$  is regular at level  $\rho \in \mathbb{R}$ , if*

$$\mu(M_\rho \setminus \{h \geq \rho\}) = 0$$

for one  $\mu$ -density (and thus all  $\mu$ -densities)  $h : X \rightarrow [0, \infty)$  of  $P$ .

Note that by (4) a probability measure  $P$  is regular at level  $\rho$ , if the boundary  $\partial\{h \geq \rho\}$  for one  $\mu$ -density  $h$  is a  $\mu$ -zero set. Let us now assume that  $P$  is regular at some level  $\rho$ . By (1) we then immediately see that

$$\mu(M_\rho \Delta \{h \geq \rho\}) = 0 \quad (5)$$

for all  $\mu$ -densities  $h$  of  $P$ . Furthermore, since  $V_\rho = X \setminus M_\rho$  and  $\{h < \rho\} = X \setminus \{h \geq \rho\}$ , the equation  $A \Delta B = (X \setminus A) \Delta (X \setminus B)$ , which holds for all  $A, B \subset X$ , shows

$$\mu(V_\rho \Delta \{h < \rho\}) = 0. \quad (6)$$

In other words, up to  $\mu$ -zero measures,  $M_\rho$  and  $V_\rho$  are the  $\rho$ -level sets of all  $\mu$ -densities  $h$  of  $P$ . The following lemma shows that there even exists a  $\mu$ -density  $h$  of  $P$  such that  $M_\rho = \{h \geq \rho\}$ .

**Lemma 5** Let  $(X, d)$  be a complete separable metric space,  $\mu$  be a  $\sigma$ -finite measure on  $X$ , and  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$ . Then, for  $\rho \in \mathbb{R}$ , the following statements are equivalent:

- i)  $P$  is regular at level  $\rho$ .
- ii) There exists a  $\mu$ -density  $h : X \rightarrow [0, \infty)$  of  $P$  such that  $\{h \geq \rho\}$  is closed.
- iii) There exists a  $\mu$ -density  $h : X \rightarrow [0, \infty)$  of  $P$  such that  $M_\rho = \{h \geq \rho\}$ .

In particular, if  $P$  has an upper semi-continuous  $\mu$ -density, then  $P$  is regular at every level.

The previous results may suggest that for continuous densities  $h$  we actually have  $M_\rho = \{h \geq \rho\}$ . In general, however, this is *not* the case. To see this, consider, e.g. a Lebesgue density that has a *strict* local maximum at  $x^* \in X$  and the corresponding level set  $\{h \geq \rho\}$  for  $\rho := h(x^*)$ . Moreover note, that not every probability measure  $P$  is regular. Indeed, it is possible to construct a Lebesgue-absolutely continuous probability measure  $P$  on  $[0, 1]$  that is not regular for a continuous range of levels  $\rho$ .

Besides the regularity, we need another notion ensuring that certain topological operations on the level sets do not change their mass.

**Definition 6** Let  $(X, d)$  be a complete separable metric space,  $\mu$  be a  $\sigma$ -finite measure on  $X$ , and  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$ . We say that  $P$  is normal at level  $\rho^* \geq 0$ , if

$$\mu(\bar{M}_{\rho^*} \setminus \dot{M}_{\rho^*}) = 0,$$

where  $\bar{M}_{\rho^*} := \bigcup_{\rho > \rho^*} M_\rho$  and  $\dot{M}_{\rho^*} := \bigcup_{\rho > \rho^*} \overset{\circ}{M}_\rho$ .

The following two lemmata provide sufficient conditions for normality. We begin with continuous densities.

**Lemma 7** Let  $(X, d)$  be a complete separable metric space,  $\mu$  be a  $\sigma$ -finite measure on  $X$  with  $\text{supp } \mu = X$ , and  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$  that has a continuous  $\mu$ -density  $h : X \rightarrow [0, \infty)$ . Then, for all  $\rho^* \geq 0$ , we have

$$\dot{M}_{\rho^*} = \bar{M}_{\rho^*},$$

and hence  $P$  is normal at every level.

The have already mentioned that regularity is ensured if there exists a  $\mu$ -density  $h$  of  $P$  with  $\mu(\partial\{h \geq \rho\}) = 0$ . The next lemma shows that this is also a sufficient for normality.

**Lemma 8** Let  $(X, d)$  be a complete separable metric space,  $\mu$  be a  $\sigma$ -finite measure on  $X$  with  $\text{supp } \mu = X$ , and  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$  that has a  $\mu$ -density  $h : X \rightarrow [0, \infty)$  such that there exists a  $\rho^* \geq 0$  with

$$\mu(\partial\{h \geq \rho\}) = 0$$

for all  $\rho > \rho^*$ . Then  $P$  is regular at every level  $\rho > \rho^*$  and normal at every level  $\rho \geq \rho^*$ .

## 2.2. Connectivity

We have already mentioned in the introduction that we will follow the idea of defining clusters by connected components. In this subsection, we introduce the necessary topological tools for this approach. Furthermore, we consider another, more quantitative notion of connectivity that is used in our algorithm.

Let us begin by introducing some notations. To this end, let  $(X, d)$  be a compact metric space. We write  $d(x, A) := \inf_{x' \in A} d(x, x')$  for the distance between some  $x \in X$  and  $A \subset X$ , and  $d(A, B) := \inf\{d(x, y) : x \in A, y \in B\}$  for the distance between  $A$  and another set  $B \subset X$ . Furthermore, for  $\delta > 0$ , we define the  $\delta$ -tube around  $A$  by

$$T_\delta(A) := \{x \in X : d(x, A) \leq \delta\}.$$

Lemma 27 in Subsection B collects some simple but useful facts about the  $\delta$ -tube around  $A$ . Let us further recall the definition of  $\tau$ -connected sets.

**Definition 9** *Let  $(X, d)$  be a compact metric space,  $A \subset X$  be a non-empty subset and  $\tau > 0$ . We say that  $x, x' \in A$  are  $\tau$ -connected in  $A$ , if there exist  $x_1, \dots, x_n \in A$  such that  $x_1 = x$ ,  $x_n = x'$  and  $d(x_i, x_{i+1}) < \tau$  for all  $i = 1, \dots, n - 1$ . Moreover, we say that  $A$  is  $\tau$ -connected, if all  $x, x' \in A$  are  $\tau$ -connected in  $A$ .*

It is easy to check that the property of being  $\tau$ -connected in  $A$  gives an equivalence relation for elements in  $A$ . We call the resulting equivalence classes the  $\tau$ -connected components of  $A$  and denote the set of all  $\tau$ -connected components of  $A$  by  $\mathcal{C}_\tau(A)$ . In addition, we define  $\mathcal{C}_\tau(\emptyset) := \emptyset$ .

Not surprisingly, the  $\tau$ -connected components of  $A \subset X$  are  $\tau$ -connected, see Lemma 28, and we always have  $|\mathcal{C}_\tau(A)| < \infty$  and  $d(A', A'') \geq \tau$  for all  $A', A'' \in \mathcal{C}_\tau(A)$ , see Lemma 29. Finally, if  $A$  is closed, so are the  $\tau$ -connected components of  $A$ .

In the following, we often have to compare the  $\tau$ -connected components of subsets  $A \subset B$ . The next lemma presents the fundamental tool for this task.

**Lemma 10** *Let  $(X, d)$  be a compact metric space,  $A \subset B$  be two non-empty subsets of  $X$  and  $\tau > 0$ . Then there exists exactly one map  $\zeta : \mathcal{C}_\tau(A) \rightarrow \mathcal{C}_\tau(B)$  such that*

$$A' \subset \zeta(A'), \quad A' \in \mathcal{C}_\tau(A).$$

We call  $\zeta$  the  $\tau$ -connected components relating map ( $\tau$ -CCRM) between  $A$  and  $B$ . Moreover, we sometimes write  $\zeta_{A,B} := \zeta$  when we have to emphasize the involved pair  $(A, B)$ .

Note that in general, the map  $\zeta_\tau$  is neither injective or surjective. Informally speaking,  $\zeta_\tau$  is injective, if and only if no  $\tau$ -connected component of  $B$  merges two  $\tau$ -connected components of  $A$ , while  $\zeta_\tau$  is surjective, if and only if  $B$  does not possess new  $\tau$ -connected components, i.e. there is no  $\tau$ -connected component  $B'$  of  $B$  with  $B' \subset B \setminus A$ . The next lemma introduces a very useful arithmetic property of  $\tau$ -CCRMs.

**Lemma 11** *Let  $(X, d)$  be a compact metric space,  $A \subset B \subset C$  be three non-empty subsets of  $X$  and  $\tau > 0$ . Then the  $\tau$ -CCRMs of these sets satisfy*

$$\zeta_{A,C} = \zeta_{B,C} \circ \zeta_{A,B}.$$

Let us now turn to the topological notion of connectivity that will be used in the definition of clusters. To this end, recall from topology that an  $A \subset X$  is called connected, if, for every pair  $A', A'' \subset A$  of closed disjoint subsets of  $A$  with  $A' \cup A'' = A$ , we have  $A' = \emptyset$  or  $A'' = \emptyset$ . Moreover, the maximal connected subsets of  $A$  are called the connected components of the space. It is well-known that these components form a partition of  $A$  and that every component is closed if  $A$  itself is closed. In the following, we denote the set of topologically connected components of  $A$  by  $\mathcal{C}(A)$ . Furthermore, to clearer distinct connected sets and components from  $\tau$ -connected sets and components, we often call the former *topologically* connected.

It can be easily shown, see Lemma 33, that, for compact metric spaces  $(X, d)$ , an  $A \subset X$  is topologically connected, if and only if it is  $\tau$ -connected for all  $\tau > 0$ . The following lemma investigates the relation between  $\mathcal{C}_\tau(A)$  and  $\mathcal{C}(A)$  in more detail.

**Lemma 12** *Let  $(X, d)$  be a compact metric space and  $A \subset X$  be a non-empty closed subset. Then the following statements hold:*

- i) *For all  $\tau > 0$ , there exists exactly one map  $\zeta : \mathcal{C}(A) \rightarrow \mathcal{C}_\tau(A)$  with*

$$A' \subset \zeta(A'), \quad A' \in \mathcal{C}(A).$$

*Moreover,  $\zeta$ , which we call the connected components relating map (CCRM) of  $A$ , is surjective.*

- ii) *If  $|\mathcal{C}(A)| < \infty$ , we have*

$$\tau_A^* := \min\{d(A', A'') : A', A'' \in \mathcal{C}(A) \text{ with } A' \neq A''\} > 0, \quad (7)$$

*where  $\min \emptyset := \infty$ . Moreover, for all  $\tau \in (0, \tau_A^*] \cap (0, \infty)$ , we have  $\mathcal{C}(A) = \mathcal{C}_\tau(A)$  and, for such  $\tau$ , the CCRM  $\zeta : \mathcal{C}(A) \rightarrow \mathcal{C}_\tau(A)$  is bijective. Finally, if  $\tau_A^* < \infty$ , that is,  $|\mathcal{C}(A)| > 1$ , we have*

$$\tau_A^* = \max\{\tau > 0 : \mathcal{C}(A) = \mathcal{C}_\tau(A)\}.$$

Note that, in general, a closed subset of  $A$  may have infinitely many topologically connected components as, e.g., the Cantor set shows. In this case, the second assertion of the lemma above is, in general, no longer true.

Our next goal is to find a connected components relating map for topologically connected components. This is done in the following lemma, which is a direct consequence of Lemma 12, and whose proof is therefore omitted.

**Lemma 13** *Let  $(X, d)$  be a compact metric space,  $A \subset B$  be two non-empty closed subsets of  $X$  that both have finitely many topologically connected components. Then there exists exactly one map  $\zeta : \mathcal{C}(A) \rightarrow \mathcal{C}(B)$  such that*

$$A' \subset \zeta(A'), \quad A' \in \mathcal{C}(A).$$

*In the following, we call  $\zeta$  the topologically connected components relating map (top-CCRM) between  $A$  and  $B$ . For  $\tau^* := \min\{\tau_A^*, \tau_B^*\}$  and all  $\tau \in (0, \tau^*]$ , we have  $\zeta = \zeta_\tau$ , where  $\zeta_\tau$  is the  $\tau$ -CCRM between  $A$  and  $B$ .*

Since top-CCRMs are  $\tau$ -CCRMs for all sufficiently small  $\tau > 0$ , the composition formula presented in Lemma 11 also holds for top-CCRMs. Moreover, by a straightforward modification of the proof of Lemma 11, we see that it also holds, if some (or all) top-CCRMs or  $\tau$ -CCRMs are replaced by the CCRMs found in part *i*) of Lemma 12.

The quantity  $\tau_A^*$  defined in (7) will play a crucial role in our analysis. However, we need to consider this quantity for more than one set, the next lemma establishes a relation between  $\tau_A^*$  and  $\tau_B^*$  if  $A \subset B$ .

**Lemma 14** *Let  $(X, d)$  be a compact metric space,  $A \subset B$  be two non-empty closed subsets of  $X$  that both have finitely many topologically connected components. If the top-CCRM  $\zeta : \mathcal{C}(A) \rightarrow \mathcal{C}(B)$  is injective, we have  $\tau_A^* \geq \tau_B^*$ .*

The following lemma establishes properties for the  $\tau$ -CCRM between a set  $A$  and  $T_\delta(A)$ . Roughly speaking, it states, that the  $\tau$ -connected component structure of  $T_\delta(A)$  is identical to that of  $A$ , if  $\tau > 0$  and  $\delta > 0$  are sufficiently small.

**Lemma 15** *Let  $(X, d)$  be a compact metric space,  $A \subset X$  be a non-empty subset of  $X$ . Then, for all  $\delta > 0$  and  $\tau > \delta$ , the following statements hold:*

- i) The set  $T_\delta(A')$  is  $\tau$ -connected for all  $A' \in \mathcal{C}_\tau(A)$ .*
- ii) The  $\tau$ -CCRM  $\zeta : \mathcal{C}_\tau(A) \rightarrow \mathcal{C}_\tau(T_\delta(A))$  is surjective.*
- iii) If  $A$  is closed and  $|\mathcal{C}(A)| < \infty$ , there exist  $\tau^* > 0$  and  $\delta^* > 0$  such that, for all  $\tau \in (0, \tau^*]$  and  $\delta \in (0, \delta^*]$  with  $\tau > \delta$ , the  $\tau$ -CCRM  $\zeta : \mathcal{C}_\tau(A) \rightarrow \mathcal{C}_\tau(T_\delta(A))$  is actually bijective. Moreover, we can choose  $\tau^*$  and  $\delta^*$  by the equation*

$$3\tau^* = 3\delta^* = \tau_A^*. \quad (8)$$

### 2.3. Clusters

In this subsection, we introduce our notion of clusters. We further present some results describing how robust the clusters are against horizontal blurriness. Finally, we introduce a notion that excludes thin bridges and cusps.

Let us begin with the following definition that describes distributions that have clusters.

**Definition 16** *Let  $(X, d)$  be a compact metric space,  $\mu$  be a finite Borel measure on  $X$  and  $P$  be a  $\mu$ -absolutely continuous Borel probability measure on  $X$ . Then we say that  $P$  can be topologically clustered between the critical levels  $\rho^* \geq 0$  and  $\rho^{**} > \rho^*$ , if  $P$  is normal at level  $\rho^*$  and, for all  $\rho \in [0, \rho^{**}]$ , the following conditions hold:*

- i) The set  $M_\rho$  has either one or two topologically connected components.*
- ii) If  $|\mathcal{C}(M_\rho)| = 1$ , then  $\rho \leq \rho^*$ .*
- iii) If  $|\mathcal{C}(M_\rho)| = 2$ , then  $\rho \geq \rho^*$  and the top-CCRM  $\zeta : \mathcal{C}(M_{\rho^{**}}) \rightarrow \mathcal{C}(M_\rho)$  is bijective.*
- iv)  $P$  is regular at level  $\rho$ .*

Note that the definition above does not exclude the case  $|\mathcal{C}(M_{\rho^*})| = 1$ , and hence the elements of  $\mathcal{C}(M_{\rho^*})$  cannot be used to define the clusters of  $P$ . On the other hand, for  $\rho > \rho^*$ , each  $A \in \mathcal{C}(M_\rho)$  should be a subset of a cluster of  $P$ . This idea is used in the following definition, which defines the clusters of  $P$  by a limit for  $\rho \searrow \rho^*$ .

**Definition 17** Let  $(X, d)$  be a compact metric space,  $\mu$  be a finite Borel measure on  $X$  and  $P$  be a  $\mu$ -absolutely continuous Borel probability measure on  $X$  that can be topologically clustered between the critical levels  $\rho^*$  and  $\rho^{**}$ . For  $\rho \in (\rho^*, \rho^{**}]$ , we write  $\zeta_\rho : \mathcal{C}(M_{\rho^{**}}) \rightarrow \mathcal{C}(M_\rho)$  for the top-CCRM. Moreover, let  $A_1$  and  $A_2$  be the topologically connected components of  $M_{\rho^{**}}$ . Then the sets

$$A_i^* := \bigcup_{\rho \in (\rho^*, \rho^{**}]} \zeta_\rho(A_i), \quad i \in \{1, 2\},$$

are called the topological clusters of  $P$ .

By the bijectivity of the maps  $\zeta_\rho$ , it is straightforward to show that  $A_1^* \cap A_2^* = \emptyset$ . In general, however, the clusters may touch each other, that is, we may have  $d(A_1^*, A_2^*) = 0$ . For example, if  $P$  is a mixture of two Gaussians with different centers but same variance, then it is easy to check that the two clusters are only separated by a hyperplane, and therefore they do touch each other.

Using finitely many samples, we can only expect estimates of the level sets  $M_\rho$  that are both vertically and *horizontally* blurry. To address the latter issue, we define, for  $\delta > 0$  and  $\rho \geq 0$ , the sets

$$\begin{aligned} M_{\rho, \delta} &:= T_\delta(M_\rho) \\ V_{\rho, \delta} &:= X \setminus T_\delta(X \setminus M_\rho) = X \setminus T_\delta(V_\rho). \end{aligned}$$

Note that  $M_{\rho, \delta}$  is obtained from  $M_\rho$  by adding a  $\delta$ -tube, while  $V_{\rho, \delta}$  is obtained from  $M_\rho$  by removing a  $\delta$ -tube. Our next goal is to investigate, how the component structure of  $M_\rho$  is preserved under these operations. The first result in this direction establishes some permanence properties that hold without further assumptions. To appreciate its rather theoretically appearing statements recall from the previous subsection that CCRMs between two sets  $A$  and  $B$  are bijective, if *a*) the components of  $A$  are not glued together in  $B$  and *b*) every component in  $B$  already appears in  $A$ .

**Theorem 18** Let  $(X, d)$  be a compact metric space,  $\mu$  be a finite Borel measure on  $X$  and  $P$  be a  $\mu$ -absolutely continuous Borel probability measure on  $X$  that can be topologically clustered between the critical levels  $\rho^*$  and  $\rho^{**}$ . For all  $\varepsilon^* > 0$  with  $\rho^* + \varepsilon^* \leq \rho^{**}$ , we define  $\delta_{\varepsilon^*} > 0$  and  $\tau_{\varepsilon^*} > 0$  by

$$3\delta_{\varepsilon^*} = 3\tau_{\varepsilon^*} = \tau_{M_{\rho^* + \varepsilon^*}}^*, \tag{9}$$

where  $\tau_{M_{\rho^* + \varepsilon^*}}^* > 0$  is the quantity considered in Lemma 12. Then, for all  $\delta \in (0, \delta_{\varepsilon^*}]$ ,  $\tau \in (0, \tau_{\varepsilon^*}]$  with  $\delta < \tau$ , and all  $\rho \in [0, \rho^{**}]$ , the following statements hold:

- i) The set  $M_{\rho, \delta}$  has either one or two  $\tau$ -connected components.
- ii) If  $\rho \geq \rho^* + \varepsilon^*$ , then  $|\mathcal{C}_\tau(M_{\rho, \delta})| = 2$  and the CCRM  $\zeta : \mathcal{C}(M_\rho) \rightarrow \mathcal{C}_\tau(M_{\rho, \delta})$  is bijective.

- iii) If  $|\mathcal{C}_\tau(M_{\rho,\delta})| = 2$ , then  $\rho \geq \rho^*$  and the  $\tau$ -CCRM  $\zeta : \mathcal{C}_\tau(M_{\rho^{**},\delta}) \rightarrow \mathcal{C}_\tau(M_{\rho,\delta})$  is bijective.
- iv) If the  $\tau$ -CCRM  $\zeta^{**} : \mathcal{C}_\tau(V_{\rho^{**},\delta}) \rightarrow \mathcal{C}_\tau(M_{\rho^{**},\delta})$  is bijective and  $|\mathcal{C}_\tau(V_{\rho,\delta})| = 1$ , then we have  $\rho < \rho^* + \varepsilon^*$ .

The first three statements of Theorem 18 basically show that the connected component structure of  $M_\rho$  is not changed when  $\delta$ -tubes are added. Moreover, the assumed bijectivity of  $\zeta^{**} : \mathcal{C}_\tau(V_{\rho^{**},\delta}) \rightarrow \mathcal{C}_\tau(M_{\rho^{**},\delta})$  in iv) means that the  $\tau$ -connected component structure of  $M_{\rho^{**}}$  is not changed by removing  $\delta$ -tubes, and the corresponding conclusion essentially states that this is actually true for all levels  $\rho \in [\rho^* + \varepsilon^*, \rho^{**}]$ .

To ensure the same stability for removing  $\delta$ -tubes, we need the following additional assumption.

**Definition 19** Let  $(X, d)$  be a compact metric space,  $\mu$  be a  $\sigma$ -finite Borel measure on  $X$  and  $P$  be a  $\mu$ -absolutely continuous Borel probability measure on  $X$  that can be topologically clustered between the critical levels  $\rho^*$  and  $\rho^{**}$ . Then we say that  $P$  has two thick clusters of order  $\gamma \in (0, 1]$ , if there exist  $c \geq 1$  and  $\tilde{\delta}_0 \in (0, 1]$  such that, for all  $\delta \in (0, \tilde{\delta}_0]$ ,  $\rho \in [0, \rho^{**}]$ , we have

$$d(x, V_{\rho,\delta}) < c \delta^\gamma, \quad x \in M_\rho.$$

In this case, we call  $\psi : (0, \infty) \rightarrow (0, \infty)$ , defined by  $\psi(\delta) := 2c\delta^\gamma$ , the corresponding thickness function.

Roughly speaking, the definition above ensures that every point of  $M_\rho$  is close to the set  $V_{\rho,\delta}$  that results from removing a  $\delta$ -tube from  $M_\rho$ . Intuitively, this excludes very thin cusps and bridges, where the thinness and length of both is controlled by  $\gamma$ . Note that an assumption of similar spirit is often made in density-based cluster analysis, we refer, e.g., to Cuevas et al. (2000); Rigolet (2007) for some examples in this direction.

Moreover note that, if  $X \subset \mathbb{R}$  is an interval and  $P$  can be topologically clustered between the critical levels  $\rho^*$  and  $\rho^{**}$ , then every level set  $M_\rho$  consists of either one or two closed intervals, since intervals are the only topologically connected sets in  $\mathbb{R}$ . Using this, it is then straightforward to show that  $P$  actually has two thick clusters of order  $\gamma = 1$ , and that we can further use every constant  $c > 1$ . Consequently, we can, e.g., consider the thickness function  $\psi(\delta) := 3\delta$ ,  $\delta > 0$ .

The next result, which is the counterpart of Theorem 18, shows that, for thick clusters, the connected component structure of  $M_\rho$  is not changed when removing  $\delta$ -tubes.

**Theorem 20** Let  $(X, d)$  be a compact metric space,  $\mu$  be a finite measure on  $X$ , and  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$  that has two thick clusters of order  $\gamma \in (0, 1]$  between the critical levels  $\rho^*$  and  $\rho^{**}$ . Let  $\psi$  be the corresponding thickness function. Moreover, for some fixed  $\varepsilon^* > 0$  with  $\varepsilon^* \leq \rho^{**} - \rho^*$  we define  $\delta_{\varepsilon^*} > 0$  and  $\tau_{\varepsilon^*} > 0$  by (9). Then, for all  $\delta \in (0, \delta_{\varepsilon^*}]$ ,  $\tau \in (0, \tau_{\varepsilon^*}]$  with  $\psi(\delta) < \tau$  and  $\delta \leq \tilde{\delta}_0$ , and all  $\rho \in [0, \rho^{**}]$ , the following statements hold:

- i) The set  $V_{\rho,\delta}$  has either one or two  $\tau$ -connected components.
- ii) The  $\tau$ -CCRM  $\zeta^{**} : \mathcal{C}_\tau(V_{\rho^{**},\delta}) \rightarrow \mathcal{C}_\tau(M_{\rho^{**},\delta})$  is bijective.

iii) If  $|\mathcal{C}_\tau(V_{\rho,\delta})| = 2$ , then  $\rho \geq \rho^*$  and the  $\tau$ -CCRM  $\zeta : \mathcal{C}_\tau(V_{\rho^{**},\delta}) \rightarrow \mathcal{C}_\tau(V_{\rho,\delta})$  is bijective.

Intuitively, considering  $\mathcal{C}_\tau(V_{\rho,\delta})$  rather than  $\mathcal{C}(V_{\rho,\delta})$  means that we add a  $\tau$ -tube around  $V_{\rho,\delta}$ . By Theorem 20, the thickness of the level sets then ensure that  $\mathcal{C}_\tau(V_{\rho,\delta})$  and  $M_\rho$  have the same component structure, or to say it in simple words, considering  $\tau$ -connected components glues together what has been accidentally cut by removing  $\delta$ -tubes.

### 3. The algorithm and its consistency

In this section, we introduce our clustering algorithm and present our main results on its clustering ability.

Let us begin by recalling that histograms are based on partitions of the input space  $X$ . In the following, we need to ensure that the partitions we use are regularly behaved. To this end, we need the diameter of a subset  $A \subset X$ , that is,

$$\text{diam } A := \sup_{x,x' \in A} d(x, x').$$

Now, the following definition describes partitions that are controlled both in size and measure.

**Definition 21** Let  $(X, d)$  be a compact metric space and  $\mu$  be a finite measure on  $X$  with  $\text{supp } \mu = X$ . If there exist constants  $d_X > 0$  and  $\kappa_X > 0$  such that, for all  $\delta \in (0, 1]$ , there exists a finite partition  $\mathcal{A}_\delta = (A_1, \dots, A_m)$  of  $X$  such that

$$\begin{aligned} \text{diam } A_i &\leq \delta, \\ m &\leq \kappa_X \delta^{-d_X}, \\ \mu(A_i) &\geq \kappa_X^{-1} \delta^{d_X} \end{aligned}$$

for all  $i = 1, \dots, m$ , then we say that the triple  $(X, d, \mu)$  admits uniform  $d_X$ -dimensional partitions. Moreover, we call each such  $\mathcal{A}_\delta$  a  $\delta$ -uniform partition of  $X$ .

The easiest yet most important examples of uniform partitions are hypercube partitions. To be more precise, let  $X := [0, 1]^d$ , the  $d$ -dimensional cube equipped with the metric defined by the supremum norm  $\|\cdot\|_{\ell_\infty^d}$ . For  $\delta \in (0, 1]$ , there then exists a unique  $\ell \in \mathbb{N}$  with  $\frac{1}{\ell+1} < \delta \leq \frac{1}{\ell}$ . We define  $h := \frac{1}{\ell+1}$  and write  $\mathcal{A}_\delta$  for the usual partition of  $[0, 1]^d$  into hypercubes of length  $h$ . Then, for each  $A_i \in \mathcal{A}_\delta$ , we clearly have  $\text{diam } A_i = h \leq \delta$  and  $\lambda^d(A_i) = h^d \geq 2^{-d} \delta^d$ , where  $\lambda^d$  denotes the  $d$ -dimensional Lebesgue measure. Moreover, we obviously have  $|\mathcal{A}_\delta| = h^{-d} \leq 2^d \delta^{-d}$ , where  $|\mathcal{A}_\delta|$  denotes the cardinality of  $\mathcal{A}$ . Consequently,  $\mathcal{A}_\delta$  is an  $\delta$ -uniform partition of  $X$  with  $d_X := d$  and  $\kappa_X := 2^d$ .

Let us now assume that  $(X, d, \mu)$  admits uniform  $d_X$ -dimensional partitions. Moreover, for fixed  $\delta > 0$ , let  $\mathcal{A}_\delta = (A_1, \dots, A_m)$  be a  $\delta$ -uniform partition of  $X$ . Given a probability measure  $P$  on  $X$ , we then define the corresponding histogram by

$$\bar{h}_{P, \mathcal{A}_\delta}(x) := \sum_{j=1}^m \frac{P(A_j)}{\mu(A_j)} \cdot \mathbf{1}_{A_j}(x), \quad x \in X,$$

where  $\mathbf{1}_A$  denotes the indicator function of a set  $A$ . If the partition is known from the context, we further write  $\bar{h}_{P,\delta} := \bar{h}_{P,\mathcal{A}_\delta}$  to simplify notation. Let us now assume that we have a data set  $D = (x_1, \dots, x_n) \in X^n$ . In a slight abuse of notation, we then denote the corresponding empirical measure by  $D$ , that is  $D := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ , where  $\delta_x$  denotes the Dirac measure at the point  $x$ . For  $A \subset X$  this gives

$$D(A) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_A(x_i),$$

and the corresponding (empirical) histogram is

$$\bar{h}_{D,\mathcal{A}_\delta}(x) = \sum_{j=1}^m \frac{D(A_j)}{\mu(A_j)} \cdot \mathbf{1}_{A_j}(x), \quad x \in X. \quad (10)$$

Our first result in this section shows, that, for i.i.d. observations  $D$ , the histogram  $\bar{h}_{D,\mathcal{A}_\delta}$  approximates  $\bar{h}_{P,\mathcal{A}_\delta}$  uniformly.

**Theorem 22** *Let  $(X, d)$  be a compact metric space and  $\mu$  be a finite measure on  $X$  such that  $(X, d, \mu)$  admits uniform  $d_X$ -dimensional partitions. Moreover, let  $P$  be a probability measure on  $X$ ,  $\delta > 0$ , and  $\mathcal{A}_\delta$  be a  $\delta$ -uniform partition of  $X$ . Then, for all  $n \geq 1$  and all  $\varepsilon > 0$ , we have*

$$P^n \left( \{D \in X^n : \|\bar{h}_{D,\mathcal{A}_\delta} - \bar{h}_{P,\mathcal{A}_\delta}\|_\infty < \varepsilon\} \right) \geq 1 - 2\kappa_X \exp \left( -d_X \ln \delta - \frac{2\delta^{2d_X} \varepsilon^2 n}{\kappa_X^2} \right).$$

Moreover, if  $P$  is  $\mu$ -absolutely continuous and there exists a bounded  $\mu$ -density  $h$  of  $P$ , then we have

$$P^n \left( \{D \in X^n : \|\bar{h}_{D,\mathcal{A}_\delta} - \bar{h}_{P,\mathcal{A}_\delta}\|_\infty < \varepsilon\} \right) \geq 1 - 2\kappa_X \exp \left( -d_X \ln \delta - \frac{3\varepsilon^2 \delta^{d_X} n}{\kappa_X (6\|h\|_\infty + 2\varepsilon)} \right).$$

Our clustering algorithm will rely on an empirical histogram. To be more precise, let us assume that, for some fixed  $\varepsilon > 0$ , we have a function  $\hat{h} : X \rightarrow \mathbb{R}$  that is a uniform  $\varepsilon$ -approximate of  $\bar{h}_{P,\mathcal{A}_\delta}$ , i.e.,

$$\|\hat{h} - \bar{h}_{P,\mathcal{A}_\delta}\|_\infty \leq \varepsilon.$$

Note that, by Theorem 22, empirical histograms are such  $\varepsilon$ -approximates with high probability. We write

$$\hat{f}_\rho := \text{sign}(\hat{h} - \rho), \quad \rho \geq 0, \quad (11)$$

where  $\text{sign}(\cdot)$  denotes the usual sign function, that is,  $\text{sign } t := 1$  if  $t \geq 0$  and  $\text{sign } t := -1$ , otherwise. Since  $\bar{h}_{P,\mathcal{A}_\delta}$  can be viewed as an approximation of the  $\mu$ -densities of  $P$ ,  $\hat{h}$  can also be viewed as such an approximation. Following this intuition,  $\{\hat{f}_\rho = -1\}$  and  $\{\hat{f}_\rho = 1\}$  can be viewed as approximations of the sets  $V_\rho$  and  $M_\rho$ , respectively. However, using finitely many samples, we can only expect estimates of the level sets  $M_\rho$  that are both horizontally and vertically blurry. The following lemma makes this intuition precise with the help of the sets  $V_{\rho,\delta}$  and  $M_{\rho,\delta}$  defined earlier.

---

**Algorithm 3.1** Estimate clusters with the help of empirical histograms

---

**Require:** Some  $\delta > 0$ ,  $\tau > 0$ , and  $\varepsilon > 0$ .A  $\delta$ -uniform partition  $\mathcal{A}_\delta$  of  $X$ .A dataset  $D \in X^n$ .**Ensure:** An estimate of the topological clusters  $A_1^*$  and  $A_2^*$ .1: Compute the empirical histogram  $\bar{h}_{D, \mathcal{A}_\delta}$ .2:  $\rho \leftarrow -\varepsilon$ 3: **repeat**4:    $\rho \leftarrow \rho + \varepsilon$ 5:   Compute  $\hat{f}_\rho$  by (11).6:   Identify the  $\tau$ -connected components  $B'_1, \dots, B'_M$  of  $\{\hat{f}_\rho = 1\}$  satisfying

$$B'_i \cap \{\hat{f}_{\rho+2\varepsilon} = 1\} \neq \emptyset.$$

7: **until**  $M \neq 1$ 8: Compute  $\hat{f}_{\rho+2\varepsilon}$  by (11).9: Identify the  $\tau$ -connected components  $B'_1, \dots, B'_M$  of  $\{\hat{f}_{\rho+2\varepsilon} = 1\}$  satisfying

$$B'_i \cap \{\hat{f}_{\rho+4\varepsilon} = 1\} \neq \emptyset.$$

10: **return**  $\rho$  and  $B'_1, \dots, B'_M$ .

---

**Lemma 23** Let  $(X, d)$  be a compact metric space and  $\mu$  be a finite measure on  $X$  such that  $(X, d, \mu)$  admits uniform  $d_X$ -dimensional partitions. Moreover, let  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$  and  $\hat{h} : X \rightarrow \mathbb{R}$  be a function with  $\|\hat{h} - \bar{h}_{P, \mathcal{A}_\delta}\|_\infty \leq \varepsilon$  for some  $\varepsilon > 0$ . Then, for all  $\delta > 0$  and  $\rho \geq 0$ , and  $\hat{f}_\rho$  defined by (11), we have:

i) If  $P$  is regular at the level  $\rho + \varepsilon$ , then  $V_{\rho+\varepsilon, \delta} \subset \{\hat{f}_\rho = 1\}$ .ii) If  $P$  is regular at the level  $\rho - \varepsilon$ , then  $\{\hat{f}_\rho = 1\} \subset M_{\rho-\varepsilon, \delta}$ .

Motivated by Lemma 23, our next goal is to relate the  $\tau$ -connected components of our estimate  $\{\hat{f}_\rho = 1\}$  to the  $\tau$ -connected components of  $V_{\rho, \delta}$ .

**Theorem 24** Let  $(X, d)$  be a compact metric space,  $\mu$  be a finite measure on  $X$  such that  $(X, d, \mu)$  admits uniform  $d_X$ -dimensional partitions, and  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$  that has two thick clusters of order  $\gamma \in (0, 1]$  between the critical levels  $\rho^*$  and  $\rho^{**}$ . Let  $\psi$  be the corresponding thickness function. Moreover, for some fixed  $\varepsilon^* > 0$  with  $\varepsilon^* \leq \rho^{**} - \rho^*$  we define  $\delta_{\varepsilon^*} > 0$  and  $\tau_{\varepsilon^*} > 0$  by (9). Let us further fix some  $\varepsilon \in (0, \varepsilon^*]$ ,  $\epsilon \geq 0$ ,  $\delta \in (0, \delta_{\varepsilon^*}]$  with  $\delta \leq \tilde{\delta}_0$ , and  $\rho \in [0, \rho^{**} - 3\varepsilon - \epsilon]$ . In addition, let  $\hat{h} : X \rightarrow \mathbb{R}$  be a uniform  $\varepsilon$ -approximate of  $\bar{h}_{P, \delta}$  and  $\hat{f}_\rho$  be the function defined by (11). Then, for all  $\tau \in (0, \tau_{\varepsilon^*}]$  with  $\psi(\delta) < \tau$ , the following disjoint union holds

$$\mathcal{C}_\tau(\{\hat{f}_\rho = 1\}) = \zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon, \delta})) \cup \{B' \in \mathcal{C}_\tau(\{\hat{f}_\rho = 1\}) : B' \cap \{\hat{f}_{\rho+2\varepsilon+\epsilon} = 1\} = \emptyset\},$$

where  $\zeta : \mathcal{C}_\tau(V_{\rho+\varepsilon, \delta}) \rightarrow \mathcal{C}_\tau(\{\hat{f}_\rho = 1\})$  is the  $\tau$ -CCRM.

Theorem 24 shows that eventually all  $\tau$ -connected components  $B'$  of our estimate  $\{\hat{f}_\rho = 1\}$  of  $M_\rho$  are either contained in  $\zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon,\delta}))$  or satisfy  $B' \cap \{\hat{f}_{\rho+2\varepsilon+\epsilon} = 1\} = \emptyset$ . Now, the latter components are easy to identify and remove, and therefore we have a device that allows us to eventually identify exactly the  $\tau$ -connected components  $B'$  that are contained in  $\zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon,\delta}))$ . This suggests that, for sufficiently small  $\delta > 0$ ,  $\tau > 0$ ,  $\varepsilon > 0$ , and  $\epsilon \geq 0$ , we only need to scan through the values of  $\rho$ . Algorithm 3.1 formalizes this idea.

Note that Algorithm 3.1 stops if either  $M > 1$  or  $M = 0$  components are identified for the current level set  $\rho$ . Moreover, the latter is eventually satisfied, since

$$\|\bar{h}_{D,\mathcal{A}_\delta}\|_\infty \leq \kappa_X \delta^{-d_X} \sum_{i=1}^m D(A_i) = \kappa_X \delta^{-d_X},$$

yields  $\{\hat{f}_\rho = 1\} = \emptyset$  for all  $\rho > \kappa_X \delta^{-d_X}$ . In the following, we denote the level returned by Algorithm 3.1 by  $\rho^*(D)$ . The following theorem shows that  $\rho^*(D)$  is close to  $\rho^*$ , whenever the empirical histogram approximates the true histogram.

**Theorem 25** *Let  $(X, d)$  be a compact metric space and  $\mu$  be a finite measure on  $X$  such that  $(X, d, \mu)$  admits uniform  $d_X$ -dimensional partitions. Moreover, let  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$  that has two thick clusters of order  $\gamma \in (0, 1]$  between the critical levels  $\rho^*$  and  $\rho^{**}$  and let  $\psi$  be the corresponding thickness function. Moreover, we fix an  $\varepsilon^* > 0$  that satisfies  $\varepsilon^* < (\rho^{**} - \rho^*)/8$  and define  $\delta_{\varepsilon^*} > 0$  and  $\tau_{\varepsilon^*} > 0$  by (9). Then, for all fixed  $n \geq 1$ ,  $\varepsilon \in (0, \varepsilon^*]$ ,  $\delta \in (0, \delta_{\varepsilon^*}]$ , and  $\tau \in (0, \tau_{\varepsilon^*}]$  with  $\psi(\delta) < \tau$  and  $\delta \leq \tilde{\delta}_0$ , and all data sets  $D \in X^n$  for which  $\|\bar{h}_{D,\mathcal{A}_\delta} - \bar{h}_{P,\mathcal{A}_\delta}\|_\infty \leq \varepsilon$  holds, the following statements are true:*

- i)  $\rho^*(D) \in [\rho^* - \varepsilon, \rho^* + \varepsilon^* + 2\varepsilon]$ .
- ii)  $|\mathcal{C}_\tau(V_{\rho^*(D)+3\varepsilon,\delta})| = 2$  and the  $\tau$ -CCRM  $\zeta : \mathcal{C}_\tau(V_{\rho^*(D)+3\varepsilon,\delta}) \rightarrow \mathcal{C}_\tau(\{\hat{f}_{\rho^*(D)+2\varepsilon} = 1\})$  is injective.
- iii) Algorithm 3.1 returns the two  $\tau$ -connected components of  $\zeta(\mathcal{C}_\tau(V_{\rho^*(D)+3\varepsilon,\delta}))$ .
- iv) There exist CCRMs  $\zeta_{\rho^{**}} : \mathcal{C}_\tau(V_{\rho^{**},\delta}) \rightarrow \mathcal{C}(M_{\rho^{**}})$  and  $\zeta_{\rho^*(D)+3\varepsilon} : \mathcal{C}_\tau(V_{\rho^*(D)+3\varepsilon,\delta}) \rightarrow \mathcal{C}(M_{\rho^*(D)+3\varepsilon})$  such that the following diagram

$$\begin{array}{ccc} \mathcal{C}_\tau(V_{\rho^{**},\delta}) & \xrightarrow{\zeta_{\rho^{**}}} & \mathcal{C}(M_{\rho^{**}}) \\ \zeta_{\rho^{**},\rho^*(D)+3\varepsilon} \downarrow & & \downarrow \tilde{\zeta} \\ \mathcal{C}_\tau(V_{\rho^*(D)+3\varepsilon,\delta}) & \xrightarrow{\zeta_{\rho^*(D)+3\varepsilon}} & \mathcal{C}(M_{\rho^*(D)+3\varepsilon}) \end{array}$$

commutes, where  $\zeta_{\rho^{**},\rho^*(D)+3\varepsilon}$  is the  $\tau$ -CCRM and  $\tilde{\zeta}$  is the top-CCRM. Moreover, every map in the diagram is bijective.

To fully appreciate Theorem 25 let us assume that we are in the situation of this theorem. Moreover, let  $A_1$  and  $A_2$  be the topologically connected components of  $M_{\rho^{**}}$  and

$V_1''$  and  $V_2''$  be the  $\tau$ -connected components of  $V_{\rho^{**}, \delta}$ . In addition, let  $V_1'$  and  $V_2'$  be the  $\tau$ -connected components of  $\mathcal{C}_\tau(V_{\rho^*(D)+3\varepsilon, \delta})$  and  $B_1(D)$  and  $B_2(D)$  be the components returned by Algorithm 3.1. By Theorem 25, we may assume without loss of generality that  $V_i'' \subset A_i$ ,  $V_i'' \subset V_i'$ , and  $V_i' \subset B_i(D)$  for  $i = 1, 2$ . This yields  $V_i'' \subset B_i(D)$  and  $V_i'' \subset A_i^*$ , that is,  $V_i'' \subset B_i(D) \cap A_i^*$ . Consequently, the returned components  $B_i(D)$  contain a chunk of the desired clusters  $A_i^*$ ,  $i = 1, 2$ . Our next and final goal is to show that  $B_i(D) \Delta A_i^*$  actually becomes arbitrarily small. To this end, we assume in the following that Algorithm 3.1 always returns two components, denoted by  $B_1(D)$  and  $B_2(D)$ . Note that this can be easily enforced by a simple modification of its return statement in line 10 of its pseudo-code.

With these preparations, we are in the position to put all pieces together. This is done in the following main result that establishes a type of clustering consistency for Algorithm 3.1.

**Theorem 26** *Let  $(X, d)$  be a compact metric space and  $\mu$  be a finite measure on  $X$  such that  $(X, d, \mu)$  admits uniform  $d_X$ -dimensional partitions. Moreover, let  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$  that has two thick clusters of order  $\gamma \in (0, 1]$  between the critical levels  $\rho^*$  and  $\rho^{**}$  and let  $\psi$  be the corresponding thickness function. Furthermore, let  $(\varepsilon_n)$ ,  $(\delta_n)$ , and  $(\tau_n)$  be strictly positive sequences converging to zero such that  $\psi(\delta_n) < \tau_n$  and*

$$d_X \kappa_X^2 \ln \delta_n + 2\delta_n^{2d_X} \varepsilon_n^2 n \rightarrow \infty.$$

*For  $n \geq 1$  consider Algorithm 3.1 with the input parameters  $\varepsilon_n$ ,  $\delta_n$ , and  $\tau_n$ . Then,  $\rho^*(D) \rightarrow \rho^*$  in probability  $P^\infty$  for  $n \rightarrow \infty$  and, for all  $\epsilon > 0$ , we have*

$$\lim_{n \rightarrow \infty} P^n \left( \{D \in X^n : \mu(B_1(D) \Delta A_1^*) + \mu(B_2(D) \Delta A_2^*) \leq \epsilon\} \right) = 1.$$

*Here we use the numbering convention of  $B_1(D)$  and  $B_2(D)$  described in the paragraph above.*

Theorem 26 shows that Algorithm 3.1 asymptotically recovers the clusters  $A_1^*$  and  $A_2^*$ , whenever the distribution  $P$  has clusters that are thicker than a pre-described order. In other words, as soon as we assume a minimal thickness, we are able to recover the clusters. Moreover, we have already mentioned previously, that, for intervals  $X \subset \mathbb{R}$ , we automatically have thickness of order  $\gamma = 1$ , and hence Algorithm 3.1 asymptotically recovers the clusters, e.g., for every distribution  $P$  on intervals that can be topologically clustered. Note that it is easy to construct distributions in this class that do not have a continuous density, for example consider the distribution  $P$  on  $X := [0, 1]$  that has the Lebesgue density  $h := \mathbf{1}_{[0, 1/4] \cup [3/4, 1]} + 0.5 \cdot \mathbf{1}_{[0, 1]}$ , and whose clusters are given by  $A_1^* := [0, 1/4]$  and  $A_2^* := [3/4, 1]$ . It is obvious, that similar constructions can also be made in higher dimensions, and finally, such examples are, of course, by no means the only examples of distributions for which the clusters can be recovered by Algorithm 3.1.

Furthermore, note that although Theorem 26 presents an asymptotic result, its entire proof uses finite-sample results and estimates, i.e., we could have also stated a result of the form: if the algorithm parameters are smaller than some thresholds determined in the proof of Theorem 26, then  $\mu(B_1(D) \Delta A_1^*) + \mu(B_2(D) \Delta A_2^*) \leq \epsilon$  holds with probability  $P^n$  not smaller than some value also determined in the proof. Since the presented algorithm

was meant to be a proof-of-concept rather than an algorithm actually used in dimensions greater than, say, 2 or 3, we decided to omit the technically rather cumbersome formulation of such a result.

#### 4. Discussion

The goal of this work was to provide the first density-based clustering algorithm that, under mild assumptions on the density  $h$ , can choose the density level in a data-dependent and asymptotically optimal way and completely recovers the corresponding clusters even when they touch each other.

Although the algorithm works in theory, we do not expect it to perform well in most practical situations. Let us therefore briefly describe what should be done to obtain a more interesting algorithm:

- The algorithm should be based on density level set estimators that are better than histograms. A natural first alternative in this direction would be kernel density rules since they have already been successfully considered in the single level clustering problem.
- The algorithm and its analysis should be extended to situations in which  $P$  has either only  $N = 1$  or  $N > 2$  clusters. Note that the first scenario can probably be rather easily analyzed with our techniques, while the second scenario probably needs a refined algorithm, first. In this direction note that our proofs already show that the algorithm recovers at least two of the  $N$  clusters. So far, however, we cannot ensure that it accidentally glues some of the  $N$  clusters together.
- The algorithm should not only determine the level  $\rho$  in a data-dependent way, but also the other algorithm parameters  $\delta$ ,  $\varepsilon$ , and  $\tau$ . Analogous, data-dependent choices should be investigated for other underlying density level set estimators.
- Finally, the algorithm does not necessarily need to stop once it leaves the loop. Instead, it could save the found clusters together with the level and reenter the loop recursively for both clusters. This way it seems plausible, that the algorithm is actually able to recover all clusters contained in the cluster tree.

Besides these issues that are more of practical interest, it would also be helpful to investigate the following, more theoretically orientated questions:

- Can we replace the thickness assumption by some other assumption such as Hölder continuity or rectifiable cluster boundaries, which have already been considered in the literature. Or, more challenging, is it possible in dimension  $d \geq 2$  to remove such assumptions at all?
- Can we replace density level sets by level sets of generalized densities in the sense of Rinaldo and Wasserman (2010)?

## Acknowledgments

I like to thank Michael Eisermann for many fruitful discussions.

## References

- J.W. Carmichael, G.A. George, and R.S. Julius. Finding natural clusters. *Systematic Zoology*, 17:144–150, 1968.
- K. Chaudhuri and S. Dasgupta. Rates of convergence for the cluster tree. In J. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 343–351. 2010.
- A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *Ann. Statist.*, 25: 2300–2312, 1997.
- A. Cuevas, M. Febrero, and R. Fraiman. Estimating the number of clusters. *The Canadian Journal of Statistics*, 28:367–382, 2000.
- L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.
- J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York, 1975.
- J.A. Hartigan. Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.*, 76:388–394, 1981.
- M. Maier, M. Hein, and U. von Luxburg. Generalized density clustering. *Optimal construction of k-nearest neighbor graphs for identifying noisy clusters*, 410:1749–1764, 2009.
- P. Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.*, 8:1369–1392, 2007.
- A. Rinaldo and L. Wasserman. Generalized density clustering. *Ann. Statist.*, 38:2678–2722, 2010.
- W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*, 20:25–47, 2003.
- W. Stuetzle and R. Nugent. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19:397–418, 2010.

## Appendix A. Proofs related to the definition of level sets

**Proof of Lemma 1** The second inclusion has already been shown in (2), and hence it suffices to show the first. To show the first inclusion we fix an  $x \in \{h \geq \rho\}$  and an open set  $U$  with  $x \in U$ . Then  $\{h \geq \rho\} \cap U$  is open and non-empty, and hence  $\text{supp } \mu = X$  yields

$$\mu_\rho(U) = \mu(\{h \geq \rho\} \cap U) \geq \mu(\{h \geq \rho\} \cap U) > 0.$$

By (3) we conclude that  $x \in M_\rho$ , that is we have shown  $\{h \geq \rho\} \subset M_\rho$ . Since  $M_\rho$  is closed, we then obtain the first inclusion. ■

**Proof of Corollary 2** Clearly, we have  $\{h > \rho\} \subset \{h \geq \rho\}$  and since  $\{h > \rho\}$  is open, we conclude that  $\{h > \rho\} \subset \{h \geq \rho\} \subset M_\rho$  by Lemma 1. This implies the first and, since  $\{h > \rho\}$  is open, also the third inclusion. The second and forth inclusion also follows from Lemma 1 and the fact that  $\{h \geq \rho\}$  is closed. ■

**Proof of Lemma 3** Obviously, it suffice to show the first inclusion. To this end, we fix an  $x \in M_{\rho_2}$  and an open set  $U \subset X$  with  $x \in U$ . Moreover, we fix a  $\mu$ -density  $h$  of  $P$ . Then we obtain

$$\mu_{\rho_1}(U) = \mu(\{h \geq \rho_1\} \cap U) \geq \mu(\{h \geq \rho_2\} \cap U) = \mu_{\rho_2}(U) > 0,$$

and hence we obtain  $x \in M_{\rho_1}$  by (3). ■

**Proof of Lemma 5** *i)*  $\Rightarrow$  *iii)*. Let  $\tilde{h} : X \rightarrow [0, \infty)$  be an arbitrary  $\mu$ -density of  $P$ . We define

$$h(x) := \begin{cases} \tilde{h}(x) & \text{if } x \notin M_\rho \Delta \{\tilde{h} \geq \rho\} \\ \rho & \text{if } x \in M_\rho \setminus \{\tilde{h} \geq \rho\} \\ 0 & \text{if } x \in \{\tilde{h} \geq \rho\} \setminus M_\rho. \end{cases}$$

Since  $\{h \neq \tilde{h}\} \subset M_\rho \Delta \{\tilde{h} \geq \rho\}$ , the regularity shows that  $\mu(\{h \neq \tilde{h}\}) = 0$ , and hence  $h$  is a  $\mu$ -density of  $P$ . Furthermore, for  $x \in \{h \geq \rho\}$ , we either have  $x \in M_\rho \setminus \{\tilde{h} \geq \rho\} \subset M_\rho$  or

$$x \in X \setminus (M_\rho \Delta \{\tilde{h} \geq \rho\}) \cap \{\tilde{h} \geq \rho\} \subset M_\rho,$$

where in the last step we used that  $x \notin A \Delta B$  together with  $x \in B$  implies  $x \in A$ . Conversely, if  $x \in M_\rho$ , then  $\tilde{h}(x) < \rho$  implies  $h(x) = \rho$  and  $\tilde{h}(x) \geq \rho$  implies  $h(x) = \tilde{h}(x) \geq \rho$ . These considerations show  $\{h \geq \rho\} = M_\rho$ .

*iii)*  $\Rightarrow$  *ii)*. Since  $M_\rho$  is closed, this implication is trivial.

*ii)*  $\Rightarrow$  *i)*. The inclusion (2) shows  $M_\rho \subset \overline{\{h \geq \rho\}} = \{h \geq \rho\}$ , and hence we obtain  $\mu(M_\rho \setminus \{h \geq \rho\}) = \mu(\emptyset) = 0$ .

Finally, if  $h$  is an upper semi-continuous  $\mu$ -density, then  $\{h \geq \rho\}$  is closed for all  $\rho \in \mathbb{R}$ . Consequently,  $P$  is regular at every level by the already proved implication from *ii)* to *i)*. ■

**Proof of Lemma 7** The inclusion  $\subset$  is trivial. To show the converse, we fix an  $\rho > \rho^*$ . Then there exists an  $\rho' \in (\rho^*, \rho)$ , and by Corollary 2 we thus find

$$M_\rho \subset \{h \geq \rho\} \subset \{h > \rho'\} \subset \overset{\circ}{M}_{\rho'}.$$

From this we easily derive the assertion. ■

**Proof of Lemma 8** The regularity follows from (4). To show that that  $P$  is normal, we fix a  $\rho_0 \geq \rho^*$ . Because of the monotonicity of  $M_\rho$  in  $\rho$ , it then suffices to show that

$\mu(M_\rho \setminus \overset{\circ}{M}_\rho) = 0$  for all  $\rho > \rho_0$ . However, Lemma 1 ensures both  $\{h \geq \rho\} \subset \overset{\circ}{M}_\rho$  and  $M_\rho \subset \{h \geq \rho\}$ , and hence we obtain  $M_\rho \setminus \overset{\circ}{M}_\rho \subset \partial\{h \geq \rho\}$ .  $\blacksquare$

## Appendix B. Proofs related to basic properties of connected components

**Lemma 27** *Let  $(X, d)$  be a compact metric space and  $A, B \subset X$  be two subsets. Then the following statements hold:*

- i) *If  $A$  is closed, then  $T_\delta(A) := \{x \in X : \exists x' \in A \text{ with } d(x, x') \leq \delta\}$ .*
- ii) *We have  $d(A, B) \leq d(T_\delta(A), T_\delta(B)) + 2\delta$ .*
- iii) *We have*

$$\bigcap_{\delta > 0} T_\delta(A) = \overline{A}. \quad (12)$$

**Proof of Lemma 27** i). For fixed  $x \in T_\delta(A)$ , there exists a sequence  $(x_n) \subset A$  with  $d(x, x_n) \leq \delta + 1/n$  for all  $n \geq 1$ . Since  $X$  is compact, we may assume without loss of generality that  $(x_n)$  converges to some  $x' \in X$ , and since we assumed that  $A$  is closed, we obtain  $x' \in A$ . Now we easily obtain the assertion from  $d(x, x') \leq d(x, x_n) + d(x_n, x')$ .

ii). Let us fix an  $x \in T_\delta(A)$  and an  $y \in T_\delta(B)$ . Then there exist two sequences  $(x_n) \subset A$  and  $(y_n) \subset B$  such that  $d(x, x_n) \leq \delta + 1/n$  and  $d(y, y_n) \leq \delta + 1/n$  for all  $n \geq 1$ . Now this construction yields

$$d(A, B) \leq d(x_n, y_n) \leq d(x_n, x) + d(x, y) + d(y, y_n) \leq d(x, y) + 2\delta + 2/n \quad n \geq 1,$$

and by first letting  $n \rightarrow \infty$  and then taking the infimum over all  $x \in T_\delta(A)$  and  $y \in T_\delta(B)$ , we obtain the assertion.

iii). To show the inclusion  $\supset$ , we fix an  $x \in \overline{A}$ . Then there exists a sequence  $(x_n) \subset A$  with  $x_n \rightarrow x$  for  $n \rightarrow \infty$ . For  $\delta > 0$  there then exists an  $n_\delta$  such that  $d(x, x_n) \leq \delta$  for all  $n \geq n_\delta$ . This shows  $x \in T_\delta(A)$ . To show the converse inclusion  $\subset$ , we fix an  $x \in X$  that satisfies  $x \in T_{1/n}(A)$  for all  $n \geq 1$ . Then there exists a sequence  $(x_n) \subset A$  with  $d(x, x_n) \leq 1/n$ , and hence we find  $x_n \rightarrow x$  for  $n \rightarrow \infty$ . This shows  $x \in \overline{A}$ .  $\blacksquare$

**Lemma 28** *Let  $(X, d)$  be a compact metric space,  $A \subset X$  be a non-empty subset and  $\tau > 0$ . Then every  $\tau$ -connected component of  $A$  is  $\tau$ -connected.*

**Proof of Lemma 28** Let  $A'$  be a  $\tau$ -connected component of  $A$  and  $x, x' \in A'$ . Then  $x$  and  $x'$  are  $\tau$ -connected in  $A$ , and hence there exist  $x_1, \dots, x_n \in A$  such that  $x_1 = x$ ,  $x_n = x'$  and  $d(x_i, x_{i+1}) < \tau$  for all  $i = 1, \dots, n - 1$ . Now,  $d(x_1, x_2) < \tau$  shows that  $x_1$  and  $x_2$  are  $\tau$ -connected in  $A$ , and hence they belong to the same  $\tau$ -connected component, i.e. we have found  $x_2 \in A'$ . Iterating this argument, we find  $x_i \in A'$  for all  $i = 1, \dots, n$ . Consequently,  $x$  and  $x'$  are not only  $\tau$ -connected in  $A$ , but also  $\tau$ -connected in  $A'$ . This shows that  $A'$  is  $\tau$ -connected.  $\blacksquare$

**Lemma 29** Let  $(X, d)$  be a compact metric space,  $A \subset X$  be a non-empty subset and  $\tau > 0$ . Then there exist only finitely many  $\tau$ -connected components  $A_1, \dots, A_m$  of  $A$ . Moreover, we have  $d(A_i, A_j) \geq \tau$  for all  $i \neq j$ . Finally, if  $A$  is closed, these components are closed, too.

**Proof of Lemma 29** Let  $A' \neq A''$  be two  $\tau$ -connected components of  $A$ . Then we have  $d(x', x'') \geq \tau$  for all  $x' \in A'$  and  $x'' \in A''$ , since otherwise  $x'$  and  $x''$  would be  $\tau$ -connected in  $A$ . Consequently, we have  $d(A', A'') \geq \tau$ , and from the latter and the compactness of  $X$ , it is straightforward to conclude that  $|\mathcal{C}_\tau(A)| < \infty$ . Finally, let  $(x_i) \subset A'$  be a sequence in some component  $A' \in \mathcal{C}_\tau(A)$  such that  $x_i \rightarrow x$  for some  $x \in X$ . Since  $A$  is closed, we have  $x \in A$ , and hence  $x \in A''$  for  $A'' \in \mathcal{C}_\tau(A)$ . By construction we find  $d(A', A'') = 0$ , and hence we obtain  $A' = A''$  by the assertion that has been shown first.  $\blacksquare$

**Lemma 30** Let  $(X, d)$  be a compact metric space,  $A \subset X$  be a non-empty subset and  $\tau > 0$ . Then the following statements are equivalent:

- i)  $A$  is  $\tau$ -connected.
- ii) For all non-empty subsets  $A^+$  and  $A^-$  of  $A$  with  $A^+ \cup A^- = A$  and  $A^+ \cap A^- = \emptyset$  we have  $d(A^+, A^-) < \tau$ .

**Proof of Lemma 30** i)  $\Rightarrow$  ii). Let us fix two non-empty subsets  $A^+$  and  $A^-$  of  $A$  with  $A^+ \cup A^- = A$  and  $A^+ \cap A^- = \emptyset$ . Let us further fix two points  $x^+ \in A^+$  and  $x^- \in A^-$ . Since  $A$  is  $\tau$ -connected, there then exist  $x_1, \dots, x_n \in A$  such that  $x_1 = x^-$ ,  $x_n = x^+$  and  $d(x_i, x_{i+1}) < \tau$  for all  $i = 1, \dots, n-1$ . Then,  $x^+ \in A^+$  and  $x^- \in A^-$  imply the existence of an  $i \in \{1, \dots, n-1\}$  with  $x_i \in A^-$  and  $x_{i+1} \in A^+$ . This yields  $d(A^+, A^-) \leq d(x_i, x_{i+1}) < \tau$ .

ii)  $\Rightarrow$  i). Assume that  $A$  is not  $\tau$ -connected. Then Lemma 29 shows that there exist finitely many  $\tau$ -connected components  $A_1, \dots, A_m$  of  $A$ . By definition, these components are non-empty, mutually disjoint, and satisfy  $A = A_1 \cup \dots \cup A_m$ . Moreover, Lemma 28 shows that each component is  $\tau$ -connected, and since we assumed that  $A$  itself is not  $\tau$ -connected, we conclude that  $m \geq 2$ . In addition, Lemma 29 shows  $d(A_j, A_{j'}) \geq \tau$ , whenever  $j \neq j'$ . Let us define  $A^- := A_1$  and  $A^+ := A_2 \cup \dots \cup A_m$ . Then our previous considerations show that the subsets  $A^+$  and  $A^-$  of  $A$  are non-empty and satisfy  $A^+ \cup A^- = A$ ,  $A^+ \cap A^- = \emptyset$ , and  $d(A^+, A^-) \geq \tau$ .  $\blacksquare$

**Corollary 31** Let  $(X, d)$  be a compact metric space,  $A \subset B \subset X$  be non-empty subsets and  $\tau > 0$ . If  $A$  is  $\tau$ -connected, then there exists exactly one  $\tau$ -connected component  $B'$  of  $B$  with  $A \cap B' \neq \emptyset$ . Moreover,  $B'$  is the only  $\tau$ -connected component  $B''$  of  $B$  that satisfies  $A \subset B''$ .

**Proof of Corollary 31** The second assertion is a direct consequence of the first, and hence it suffice to show the first assertion. Now, by Lemma 29, there exist finitely many  $\tau$ -connected components  $B_1, \dots, B_m$  of  $B$ . Since we obviously have  $A \subset B_1 \cup \dots \cup B_m$  it suffices to show  $A \cap B_i = \emptyset$  for all but one index  $i \in \{1, \dots, m\}$ . Let us assume the converse,

that is, there exist two indices  $i, j \in \{1, \dots, m\}$  with  $i \neq j$ ,  $A \cap B_i \neq \emptyset$ , and  $A \cap B_j \neq \emptyset$ . We write  $A^- := A \cap B_i$  and  $A^+ := A \cap (B \setminus B_i)$ . Since  $B_j \subset B \setminus B_i$ , we obtain  $A^+ \neq \emptyset$ , and therefore, Lemma 30 shows  $d(A^-, A^+) < \tau$ . Consequently, there exist  $x^- \in A^-$  and  $x^+ \in A^+$  with  $d(x^+, x^-) < \tau$ . Now we obviously have  $x^- \in B_i$ , and by construction, we also find an index  $i' \neq i$  with  $x^+ \in B_{i'}$ . Our previous inequality then yields  $d(B_i, B_{i'}) < \tau$ , while Lemma 29 shows  $d(B_i, B_{i'}) \geq \tau$ , that is, we have found a contradiction. ■

**Lemma 32** *Let  $(X, d)$  be a compact metric space,  $A \subset X$  be a non-empty subset and  $\tau > 0$ . Then, for a partition  $A_1, \dots, A_m$  of  $A$ , the following statements are equivalent:*

- i)  $\mathcal{C}_\tau(A) = \{A_1, \dots, A_m\}$ .
- ii) For all  $i = 1, \dots, m$ , the set  $A_i$  is  $\tau$ -connected and  $d(A_i, A_j) \geq \tau$  for all  $i \neq j$ .

**Proof of Lemma 32** i)  $\Rightarrow$  ii). Follows from Lemma 29.

ii)  $\Rightarrow$  i). Let us fix an  $A' \in \mathcal{C}_\tau(A)$ . Then, by Corollary 31, every  $A_i$  with  $A_i \cap A' \neq \emptyset$  satisfies  $A_i \subset A'$ . Since  $A_1, \dots, A_m$  is a partition, we conclude that

$$A' = \bigcup_{i \in I} A_i,$$

where  $I := \{i : A_i \cap A' \neq \emptyset\}$ . Now let us assume that  $|I| \geq 2$ . We fix an  $i_0 \in I$  and write  $A^+ := A_{i_0}$  and  $A^- := \bigcup_{i \in I \setminus \{i_0\}} A_i$ . Since  $|I| \geq 2$ , we obtain  $A^- \neq \emptyset$ , and hence Lemma 30 shows  $d(A^+, A^-) < \tau$ . On the other hand, our assumption ensures  $d(A^+, A^-) \geq \tau$ , and hence  $|I| \geq 2$  cannot be true. Consequently, there exists a unique index  $i$  with  $A' = A_i$ , that is, we have shown the assertion. ■

**Proof of Lemma 10** For  $A' \in \mathcal{C}_\tau(A)$ , Corollary 31 shows that there exists exactly  $B' \in \mathcal{C}_\tau(B)$  with  $A' \subset B'$ . Setting  $\zeta(A') := B'$  then gives the desired map and this map is uniquely determined since  $B'$  is. ■

**Proof of Lemma 11** Clearly,  $\zeta_{B,C} \circ \zeta_{A,B}$  maps from  $\mathcal{C}_\tau(A)$  to  $\mathcal{C}_\tau(C)$ . Moreover, for  $A' \in \mathcal{C}_\tau(A)$  we have  $A' \subset \zeta_{A,B}(A')$  and for  $B' := \zeta_{A,B}(A') \in \mathcal{C}_\tau(B)$  we have  $B' \subset \zeta_{B,C}(B')$ . Combining these inclusions we find  $A' \subset \zeta_{B,C}(\zeta_{A,B}(A')) = \zeta_{B,C} \circ \zeta_{A,B}(A')$  for all  $A' \in \mathcal{C}_\tau(A)$ . By Lemma 10,  $\zeta_{A,C}$  is the only map satisfying this property, and hence we conclude that  $\zeta_{A,C} = \zeta_{B,C} \circ \zeta_{A,B}$ . ■

**Lemma 33** *Let  $(X, d)$  be a compact metric space and  $A \subset X$  be a non-empty closed subset. Then the following statements are equivalent:*

- i)  $A$  is connected.
- ii)  $A$  is  $\tau$ -connected for all  $\tau > 0$ .

**Proof of Lemma 33** *i)  $\Rightarrow$  ii).* Let us assume that  $A$  is not  $\tau$ -connected for some  $\tau > 0$ . Then, by Lemma 29, there are finitely many  $\tau$ -connected components  $A_1, \dots, A_m$  of  $A$  with  $m > 1$ . We write  $A' := A_1$  and  $A'' := A_2 \cup \dots \cup A_m$ . Then  $A'$  and  $A''$  are non-empty, disjoint and  $A' \cup A'' = A$  by construction. Moreover, Lemma 29 shows that  $A'$  and  $A''$  are closed since  $A$  is closed, and hence  $A$  cannot be connected.

*ii)  $\Rightarrow$  i).* Let us assume that  $A$  is not connected. Then there exist two non-empty closed disjoint subsets of  $A$  with  $A' \cup A'' = A$ . Since  $X$  is compact,  $A'$  and  $A''$  are also compact, and hence  $A' \cap A'' = \emptyset$  implies  $\tau := d(A', A'') > 0$ . Lemma 30 then shows that  $A$  is not  $\tau$ -connected. ■

**Proof of Lemma 12** *i).* Let  $A' \subset A$  be a topologically connected component of  $A$  and  $\tau > 0$ . Then we have already seen in Lemma 33 that  $A'$  is  $\tau$ -connected, and since  $A' \subset A$ , Corollary 31 shows that there exists exactly one  $A'' \in \mathcal{C}_\tau(A)$  with  $A' \subset A''$ . Consequently,  $\zeta(A') := A''$  is the only possible definition of  $\zeta$ . Moreover, if  $A'' \in \mathcal{C}_\tau(A)$  is an arbitrary  $\tau$ -connected component, then there exists an  $x \in A''$ , and to this  $x$ , there exists an  $A' \in \mathcal{C}(A)$  with  $x \in A'$ . Corollary 31 then shows that  $A' \subset A''$ , and hence we obtain  $\zeta(A') = A''$ . In other words,  $\zeta$  is surjective.

*ii).* Let  $A_1, \dots, A_m$  be the topologically connected components of  $A$ . Then the components are closed, and since  $A$  is a closed and thus compact subset of  $X$ , the components are compact, too. This shows  $d(A_i, A_j) > 0$  for all  $i \neq j$ , and consequently we obtain  $\tau_A^* > 0$ . Let us fix a  $\tau \in (0, \tau_A^*] \cap (0, \infty)$ . Then, Lemma 33 shows that each  $A_i$  is  $\tau$ -connected, and therefore Lemma 32 together with  $d(A_i, A_j) \geq \tau_A^* \geq \tau$  for all  $i \neq j$  yields  $\mathcal{C}_\tau(A) = \{A_1, \dots, A_m\}$ . Consequently, we have proved  $\mathcal{C}(A) = \mathcal{C}_\tau(A)$ . The bijectivity of  $\zeta$  now follows from its surjectivity. For the proof of the last equation, we define  $\tau^* := \sup\{\tau > 0 : \mathcal{C}(A) = \mathcal{C}_\tau(A)\}$ . Then we have already seen that  $\tau_A^* \leq \tau^*$ . Now suppose that  $\tau_A^* < \tau^*$ . Then there exists a  $\tau \in (\tau_A^*, \tau^*)$  with  $\mathcal{C}(A) = \mathcal{C}_\tau(A)$ . On the one hand, we then find  $d(A_i, A_j) \geq \tau$  for all  $i \neq j$  by Lemma 29, while on the other hand  $\tau > \tau_A^*$  shows that there exist  $i_0 \neq j_0$  with  $d(A_{i_0}, A_{j_0}) < \tau$ . In other words, the assumption  $\tau_A^* < \tau^*$  leads to a contradiction, and hence we have  $\tau_A^* = \tau^*$ . ■

**Proof of Lemma 14** Let  $A', A'' \in \mathcal{C}(A)$  with  $A' \neq A''$ . Since  $\zeta$  is injective, we then obtain  $\zeta(A') \neq \zeta(A'')$ . Combining this with  $A' \subset \zeta(A')$  and  $A'' \subset \zeta(A'')$ , we find

$$d(A', A'') \geq d(\zeta(A'), \zeta(A'')) \geq \tau_B^*,$$

where the last inequality follows from Lemma 12. Taking the infimum over all  $A'$  and  $A''$  with  $A' \neq A''$  yields the assertion. ■

**Proof of Lemma 15** *i).* Since  $\tau > \delta$ , there exist an  $\varepsilon > 0$  with  $\delta + \varepsilon < \tau$ . For  $x \in T_\delta(A')$ , there thus exists an  $x' \in A'$  with  $d(x, x') \leq \delta + \varepsilon < \tau$ , i.e.  $x$  and  $x'$  are  $\tau$ -connected. Since  $A'$  is  $\tau$ -connected, it is then easy to show that every pair  $x, x'' \in T_\delta(A')$  is  $\tau$ -connected.

*ii).* Let us fix an  $A' \in \mathcal{C}_\tau(T_\delta(A))$  and an  $x \in A'$ . For  $n \geq 1$  there then exists an  $x_n \in A$  with  $d(x, x_n) \leq \delta + 1/n$  and since by Lemma 29 there only exist finitely many  $\tau$ -connected components of  $A$ , we may assume without loss of generality that there exists

an  $A'' \in \mathcal{C}_\tau(A)$  with  $x_n \in A''$  for all  $n \geq 1$ . This yields  $d(x, A'') \leq \delta + 1/n$  for all  $n \geq 1$ , and hence  $d(x, A'') \leq \delta$ . Consequently, we obtain  $x \in T_\delta(A'')$ , i.e. we have  $T_\delta(A'') \cap A' \neq \emptyset$ . Since  $T_\delta(A'') \subset T_\delta(A)$ , we then conclude that  $T_\delta(A'') \subset A'$  by Corollary 31 and part *i*). Furthermore, we clearly have  $A'' \subset T_\delta(A'')$ , and hence  $\zeta(A'') = A'$ .

*iii).* We write  $A_1, \dots, A_m$  for the  $\tau$ -connected components of  $A$ . For arbitrary  $\tau > 0$  and  $\delta > 0$ , we first show that

$$T_\delta(A) = \bigcup_{i=1}^m T_\delta(A_i). \quad (13)$$

Obviously, the inclusion „ $\supset$ “ is trivial. To show the converse inclusion, we fix an  $x \in T_\delta(A)$ . Since  $A$  is compact, there then exists an  $x' \in A$  with  $d(x, x') \leq \delta$ . Obviously, we further have  $x' \in A_i$  for some component  $A_i$ , and hence we find  $x \in T_\delta(A_i)$ .

We now choose  $\tau^*$  and  $\delta^*$  by (8) and fix some  $\tau \in (0, \tau^*]$  and  $\delta \in (0, \delta^*]$ . Moreover, let  $A_1, \dots, A_m$  again be the  $\tau$ -connected components of  $A$ . Since  $\tau \leq \tau^* \leq \tau_A^*$ , part *ii*) of Lemma 12 shows  $\mathcal{C}(A) = \mathcal{C}_\tau(A)$ , and consequently we obtain  $d(A_i, A_j) \geq \tau_A^* = 3\tau^*$  for all  $i \neq j$  by another application of part *ii*) of Lemma 12. Our next goal is to show that

$$d(T_\delta(A_i), T_\delta(A_j)) \geq \tau, \quad i \neq j. \quad (14)$$

To this end, we fix an  $x_i \in T_\delta(A_i)$  and an  $x_j \in T_\delta(A_j)$ . Then there exist  $x'_i \in A_i$  and  $x'_j \in A_j$  with  $d(x_i, x'_i) \leq \delta$  and  $d(x_j, x'_j) \leq \delta$ , and therefore using  $\delta \leq \delta^* = \tau^*$  we obtain

$$3\tau^* \leq d(x'_i, x'_j) \leq d(x'_i, x_i) + d(x_i, x_j) + d(x_j, x'_j) \leq 2\tau^* + d(x_i, x_j).$$

Obviously, the latter together with  $\tau^* \geq \tau$  implies (14).

Now part *i*) showed that each  $T_\delta(A_i)$ ,  $i = 1, \dots, m$ , is  $\tau$ -connected whenever  $\tau > \delta$ . Combining this with (13), (14), and Lemma 32, we thus see that  $T_\delta(A_1), \dots, T_\delta(A_m)$  are the  $\tau$ -connected components of  $T_\delta(A)$ . The bijectivity of  $\zeta$  then follows from the surjectivity and a simple cardinality argument. ■

## Appendix C. Proofs related to the identification of components

**Proof of Theorem 18** *i).* Since  $\tau > \delta$ , part *ii*) of Lemma 15 and part *i*) of Lemma 12 yield

$$|\mathcal{C}_\tau(M_{\rho, \delta})| \leq |\mathcal{C}_\tau(M_\rho)| \leq |\mathcal{C}(M_\rho)| \leq 2. \quad (15)$$

*ii).* Let us fix a  $\rho \in [\rho^* + \varepsilon^*, \rho^{**}]$ . Then Definition 16 guarantees that both  $M_{\rho^* + \varepsilon^*}$  and  $M_\rho$  have two topologically connected components and that the top-CCRM  $\zeta : \mathcal{C}(M_\rho) \rightarrow \mathcal{C}(M_{\rho^* + \varepsilon^*})$  is bijective. From Lemma 14 we thus obtain  $\tau_{M_\rho}^* \geq \tau_{M_{\rho^* + \varepsilon^*}}^*$ , and consequently, we find  $\tau \leq \tau_{M_{\rho^* + \varepsilon^*}}^* \leq \tau_{M_\rho}^*$ . This implies  $\mathcal{C}(M_\rho) = \mathcal{C}_\tau(M_\rho)$  by part *ii*) of Lemma 12, that is,  $|\mathcal{C}_\tau(M_\rho)| = 2$ . Furthermore,  $\tau_{M_\rho}^* \geq \tau_{M_{\rho^* + \varepsilon^*}}^*$  implies  $\delta \leq \tau_{M_\rho}^*/3$  and  $\tau \leq \tau_{M_\rho}^*/3$ , and hence part *iii*) of Lemma 15 shows that the  $\tau$ -CCRM  $\zeta : \mathcal{C}_\tau(M_\rho) \rightarrow \mathcal{C}_\tau(M_{\rho, \delta})$  is bijective. This implies  $|\mathcal{C}_\tau(M_{\rho, \delta})| = 2$ .

*iii).* If  $|\mathcal{C}_\tau(M_{\rho, \delta})| > 1$ , then (15) implies  $|\mathcal{C}(M_\rho)| > 1$ , and hence Definition 16 yields  $\rho \geq \rho^*$ . Moreover, for  $\rho^{**}$ , we have already seen in part *ii*) that the  $\tau$ -CCRM  $\zeta_M :$

$\mathcal{C}_\tau(M_{\rho^{**}}) \rightarrow \mathcal{C}_\tau(M_{\rho^{**},\delta})$  is bijective, and the proof of *ii*) further showed  $\mathcal{C}(M_{\rho^{**}}) = \mathcal{C}_\tau(M_{\rho^{**}})$ . Consequently, we can identify  $\zeta_M$  with the CCRM  $\mathcal{C}(M_{\rho^{**}}) \rightarrow \mathcal{C}_\tau(M_{\rho^{**},\delta})$ . Moreover, by (15) we obtain  $|\mathcal{C}(M_\rho)| = 2$ , and hence Definition 16 ensures that the top-CCRM  $\zeta^{**} : \mathcal{C}(M_{\rho^{**}}) \rightarrow \mathcal{C}(M_\rho)$  is bijective. In addition,  $\tau > \delta$  together with part *ii*) of Lemma 15 and part *i*) of Lemma 12 shows that the CCRM  $\zeta_\rho : \mathcal{C}(M_\rho) \rightarrow \mathcal{C}_\tau(M_{\rho,\delta})$  is surjective. Now, by Lemma 11 these maps commute in the sense of the following diagram

$$\begin{array}{ccc} \mathcal{C}(M_{\rho^{**}}) & \xrightarrow{\zeta^{**}} & \mathcal{C}(M_\rho) \\ \zeta_M \downarrow & & \downarrow \zeta_\rho \\ \mathcal{C}_\tau(M_{\rho^{**},\delta}) & \xrightarrow{\zeta} & \mathcal{C}_\tau(M_{\rho,\delta}) \end{array}$$

and consequently,  $\zeta$  is surjective. Since  $|\mathcal{C}_\tau(M_{\rho^{**},\delta})| = |\mathcal{C}(M_{\rho^{**}})| = 2$  and  $|\mathcal{C}_\tau(M_{\rho,\delta})| = 2$ , we then conclude that  $\zeta$  is bijective.

*iv).* Let us fix an  $\rho \in [\rho^* + \varepsilon^*, \rho^{**}]$ . By part *ii*) and *i*) we then see that  $M_{\rho,\delta}$  has two  $\tau$ -connected components and part *iii*) thus shows that the  $\tau$ -CCRM  $\zeta_M : \mathcal{C}_\tau(M_{\rho^{**},\delta}) \rightarrow \mathcal{C}_\tau(M_{\rho,\delta})$  is bijective. Moreover, Lemma 11 yields the following diagram

$$\begin{array}{ccc} \mathcal{C}_\tau(V_{\rho^{**},\delta}) & \xrightarrow{\zeta^{**}} & \mathcal{C}_\tau(M_{\rho^{**},\delta}) \\ \zeta_V \downarrow & & \downarrow \zeta_M \\ \mathcal{C}_\tau(V_{\rho,\delta}) & \xrightarrow{\zeta_{V,M}} & \mathcal{C}_\tau(M_{\rho,\delta}) \end{array}$$

where  $\zeta_V$  and  $\zeta_{V,M}$  are the corresponding  $\tau$ -CCRMs. Now our assumption guarantees that  $\zeta^{**}$  is bijective, and hence the diagram shows that  $\zeta_{V,M} \circ \zeta_V$  is bijective. Consequently,  $\zeta_V$  is injective, and we obtain  $2 = |\mathcal{C}_\tau(M_{\rho,\delta})| = |\mathcal{C}_\tau(V_{\rho^{**},\delta})| \leq |\mathcal{C}_\tau(V_{\rho,\delta})|$ . ■

**Lemma 34** *Let  $(X, d)$  be a compact metric space,  $\mu$  be a finite measure on  $X$ , and  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$  that has two thick clusters of order  $\gamma \in (0, 1]$  between the critical levels  $\rho^*$  and  $\rho^{**}$ . We write  $\psi$  for the corresponding thickness function. Then, for all  $\rho \in [0, \rho^{**}]$ ,  $\delta \in (0, \tilde{\delta}_0]$ , and  $\tau > \psi(\delta)$ , the following statements hold:*

- i) For all  $B' \in \mathcal{C}(M_\rho)$ , there exists at most one  $A' \in \mathcal{C}_\tau(V_{\rho,\delta})$  such that  $A' \cap B' \neq \emptyset$ .*
- ii) We have  $|\mathcal{C}_\tau(V_{\rho,\delta})| \leq |\mathcal{C}(M_\rho)|$ .*
- iii) If  $|\mathcal{C}_\tau(V_{\rho,\delta})| = |\mathcal{C}(M_\rho)|$ , then there exists a unique map  $\zeta : \mathcal{C}_\tau(V_{\rho,\delta}) \rightarrow \mathcal{C}(M_\rho)$  that satisfies*

$$A' \subset \zeta(A') , \quad A' \in \mathcal{C}_\tau(V_{\rho,\delta}) . \quad (16)$$

*Moreover,  $\zeta$  is bijective.*

**Proof of Lemma 34** *i).* Let us fix a  $\tau' \in (0, \tau_{M_\rho}^*]$  such that  $\psi(\delta) + \tau' < \tau$ , where  $\tau_{M_\rho}^*$  is the constant defined in Lemma 12 and  $c \geq 1$  is the constant appearing in Definition

19. Moreover, we fix a  $B' \in \mathcal{C}(M_\rho)$ . By Lemma 12 we then see that  $\mathcal{C}(M_\rho) = \mathcal{C}_{\tau'}(M_\rho)$ , and hence  $B'$  is  $\tau'$ -connected. Now let  $A_1, \dots, A_m$  be the  $\tau$ -connected components of  $V_{\rho, \delta}$ . Clearly, Lemma 29 yields  $d(A_i, A_j) \geq \tau$  for all  $i \neq j$ . Assume that the assertion of the lemma is not true, that is, there exist  $i_0 \neq j_0$  with  $A_{i_0} \cap B' \neq \emptyset$  and  $A_{j_0} \cap B' \neq \emptyset$ . Then there exist  $x' \in A_{i_0} \cap B'$  and  $x'' \in A_{j_0} \cap B'$ , and since  $B'$  is  $\tau'$ -connected, there further exist  $x_0, \dots, x_{n+1} \in B' \subset M_\rho$  with  $x_0 = x'$ ,  $x_{n+1} = x''$  and  $d(x_i, x_{i+1}) < \tau'$  for all  $i = 0, \dots, n$ . Moreover, our assumptions guarantee  $d(x_i, V_{\rho, \delta}) < \psi(\delta)/2$  for all  $i = 0, \dots, n+1$ . For all  $i = 0, \dots, n+1$ , there thus exists an index  $\ell_i$  such that

$$d(x_i, A_{\ell_i}) < \psi(\delta)/2.$$

In addition, we have  $x_0 \in A_{i_0}$  and  $x_{n+1} \in A_{j_0}$  by construction, and hence we may choose  $\ell_0 = i_0$  and  $\ell_{n+1} = j_0$ . Since we assumed  $\ell_0 \neq \ell_{n+1}$ , there then exists an  $i \in \{0, \dots, n+1\}$  with  $\ell_i \neq \ell_{i+1}$ . For this index, our construction now yields

$$d(A_{\ell_i}, A_{\ell_{i+1}}) \leq d(x_i, A_{\ell_i}) + d(x_i, x_{i+1}) + d(x_{i+1}, A_{\ell_{i+1}}) < \psi(\delta) + \tau' < \tau,$$

which contradicts the earlier established  $d(A_{\ell_i}, A_{\ell_{i+1}}) \geq \tau$ .

*ii).* Since  $V_{\rho, \delta} \subset M_\rho$ , we have, for every  $A' \in \mathcal{C}_\tau(V_{\rho, \delta})$ , a  $B' \in \mathcal{C}(M_\rho)$  with  $A' \cap B' \neq \emptyset$ . We pick one such  $B'$  and define  $\zeta(A') := B'$ . Now part *i*) shows that  $\zeta : \mathcal{C}_\tau(V_{\rho, \delta}) \rightarrow \mathcal{C}(M_\rho)$  is injective, and hence we conclude  $|\mathcal{C}_\tau(V_{\rho, \delta})| \leq |\mathcal{C}(M_\rho)|$ .

*iii).* As mentioned in part *ii*), we have an injective map  $\zeta : \mathcal{C}_\tau(V_{\rho, \delta}) \rightarrow \mathcal{C}(M_\rho)$  that satisfies

$$A' \cap \zeta(A') \neq \emptyset, \quad A' \in \mathcal{C}_\tau(V_{\rho, \delta}). \quad (17)$$

Now,  $|\mathcal{C}_\tau(V_{\rho, \delta})| = |\mathcal{C}(M_\rho)|$  implies that  $\zeta$  is actually bijective. Let us show that  $\zeta$  is the only map that satisfies (17). To this end, assume the converse, that is, for some  $A' \in \mathcal{C}_\tau(V_{\rho, \delta})$ , there exists an  $B' \in \mathcal{C}(M_\rho)$  with  $B' \neq \zeta(A')$  and  $A' \cap B' \neq \emptyset$ . Since  $\zeta$  is bijective, there then exists an  $A'' \in \mathcal{C}_\tau(V_{\rho, \delta})$  with  $\zeta(A'') = B'$ , and hence we have  $A'' \cap B' \neq \emptyset$ . By part *i*), we conclude that  $A' = A''$ , which in turn yields  $\zeta(A') = \zeta(A'') = B'$ . In other words, we have found a contradiction, and hence  $\zeta$  is indeed the only map that satisfies (17). From this it is easy to conclude, that there exists at most one map that satisfies (16). Let us therefore finally show that  $\zeta$  satisfies (16). To this end, we pick an  $A' \in \mathcal{C}_\tau(V_{\rho, \delta})$  and write  $B_1, \dots, B_m$  for the topologically connected components of  $M_\rho$ . Since  $V_{\rho, \delta} \subset M_\rho$ , we then have  $A' \subset B_1 \cup \dots \cup B_m$ , where the latter union is disjoint. Now, we have just seen that  $\zeta(A') \in \{B_1, \dots, B_m\}$  is the only component satisfying  $A' \cap \zeta(A') \neq \emptyset$ , and therefore we can conclude  $A' \subset \zeta(A')$ . ■

**Proof of Theorem 20** *i).* This follows from  $|\mathcal{C}_\tau(V_{\rho, \delta})| \leq |\mathcal{C}(M_\rho)| \leq 2$ , where the first inequality was established in part *ii*) of Lemma 34.

*ii).* Our definition of  $\varepsilon^*$  yields  $\delta_{\varepsilon^*} = \tau_{\varepsilon^*} = \tau_{M_{\rho^*+\varepsilon^*}}^*/3 \leq \tau_{M_{\rho^{**}}}^*/3$ . By part *iii*) of Lemma 15 we then conclude that the  $\tau$ -CCRM  $\mathcal{C}_\tau(M_{\rho^{**}}) \rightarrow \mathcal{C}_\tau(M_{\rho^{**}, \delta})$  is bijective. By Lemma 11 it thus suffices to show that the  $\tau$ -CCRM  $\zeta : \mathcal{C}_\tau(V_{\rho^{**}, \delta}) \rightarrow \mathcal{C}_\tau(M_{\rho^{**}})$  is bijective. Furthermore, if  $|\mathcal{C}_\tau(V_{\rho^{**}, \delta})| = 1$ , the map  $\zeta$  is automatically injective, and if  $|\mathcal{C}_\tau(V_{\rho^{**}, \delta})| = 2$ , the injectivity follows from the surjectivity. Consequently, it actually suffices to show that  $\zeta$  is surjective. To this end, we fix a  $B' \in \mathcal{C}_\tau(M_{\rho^{**}})$  and an  $x \in B'$ . Then our assumption

ensures  $d(x, V_{\rho^{**}, \delta}) < c\delta^\gamma$ , and hence there exists an  $A' \in \mathcal{C}_\tau(V_{\rho^{**}, \delta})$  with  $d(x, A') < c\delta^\gamma$ . Therefore,  $c\delta^\gamma < \psi(\delta) < \tau$  implies that  $x$  and  $A'$  are  $\tau$ -connected, which yields  $x \in A'$ . In other words, we have shown  $A' \cap B' \neq \emptyset$ . By Lemma 31 and the definition of  $\zeta$ , we conclude that  $\zeta(A') = B'$ .

iii). By part ii) of Lemma 34, we conclude that  $|\mathcal{C}_\tau(V_{\rho, \delta})| = |\mathcal{C}(M_\rho)| = 2$ , and hence the definition of topological clustering ensures  $\rho \geq \rho^*$ . Furthermore, part iii) of Lemma 34 yields a unique map  $\zeta_\rho : \mathcal{C}_\tau(V_{\rho, \delta}) \rightarrow \mathcal{C}(M_\rho)$  satisfying (16). Moreover, part ii) of Theorem 18 shows  $|\mathcal{C}_\tau(M_{\rho^{**}, \delta})| = 2$  and the already established bijectivity of  $\zeta^{**}$  then gives  $|\mathcal{C}_\tau(V_{\rho^{**}, \delta})| = |\mathcal{C}_\tau(M_{\rho^{**}, \delta})| = 2$ . Consequently, part iii) of Lemma 34 yields a unique map  $\zeta_{\rho^{**}} : \mathcal{C}_\tau(V_{\rho^{**}, \delta}) \rightarrow \mathcal{C}(M_{\rho^{**}})$  satisfying (16). Finally, let  $\zeta_{\text{top}} : \mathcal{C}(M_{\rho^{**}}) \rightarrow \mathcal{C}(M_\rho)$  be the top-CCRM, which is bijective according to the definition of topological clustering. Then the  $\tau$ -CCRM  $\zeta : \mathcal{C}_\tau(V_{\rho^{**}, \delta}) \rightarrow \mathcal{C}_\tau(V_{\rho, \delta})$  enjoys the following diagram

$$\begin{array}{ccc} \mathcal{C}_\tau(V_{\rho^{**}, \delta}) & \xrightarrow{\zeta_{\rho^{**}}} & \mathcal{C}(M_{\rho^{**}}) \\ \zeta \downarrow & & \downarrow \zeta_{\text{top}} \\ \mathcal{C}_\tau(V_{\rho, \delta}) & \xrightarrow{\zeta_\rho} & \mathcal{C}(M_\rho) \end{array}$$

whose commutativity can be checked analogously to the proof of Lemma 11. Then the bijectivity of  $\zeta_{\rho^{**}}$ ,  $\zeta_{\text{top}}$ , and  $\zeta_\rho$  yields the bijectivity of  $\zeta$ , which completes the proof. ■

## Appendix D. Proofs related to basic properties of histograms

**Proof of Theorem 22** We fix an  $A \in \mathcal{A}_\delta$  and write  $f := \mu(A)^{-1}\mathbf{1}_A$ . Then  $f$  is non-negative, bounded, and our assumptions ensure  $\|f\|_\infty \leq \kappa_X \delta^{-d_X}$ . Consequently, Hoeffding's inequality, see e.g. (Devroye et al., 1996, Theorem 8.1), yields

$$P^n \left( \left\{ D \in X^n : \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E}_P f \right| < \varepsilon \right\} \right) \geq 1 - 2e^{-2\kappa_X^{-2}\delta^{2d_X}\varepsilon^2 n}$$

for all  $n \geq 1$  and  $\varepsilon > 0$ , where we assumed  $D = (x_1, \dots, x_n)$ . Furthermore, we have  $\frac{1}{n} \sum_{i=1}^n f(x_i) = \mu(A)^{-1}D(A)$  and  $\mathbb{E}_P f = \mu(A)^{-1}P(A)$ . By a union bound argument and  $|\mathcal{A}_\delta| \leq \kappa_X \delta^{-d_X}$ , we thus obtain

$$P^n \left( \left\{ D \in X^n : \sup_{A \in \mathcal{A}_\delta} \left| \frac{D(A)}{\mu(A)} - \frac{P(A)}{\mu(A)} \right| < \varepsilon \right\} \right) \geq 1 - 2\kappa_X \delta^{-d_X} e^{-2\kappa_X^{-2}\delta^{2d_X}\varepsilon^2 n}.$$

For  $A \in \mathcal{A}_\delta$  and  $x \in A$ , we have  $\bar{h}_{D, \mathcal{A}_\delta}(x) = \mu(A)^{-1}D(A)$  and  $\bar{h}_{P, \mathcal{A}_\delta}(x) = \mu(A)^{-1}P(A)$ , and hence find the first assertion.

To show the second inequality, we fix an  $A \in \mathcal{A}_\delta$  and write  $f := \mu(A)^{-1}(\mathbf{1}_A - P(A))$ . This yields  $\|f\|_\infty \leq \kappa_X \delta^{-d_X}$  and

$$\mathbb{E}f^2 \leq \mu(A)^{-2}P(A) \leq \mu(A)^{-1}\|h\|_\infty \leq \kappa_X \delta^{-d_X} \|h\|_\infty.$$

Consequently, Bernstein's inequality, see e.g. (Devroye et al., 1996, Theorem 8.2), yields

$$P^n \left( \left\{ D \in X^n : \left| \frac{1}{n} \sum_{i=1}^n f(x_i) \right| < \varepsilon \right\} \right) \geq 1 - 2e^{-\frac{3\varepsilon^2 \delta^d n}{\kappa_X (6\|h\|_\infty + 2\varepsilon)}}.$$

The rest of the proof follows the lines of the proof of the first inequality.  $\blacksquare$

**Proof of Lemma 23** *i).* We will show the equivalent inclusion  $\{\hat{f}_\rho = -1\} \subset T_\delta(V_{\rho+\varepsilon})$ . To this end, we fix an  $x \in X$  with  $\hat{f}_\rho(x) = -1$ . If  $x \in V_{\rho+\varepsilon}$ , we immediately obtain  $x \in T_\delta(V_{\rho+\varepsilon})$ , and hence we may restrict our considerations to the case  $x \in M_{\rho+\varepsilon}$ . Then,  $\hat{f}_\rho(x) = -1$  implies  $\hat{h}(x) < \rho$  and from  $\|\hat{h} - \bar{h}_{P,\mathcal{A}_\delta}\|_\infty \leq \varepsilon$ , we thus conclude  $\bar{h}_{P,\mathcal{A}_\delta}(x) \leq \hat{h}(x) + \varepsilon < \rho + \varepsilon$ . Now let  $A_i$  be the unique cell of the partition  $\mathcal{A}_\delta$  satisfying  $x \in A_i$ . The definition of  $\bar{h}_{P,\mathcal{A}_\delta}$  together with the assumed  $0 < \mu(A_i) < \infty$  then yields

$$\int_{A_i} h d\mu = P(A_i) < (\rho + \varepsilon)\mu(A_i), \quad (18)$$

where  $h : X \rightarrow [0, \infty)$  is an arbitrary  $\mu$ -density of  $P$ . Our next goal is to show that there exists an  $x' \in V_{\rho+\varepsilon} \cap A_i$ . Suppose the converse, that is  $A_i \subset M_{\rho+\varepsilon}$ . Then the regularity of  $P$  at the level  $\rho + \varepsilon$  yields  $\mu(A_i \setminus \{h \geq \rho + \varepsilon\}) \leq \mu(M_{\rho+\varepsilon} \setminus \{h \geq \rho + \varepsilon\}) = 0$ , and hence we conclude that  $\mu(A_i \cap \{h \geq \rho + \varepsilon\}) = \mu(A_i)$ . This leads to

$$\int_{A_i} h d\mu = \int_{A_i \cap \{h \geq \rho + \varepsilon\}} h d\mu + \int_{A_i \setminus \{h \geq \rho + \varepsilon\}} h d\mu = \int_{A_i \cap \{h \geq \rho + \varepsilon\}} h d\mu \geq (\rho + \varepsilon)\mu(A_i).$$

However, this inequality contradicts (18), and hence there does exist an  $x' \in V_{\rho+\varepsilon} \cap A_i$ . This implies

$$d(x, V_{\rho+\varepsilon}) \leq d(x, x') \leq \text{diam } A_i \leq \delta,$$

i.e. we have shown  $x \in T_\delta(V_{\rho+\varepsilon})$ .

*ii).* Let us fix an  $x \in X$  with  $\hat{f}_\rho(x) = 1$ . If  $x \in M_{\rho-\varepsilon}$ , we immediately obtain  $x \in T_\delta(M_{\rho-\varepsilon})$ , and hence it remains to consider the case  $x \in V_{\rho-\varepsilon}$ . Clearly, if  $\rho - \varepsilon \leq 0$ , this case is impossible, and hence we may additionally assume  $\rho - \varepsilon > 0$ . Then,  $\hat{f}_\rho(x) = 1$  implies  $\hat{h}(x) \geq \rho$  and from  $\|\hat{h} - \bar{h}_{P,\mathcal{A}_\delta}\|_\infty \leq \varepsilon$ , we thus conclude  $\bar{h}_{P,\mathcal{A}_\delta}(x) \geq \hat{h}(x) - \varepsilon \geq \rho - \varepsilon$ . Now let  $A_i$  be the unique cell of the partition  $\mathcal{A}_\delta$  satisfying  $x \in A_i$ . By the definition of  $\bar{h}_{P,\mathcal{A}_\delta}$  and  $\mu(A_i) > 0$  we then obtain

$$\int_{A_i} h d\mu = P(A_i) \geq (\rho - \varepsilon)\mu(A_i), \quad (19)$$

where, again,  $h : X \rightarrow [0, \infty)$  is an arbitrary  $\mu$ -density of  $P$ . Our next goal is to show that there exists an  $x' \in M_{\rho-\varepsilon} \cap A_i$ . Suppose the converse holds, that is  $A_i \subset V_{\rho-\varepsilon}$ . Then the assumed regularity of  $P$  at the level  $\rho - \varepsilon$  yields (6), and hence we conclude that  $\mu(A_i \setminus \{h < \rho - \varepsilon\}) \leq \mu(V_{\rho-\varepsilon} \setminus \{h < \rho - \varepsilon\}) = 0$ . This implies

$$\int_{A_i} h d\mu = \int_{A_i \cap \{h < \rho - \varepsilon\}} h d\mu + \int_{A_i \setminus \{h < \rho - \varepsilon\}} h d\mu = \int_{A_i \cap \{h < \rho - \varepsilon\}} h d\mu < (\rho - \varepsilon)\mu(A_i).$$

However, this inequality contradicts (19), and hence there does exist an  $x' \in M_{\rho-\varepsilon} \cap A_i$ . This yields

$$d(x, M_{\rho-\varepsilon}) \leq d(x, x') \leq \text{diam } A_i \leq \delta,$$

i.e. we have shown  $x \in T_\delta(M_{\rho-\varepsilon})$ . ■

**Lemma 35** *Let  $(X, d)$  be a compact metric space and  $\mu$  be a finite measure on  $X$  such that  $(X, d, \mu)$  admits uniform  $d_X$ -dimensional partitions. Moreover, let  $P$  be a  $\mu$ -absolutely continuous probability measure on  $X$  and  $\hat{h} : X \rightarrow \mathbb{R}$  be a function with  $\|\hat{h} - \bar{h}_{P, A_\delta}\|_\infty \leq \varepsilon$  for some  $\varepsilon > 0$ . Furthermore, for fixed  $\delta > 0$ ,  $\epsilon \geq 0$ ,  $\tau > 0$ , and  $\rho \geq 0$ , let  $\zeta : \mathcal{C}_\tau(V_{\rho+\varepsilon, \delta}) \rightarrow \mathcal{C}_\tau(\{\hat{f}_\rho = 1\})$  be the  $\tau$ -CCRM. Then the following statements hold:*

i) *If, for some  $V' \in \mathcal{C}_\tau(V_{\rho+\varepsilon, \delta})$ , we have  $V' \cap V_{\rho+3\varepsilon+\epsilon, \delta} \neq \emptyset$ , then we obtain*

$$\zeta(V') \cap \{\hat{f}_{\rho+2\varepsilon+\epsilon} = 1\} \neq \emptyset.$$

ii) *For all  $B' \in \mathcal{C}_\tau(\{\hat{f}_\rho = 1\})$  with  $B' \notin \zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon, \delta}))$ , we have*

$$B' \subset T_\delta(X \setminus M_{\rho+\varepsilon}) \cap T_\delta(M_{\rho-\varepsilon}).$$

*Moreover, every  $A \subset B' \cap \{\hat{f}_{\rho+2\varepsilon+\epsilon} = 1\}$  satisfies  $A \subset T_\delta(X \setminus M_{\rho+\varepsilon}) \cap T_\delta(M_{\rho+\varepsilon+\epsilon})$ .*

**Proof of Lemma 35** i). This assertion follows from the  $\tau$ -CCRM property  $V' \subset \zeta(V')$  and the inclusion  $V_{\rho+3\varepsilon+\epsilon, \delta} \subset \{\hat{f}_{\rho+2\varepsilon+\epsilon} = 1\}$  established in Lemma 23.

ii). For  $x \in B'$  we have  $x \notin \bigcup_{V' \in \mathcal{C}_\tau(V_{\rho+\varepsilon, \delta})} \zeta(V')$ , and hence the  $\tau$ -CCRM property yields

$$x \notin \bigcup_{V' \in \mathcal{C}_\tau(V_{\rho+\varepsilon, \delta})} V' = V_{\rho+\varepsilon, \delta}.$$

This shows  $x \in T_\delta(X \setminus M_{\rho+\varepsilon})$ , i.e. we have proved  $B' \subset T_\delta(X \setminus M_{\rho+\varepsilon})$ . The inclusion  $B' \subset T_\delta(M_{\rho-\varepsilon})$  directly follows from  $B' \subset \{\hat{f}_\rho = 1\}$  and Lemma 23. The last assertion follows from the inclusion  $\{\hat{f}_{\rho+2\varepsilon+\epsilon} = 1\} \subset T_\delta(M_{\rho+\varepsilon+\epsilon})$  established in Lemma 23 and the previously shown inclusion. ■

**Proof of Theorem 24** Our first goal is to establish the following *disjoint* union:

$$\begin{aligned} \mathcal{C}_\tau(\{\hat{f}_\rho = 1\}) &= \zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon, \delta})) \\ &\cup \{B' \in \mathcal{C}_\tau(\{\hat{f}_\rho = 1\}) \setminus \zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon, \delta})) : B' \cap \{\hat{f}_{\rho+2\varepsilon+\epsilon} = 1\} \neq \emptyset\} \\ &\cup \{B' \in \mathcal{C}_\tau(\{\hat{f}_\rho = 1\}) : B' \cap \{\hat{f}_{\rho+2\varepsilon+\epsilon} = 1\} = \emptyset\}. \end{aligned} \tag{20}$$

We begin by showing the auxiliary result

$$V' \cap V_{\rho+3\varepsilon+\epsilon, \delta} \neq \emptyset, \quad V' \in \mathcal{C}_\tau(V_{\rho+\varepsilon, \delta}). \tag{21}$$

To this end, we observe that parts *i*) and *ii*) of Theorem 18 yield  $|\mathcal{C}_\tau(M_{\rho^{**}, \delta})| = 2$ , and hence part *ii*) of Theorem 20 implies  $|\mathcal{C}_\tau(V_{\rho^{**}, \delta})| = 2$ , and thus  $V_{\rho^{**}, \delta} \neq \emptyset$ . Let  $W'$  and  $W''$  be the two  $\tau$ -connected components of  $V_{\rho^{**}, \delta}$ . Let us first assume that  $V_{\rho+\varepsilon, \delta}$  has exactly one  $\tau$ -connected component  $V'$ , i.e.  $V' = V_{\rho+\varepsilon, \delta}$ . Then  $\rho + 3\varepsilon + \epsilon \leq \rho^{**}$  and  $\rho + \varepsilon \leq \rho + 3\varepsilon + \epsilon$  imply

$$\emptyset \neq V_{\rho^{**}, \delta} \subset V_{\rho+3\varepsilon+\epsilon, \delta} = V_{\rho+\varepsilon, \delta} \cap V_{\rho+3\varepsilon+\epsilon, \delta} = V' \cap V_{\rho+3\varepsilon+\epsilon, \delta},$$

i.e. we have shown (21). Let us now assume that  $V_{\rho+\varepsilon, \delta}$  has more than one  $\tau$ -component. Then it has exactly two such components  $V'$  and  $V''$  by  $\rho + \varepsilon < \rho^{**}$  and part *i*) of Theorem 20. By part *iii*) of Theorem 20 we may then assume without loss of generality that we have  $W' \subset V'$  and  $W'' \subset V''$ . Since  $\rho + 3\varepsilon + \epsilon \leq \rho^{**}$  implies  $V_{\rho^{**}, \delta} \subset V_{\rho+3\varepsilon+\epsilon, \delta}$ , these inclusions yield  $\emptyset \neq W' = W' \cap V_{\rho^{**}, \delta} \subset V' \cap V_{\rho+3\varepsilon+\epsilon, \delta}$  and  $\emptyset \neq W'' = W'' \cap V_{\rho^{**}, \delta} \subset V'' \cap V_{\rho+3\varepsilon+\epsilon, \delta}$ . Consequently, we have proved (21) in this case, too.

Now, from (21) we conclude by part *i*) of Lemma 35 that  $B' \cap \{\hat{f}_{\rho+2\varepsilon+\epsilon} = 1\} \neq \emptyset$  for all  $B' \in \zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon, \delta}))$ . This yields

$$\begin{aligned} & \{B' \in \mathcal{C}_\tau(\{\hat{f}_\rho = 1\}) \setminus \zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon, \delta})) : B' \cap \{\hat{f}_{\rho+2\varepsilon+\epsilon} = 1\} = \emptyset\} \\ &= \{B' \in \mathcal{C}_\tau(\{\hat{f}_\rho = 1\}) : B' \cap \{\hat{f}_{\rho+2\varepsilon+\epsilon} = 1\} = \emptyset\}, \end{aligned}$$

and the latter immediately implies (20).

Now, using (20) and  $\{\hat{f}_{\rho+2\varepsilon} = 1\} \supset \{\hat{f}_{\rho+2\varepsilon+\epsilon} = 1\}$  it remains to show

$$B' \cap \{\hat{f}_{\rho+2\varepsilon} = 1\} = \emptyset,$$

for all  $B' \in \mathcal{C}_\tau(\{\hat{f}_\rho = 1\}) \setminus \zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon, \delta}))$ . Let us assume the converse, that is, there exists some  $B' \in \mathcal{C}_\tau(\{\hat{f}_\rho = 1\})$  with  $B' \notin \zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon, \delta}))$  and  $B' \cap \{\hat{f}_{\rho+2\varepsilon} = 1\} \neq \emptyset$ . Since  $\{\hat{f}_{\rho+2\varepsilon} = 1\} \subset T_\delta(M_{\rho+\varepsilon})$  by Lemma 23, there then exists an  $x \in B' \cap T_\delta(M_{\rho+\varepsilon})$ . The latter yields an  $x' \in M_{\rho+\varepsilon}$  with  $d(x, x') \leq \delta$ , and since  $P$  has thick clusters we obtain

$$d(x', V_{\rho+\varepsilon, \delta}) < c \delta^\gamma.$$

From this inequality we conclude that there exists an  $x'' \in V_{\rho+\varepsilon, \delta}$  satisfying  $d(x', x'') < c \delta^\gamma$ . Let  $V'' \in \mathcal{C}_\tau(V_{\rho+\varepsilon, \delta})$  be the unique  $\tau$ -connected component satisfying  $x'' \in V''$ . The  $\tau$ -CCRM property then yields  $x'' \in V'' \subset \zeta(V'') =: B''$ , and hence, using  $c \geq 1$ , we find

$$d(B', B'') \leq d(x, x'') \leq d(x, x') + d(x', x'') < \delta + c \delta^\gamma \leq \tau.$$

However, since  $B' \notin \zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon, \delta}))$  and  $B'' \in \zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon, \delta}))$  we obtain  $B' \neq B''$ , and therefore, Lemma 29 yields  $d(B', B'') \geq \tau$ , i.e. we have found a contradiction. ■

**Proof of Theorem 25** *i*). Let  $D \in X^n$  be a dataset such that  $\|\bar{h}_{D, \mathcal{A}_\delta} - \bar{h}_{P, \mathcal{A}_\delta}\|_\infty < \varepsilon$ . Moreover, let  $\epsilon := 0$  and  $\rho \geq 0$  be the current level that is considered by Algorithm 3.1. Then, Theorem 24 shows that, for  $\rho \in [0, \rho^{**} - 3\varepsilon]$ , Algorithm 3.1 identifies exactly the  $\tau$ -connected components of  $\{\hat{f}_\rho = 1\}$  in its loop that belong to the set  $\zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon, \delta}))$ , where  $\zeta : \mathcal{C}_\tau(V_{\rho+\varepsilon, \delta}) \rightarrow \mathcal{C}_\tau(\{\hat{f}_\rho = 1\})$  is the  $\tau$ -CCRM. In the following, we thus consider the set  $\zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon, \delta}))$  for  $\rho \in [0, \rho^{**} - 3\varepsilon]$ .

Let us first consider the case  $\rho \in [0, \rho^* - \varepsilon]$ . Then, part *i*) and *iii*) of Theorem 20 together with the assumed  $\rho + \varepsilon < \rho^*$  show  $|\mathcal{C}_\tau(V_{\rho+\varepsilon,\delta})| = 1$ . This yields  $|\zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon,\delta}))| = 1$ , and hence Algorithm 3.1 does not stop. Consequently, we have  $\rho^*(D) \geq \rho^* - \varepsilon$ .

Let us now consider the case  $\rho \in [\rho^* + \varepsilon^* + \varepsilon, \rho^* + \varepsilon^* + 2\varepsilon]$ . Then we first note that Algorithm 3.1 actually inspects such an  $\rho$ , since it iteratively inspects all  $\rho = i\varepsilon$ ,  $i = 0, 1, \dots$ , and the width of the interval above is  $\varepsilon$ . Moreover, our assumptions on  $\varepsilon^*$  and  $\varepsilon$  guarantee  $\rho^* + \varepsilon^* + 2\varepsilon \leq \rho^{**} - 3\varepsilon$  and hence we have  $\rho \in [\rho^* + \varepsilon^* + \varepsilon, \rho^{**} - 3\varepsilon]$ . Let us write  $\zeta_V : \mathcal{C}_\tau(V_{\rho^{**},\delta}) \rightarrow \mathcal{C}_\tau(V_{\rho+\varepsilon,\delta})$ ,  $\zeta_M : \mathcal{C}_\tau(M_{\rho^{**},\delta}) \rightarrow \mathcal{C}_\tau(M_{\rho-\varepsilon,\delta})$ , and  $\zeta_{V,M} : \mathcal{C}_\tau(V_{\rho+\varepsilon,\delta}) \rightarrow \mathcal{C}_\tau(M_{\rho-\varepsilon,\delta})$  for the  $\tau$ -CCRMs between the involved sets. Using Lemma 11 twice, we then obtain the following diagram:

$$\begin{array}{ccc} \mathcal{C}_\tau(V_{\rho+\varepsilon,\delta}) & \xrightarrow{\zeta_{V,M}} & \mathcal{C}_\tau(M_{\rho-\varepsilon,\delta}) \\ \zeta_V \uparrow & & \uparrow \zeta_M \\ \mathcal{C}_\tau(V_{\rho^{**},\delta}) & \xrightarrow{\zeta^{**}} & \mathcal{C}_\tau(M_{\rho^{**},\delta}) \end{array}$$

Moreover, we have  $\rho - \varepsilon \geq \rho^* + \varepsilon^*$  and  $\rho + \varepsilon \geq \rho^* + \varepsilon^*$ , and hence part *ii*) and *iv*) of Theorem 18 together with part *i*) and *ii*) of Theorem 20 show that the sets  $M_{\rho-\varepsilon,\delta}$  and  $V_{\rho+\varepsilon,\delta}$  both have two  $\tau$ -connected components. Consequently, part *iii*) of Theorem 18 and part *iii*) of Theorem 20 ensure that the maps  $\zeta_V$  and  $\zeta_M$  are bijective, and, in addition, part *ii*) of Theorem 20 shows that  $\zeta^{**}$  is bijective. Consequently,  $\zeta_{V,M}$  is bijective. Let us further consider the  $\tau$ -CCRM  $\zeta' : \mathcal{C}_\tau(\{\hat{f}_\rho = 1\}) \rightarrow \mathcal{C}_\tau(M_{\rho-\varepsilon,\delta})$ . Then Lemma 11 yields another diagram:

$$\begin{array}{ccc} \mathcal{C}_\tau(V_{\rho+\varepsilon,\delta}) & \xrightarrow{\zeta_{V,M}} & \mathcal{C}_\tau(M_{\rho-\varepsilon,\delta}) \\ \zeta \searrow & & \swarrow \zeta' \\ & \mathcal{C}_\tau(\{\hat{f}_\rho = 1\}) & \end{array}$$

Since  $\zeta_{V,M}$  is bijective, we then find that  $\zeta$  is injective, and since we have already seen that  $V_{\rho+\varepsilon,\delta}$  has two  $\tau$ -connected components, we conclude that  $\zeta(\mathcal{C}_\tau(V_{\rho+\varepsilon,\delta}))$  contains two elements. Consequently, the stopping criterion of Algorithm 3.1 is satisfied, that is,  $\rho^*(D) \leq \rho^* + \varepsilon^* + 2\varepsilon$ .

*ii).* Theorem 24 shows that in its last run through the loop Algorithm 3.1 identifies exactly the  $\tau$ -connected components of  $\{\hat{f}_{\rho^*(D)} = 1\}$  that belong to the set  $\zeta_\varepsilon(\mathcal{C}_\tau(V_{\rho^*(D)+\varepsilon,\delta}))$ , where  $\zeta_\varepsilon : \mathcal{C}_\tau(V_{\rho^*(D)+\varepsilon,\delta}) \rightarrow \mathcal{C}_\tau(\{\hat{f}_{\rho^*(D)} = 1\})$  is the  $\tau$ -CCRM. Moreover, since Algorithm 3.1 stops at  $\rho^*(D)$ , we have  $|\zeta_\varepsilon(\mathcal{C}_\tau(V_{\rho^*(D)+\varepsilon,\delta}))| \neq 1$  and thus  $|\mathcal{C}_\tau(V_{\rho^*(D)+\varepsilon,\delta})| \neq 1$ . From  $\rho^*(D) + \varepsilon \leq \rho^{**}$  and part *i*) of Theorem 20 we thus conclude that  $|\mathcal{C}_\tau(V_{\rho^*(D)+\varepsilon,\delta})| = 2$ . For later purposes, note that the latter implies the injectivity of  $\zeta_\varepsilon$ . Therefore, *iii*) of Theorem 20 shows that the  $\tau$ -CCRM  $\zeta_{\rho^{**},\rho^*(D)+\varepsilon} : \mathcal{C}_\tau(V_{\rho^{**},\delta}) \rightarrow \mathcal{C}_\tau(V_{\rho^*(D)+\varepsilon,\delta})$  is bijective. Let us now consider the following commutative diagram:

$$\begin{array}{ccc}
\mathcal{C}_\tau(V_{\rho^{**}, \delta}) & \xrightarrow{\zeta_{\rho^{**}, \rho^*(D)+\varepsilon}} & \mathcal{C}_\tau(V_{\rho^*(D)+\varepsilon, \delta}) \\
& \searrow \zeta_{\rho^{**}, \rho^*(D)+3\varepsilon} & \swarrow \zeta \\
& \mathcal{C}_\tau(V_{\rho^*(D)+3\varepsilon, \delta}) &
\end{array}$$

where the remaining two maps are the corresponding  $\tau$ -CCRMs. Now the bijectivity of  $\zeta_{\rho^{**}, \rho^*(D)+\varepsilon}$  shows that  $\zeta_{\rho^{**}, \rho^*(D)+3\varepsilon}$  is injective, and since  $\rho^*(D) + 3\varepsilon \leq \rho^{**}$  implies  $|\mathcal{C}_\tau(V_{\rho^*(D)+3\varepsilon, \delta})| \leq 2 = |\mathcal{C}_\tau(V_{\rho^{**}, \delta})|$  by part *i*) of Theorem 20,  $\zeta_{\rho^{**}, \rho^*(D)+3\varepsilon}$  is actually bijective. This yields  $|\mathcal{C}_\tau(V_{\rho^*(D)+3\varepsilon, \delta})| = 2$  and the bijectivity of  $\zeta$ . Let us consider yet another commutative diagram

$$\begin{array}{ccc}
\mathcal{C}_\tau(V_{\rho^*(D)+3\varepsilon, \delta}) & \xrightarrow{\zeta} & \mathcal{C}_\tau(V_{\rho^*(D)+\varepsilon, \delta}) \\
\downarrow \zeta_{3\varepsilon} & & \downarrow \zeta_\varepsilon \\
\mathcal{C}_\tau(\{\hat{f}_{\rho^*(D)+2\varepsilon} = 1\}) & \xrightarrow{\zeta_f} & \mathcal{C}_\tau(\{\hat{f}_{\rho^*(D)} = 1\})
\end{array}$$

where again, all occurring maps are the  $\tau$ -CCRMs between the respective sets. Now we have already shown that  $\zeta_\varepsilon$  is injective and  $\zeta$  is bijective. Consequently,  $\zeta_{3\varepsilon}$  is injective.

*iii).* Follows from Theorem 24 and  $\rho^*(D) + 2\varepsilon \leq \rho^{**} - 3\varepsilon$ .

*iv).* By part *iii)* of Lemma 34 there exist bijective CCRMs  $\zeta_{\rho^{**}} : \mathcal{C}_\tau(V_{\rho^{**}, \delta}) \rightarrow \mathcal{C}(M_{\rho^{**}})$  and  $\zeta_{\rho^*(D)+3\varepsilon} : \mathcal{C}_\tau(V_{\rho^*(D)+3\varepsilon, \delta}) \rightarrow \mathcal{C}(M_{\rho^*(D)+3\varepsilon})$ . Moreover, in the proof of *ii)* we have already seen that  $\tau$ -CCRM  $\zeta_{\rho^{**}, \rho^*(D)+3\varepsilon}$  is bijective. This gives the diagram. ■

## Appendix E. Proofs related to large sample sizes

**Lemma 36** *Let  $(X, d)$  be a complete separable metric space,  $\mu$  be a finite measure on  $X$ , and  $(A_\rho)_{\rho \in \mathbb{R}}$  be a family of closed subsets of  $X$  with  $A_\rho \subset A_{\rho'}$  for all  $\rho' \leq \rho$ . For  $\rho^* \in \mathbb{R}$ , we write*

$$\bar{A}_{\rho^*} := \bigcup_{\rho > \rho^*} A_\rho \quad \text{and} \quad \dot{A}_{\rho^*} := \bigcup_{\rho > \rho^*} \dot{A}_\rho.$$

*Then we have*

$$\begin{aligned}
\dot{A}_{\rho^*} &= \bigcup_{\rho > \rho^*} \bigcup_{\varepsilon > 0} \bigcup_{\delta > 0} (X \setminus T_\delta(X \setminus A_{\rho+\varepsilon})) \\
\bar{A}_{\rho^*} &= \bigcup_{\rho > \rho^*} \bigcap_{\varepsilon > 0} \bigcap_{\delta > 0} T_\delta(A_{\rho-\varepsilon}).
\end{aligned}$$

*Moreover, the following statements are equivalent:*

$$i) \quad \mu(\bar{A}_{\rho^*} \setminus \dot{A}_{\rho^*}) = 0.$$

ii) For all  $\varepsilon > 0$ , there exists a  $\rho_\varepsilon > \rho^*$  such that, for all  $\rho \in (\rho^*, \rho_\varepsilon]$ , we have  $\mu(A_\rho \setminus \mathring{A}_\rho) \leq \varepsilon$ .

**Proof of Lemma 36** To show the first equality, we observe that (12) implies

$$\bigcap_{\rho > \rho^*} \bigcap_{\varepsilon > 0} \bigcap_{\delta > 0} T_\delta(X \setminus A_{\rho+\varepsilon}) = \bigcap_{\varepsilon > 0} \bigcap_{\rho > \rho^*} \overline{X \setminus A_{\rho+\varepsilon}} = \bigcap_{\rho > \rho^*} \overline{X \setminus A_\rho}.$$

Moreover, every set  $A \subset X$  satisfies  $\overline{X \setminus A} = X \setminus \mathring{A}$ , and hence we obtain

$$\bigcap_{\rho > \rho^*} \overline{X \setminus A_\rho} = \bigcap_{\rho > \rho^*} (X \setminus \mathring{A}_\rho) = X \setminus \bigcup_{\rho > \rho^*} \mathring{A}_\rho.$$

Combining both equalities and then taking the complement, we find the first assertion. Analogously, (12) shows

$$\bigcup_{\rho > \rho^*} \bigcap_{\varepsilon > 0} \bigcap_{\delta > 0} T_\delta(A_{\rho-\varepsilon}) = \bigcup_{\rho > \rho^*} \bigcap_{\varepsilon > 0} \overline{A_{\rho-\varepsilon}} = \bigcup_{\rho > \rho^*} \bigcap_{\varepsilon > 0} A_{\rho-\varepsilon}.$$

Moreover, using the monotonicity of the family  $(A_\rho)$ , it is straightforward to show that  $\bigcup_{\rho > \rho^*} \bigcap_{\varepsilon > 0} A_{\rho-\varepsilon} = \bigcup_{\rho > \rho^*} A_\rho$ , which finishes the proof.

i)  $\Rightarrow$  ii). Let us fix an  $\varepsilon > 0$ . Since  $\bigcup_{\rho' \geq \rho} \mathring{A}_{\rho'} = \mathring{A}_\rho \nearrow \mathring{A}_{\rho^*}$  for  $\rho \searrow \rho^*$ , the  $\sigma$ -continuity of finite measures yields an  $\rho_\varepsilon > \rho^*$  such that  $\mu(\bar{A}_{\rho^*} \setminus \mathring{A}_\rho) \leq \varepsilon$  for all  $\rho \in (\rho^*, \rho_\varepsilon]$ . Using  $A_\rho \subset \bar{A}_{\rho^*}$  for  $\rho > \rho^*$ , we then obtain the assertion.

ii)  $\Rightarrow$  i). Let us fix an  $\varepsilon > 0$ . For  $\rho \in (\rho^*, \rho_\varepsilon]$ , we then have  $\mathring{A}_\rho \subset \mathring{A}_{\rho^*}$ , and hence our assumption yields  $\mu(A_\rho \setminus \mathring{A}_{\rho^*}) \leq \varepsilon$ . In other words, we have  $\lim_{\rho \searrow \rho^*} \mu(A_\rho \setminus \mathring{A}_{\rho^*}) = 0$ . Moreover, we have  $A_\rho \nearrow \bar{A}_{\rho^*}$  for  $\rho \searrow \rho^*$ , and hence the continuity of  $\mu$  yields  $\lim_{\rho \searrow \rho^*} \mu(A_\rho \setminus \mathring{A}_{\rho^*}) = \mu(\bar{A}_{\rho^*} \setminus \mathring{A}_{\rho^*})$ . ■

**Proof of Theorem 26** Let us write  $A_{\rho^{**},i}$ ,  $i = 1, 2$ , for the two topologically connected components of  $M_{\rho^{**}}$ . Moreover, for  $\rho \in (\rho^*, \rho^{**}]$ , we define  $A_{\rho,i} := \zeta_\rho(A_{\rho^{**},i})$ , where  $\zeta_\rho : \mathcal{C}(M_{\rho^{**}}) \rightarrow \mathcal{C}(M_\rho)$  is the top-CCRM. In addition, we write  $A_{\rho,i} := \emptyset$  for  $\rho > \rho^{**}$  and  $A_{\rho,i} := X$  for  $\rho \leq \rho^*$ . Our first goal is to show that

$$\mu(\bar{A}_{\rho^*,i} \setminus \mathring{A}_{\rho^*,i}) = 0 \tag{22}$$

for  $i = 1, 2$ , where we used the notation of Lemma 36. To this end, we fix an  $\epsilon > 0$ . Since the definition of clusters ensures that  $P$  is normal at level  $\rho^*$ , we have  $\mu(\bar{M}_{\rho^*} \setminus \mathring{M}_{\rho^*}) = 0$ , Lemma 36 then shows that there exists a  $\rho_\epsilon > \rho^*$  such that  $\mu(M_\rho \setminus \mathring{M}_\rho) \leq \epsilon$  for all  $\rho \in (\rho^*, \rho_\epsilon]$ , where we may assume without loss of generality that  $\rho_\epsilon \leq \rho^{**}$ . Let us fix a  $\rho \in (\rho^*, \rho_\epsilon]$ . Then the fact that  $M_\rho = A_{\rho,1} \cup A_{\rho,2}$  is a disjoint union of closed sets yields  $\bar{M}_\rho = A_{\rho,1} \cup \mathring{A}_{\rho,2}$ . Consequently, we obtain

$$M_\rho \setminus \mathring{M}_\rho = (A_{\rho,1} \setminus (\mathring{A}_{\rho,1} \cup \mathring{A}_{\rho,2})) \cup (A_{\rho,2} \setminus (\mathring{A}_{\rho,1} \cup \mathring{A}_{\rho,2})) = (A_{\rho,1} \setminus \mathring{A}_{\rho,1}) \cup (A_{\rho,2} \setminus \mathring{A}_{\rho,2}).$$

This implies  $\mu(A_{\rho,i} \setminus \mathring{A}_{\rho,i}) \leq \epsilon$ , and hence Lemma 36 shows (22).

Let us now fix an  $\epsilon > 0$ . We define  $V_{\rho,\delta,i} := X \setminus T_\delta(X \setminus A_{\rho,i})$  for all  $\delta > 0$ ,  $\rho \in \mathbb{R}$ , and  $i = 1, 2$ . By the first equality of Lemma 36, Equation (22), and the  $\sigma$ -continuity of finite measures there then exist  $\delta_\epsilon > 0$ ,  $\varepsilon_\epsilon > 0$ , and  $\rho_\epsilon > \rho^*$  such that, for all  $\varepsilon \in (0, \varepsilon_\epsilon]$ ,  $\delta \in (0, \delta_\epsilon]$ ,  $\rho \in (\rho^*, \rho_\epsilon]$ , and  $i = 1, 2$  we have

$$\mu(\bar{A}_{\rho^*,i} \setminus V_{\rho+\varepsilon,\delta,i}) = \mu(\dot{A}_{\rho^*,i} \setminus V_{\rho+\varepsilon,\delta,i}) \leq \epsilon. \quad (23)$$

Moreover, the second equality of Lemma 36 shows that, for all  $\rho > \rho^*$ , we have

$$\bigcap_{\varepsilon > 0} \bigcap_{\delta > 0} M_{\rho-\varepsilon,\delta} \subset \bar{M}_{\rho^*}.$$

Clearly, this implies  $\bigcap_{\varepsilon > 0} \bigcap_{\delta > 0} M_{\rho-\varepsilon,\delta} \setminus \bar{M}_{\rho^*} = \emptyset$ . Consequently, we have

$$\mu(M_{\rho-\varepsilon,\delta} \setminus \bar{M}_{\rho^*}) \leq \epsilon \quad (24)$$

for all  $\rho > \rho^*$  and all sufficiently small  $\varepsilon > 0$ ,  $\delta > 0$ . Without loss of generality, we may thus assume that (24) holds for all  $\varepsilon \in (0, \varepsilon_\epsilon]$ ,  $\delta \in (0, \delta_\epsilon]$  and all  $\rho > \rho^*$ . We now define  $\varepsilon^* := \min\{\frac{\rho_\epsilon - \rho^*}{5}, \frac{\rho^{**} - \rho^*}{8}\}$ ,  $\varepsilon_\epsilon^* := \min\{\varepsilon^*, \varepsilon_\epsilon\}$ ,  $\delta^* := \min\{\delta_\epsilon, \delta_{\varepsilon^*}, \tilde{\delta}_0\}$ , and  $\tau^* := \tau_{\varepsilon^*}$ . Then, for all sufficiently large  $n$ , we have  $\varepsilon_n \in (0, \varepsilon^*]$ ,  $\delta_n \in (0, \delta^*]$ ,  $\tau_n \in (0, \tau^*]$ , and by Theorem 22 we further know that the probability  $P^n$  of  $\|\bar{h}_{D,\mathcal{A}_{\delta_n}} - \bar{h}_{P,\mathcal{A}_{\delta_n}}\|_\infty < \varepsilon_n$  converges to 1 for  $n \rightarrow \infty$ . Let us therefore only consider such data sets  $D$  and parameters satisfying  $\varepsilon_n \in (0, \varepsilon^*]$ ,  $\delta_n \in (0, \delta^*]$ ,  $\tau_n \in (0, \tau^*]$ . Then our construction ensures that we can apply Theorem 25. In particular, we have  $\rho^* < \rho^*(D) + 2\varepsilon_n \leq \rho^* + \varepsilon^* + 4\varepsilon_n \leq \rho^* + 5\varepsilon^* \leq \rho_\epsilon$ , and hence (23) and (24) hold for  $\rho := \rho^*(D) + 2\varepsilon_n$ . Following the discussion in front of Theorem 26, we further have two  $\tau_n$ -connected components  $V'_1$  and  $V'_2$  of  $V_{\rho+\varepsilon_n,\delta_n}$  and two  $\tau_n$ -connected components  $V''_1$  and  $V''_2$  of  $V_{\rho^{**},\delta_n}$  such that  $V''_i \subset V'_i$ ,  $V''_i \subset A_{\rho^{**},i}$ , and  $V'_i \subset B_i(D)$  for  $i = 1, 2$ . Let us next show that, for  $i = 1, 2$ , we have

$$V_{\rho+\varepsilon_n,\delta_n,i} \subset V'_i. \quad (25)$$

To this end, we fix an  $x \in V_{\rho+\varepsilon_n,\delta_n,1} = X \setminus T_{\delta_n}(X \setminus A_{\rho+\varepsilon_n,1})$ . Since  $V_{\rho+\varepsilon_n,\delta_n,1} \subset A_{\rho+\varepsilon_n,1}$  and  $V_{\rho+\varepsilon_n,\delta_n,1} \subset V_{\rho+\varepsilon_n,\delta_n}$ , we then have  $x \in A_{\rho+\varepsilon_n,1}$  and  $x \in V'_1 \cup V'_2$ . Let us assume that  $x \in V'_2$ . Then we have  $V'_2 \cap A_{\rho+\varepsilon_n,1} \neq \emptyset$ . Now, the diagram of Theorem 25 shows that  $\zeta_{\rho+\varepsilon_n} : \mathcal{C}_{\tau_n}(V_{\rho+\varepsilon_n,\delta_n}) \rightarrow \mathcal{C}(M_{\rho+\varepsilon_n})$  satisfies  $\zeta_{\rho+\varepsilon_n}(V'_i) = A_{\rho+\varepsilon_n,i}$ , and hence we have  $V'_2 \subset A_{\rho+\varepsilon_n,2}$ . Consequently,  $V'_2 \cap A_{\rho+\varepsilon_n,1} \neq \emptyset$  implies  $A_{\rho+\varepsilon_n,2} \cap A_{\rho+\varepsilon_n,1} \neq \emptyset$ , which is a contradiction. Therefore, we have  $x \in V'_1$ , that is, we have shown (25) for  $i = 1$ . The case  $i = 2$  can be shown analogously.

Using  $A_i^* = \bar{A}_{\rho^*,i}$ ,  $V'_i \subset B_i(D)$ , (25), and (23) we now obtain

$$\mu(A_i^* \setminus B_i(D)) = \mu(\bar{A}_{\rho^*,i} \setminus B_i(D)) \leq \mu(\bar{A}_{\rho^*,i} \setminus V'_i) \leq \mu(\bar{A}_{\rho^*,i} \setminus V_{\rho+\varepsilon_n,\delta_n,i}) \leq \epsilon. \quad (26)$$

Conversely, using  $\mu(B \setminus A) = \mu(B) - \mu(A \cap B)$  twice, we obtain

$$\begin{aligned} \mu(B_1(D) \setminus (A_1^* \cup A_2^*)) &= \mu(B_1(D)) - \mu(B_1 \cap (A_1^* \cup A_2^*)) \\ &\geq \mu(B_1(D)) - \mu(B_1(D) \cap A_1^*) - \mu(B_1(D) \cap A_2^*) \\ &= \mu(B_1(D) \setminus A_1^*) - \mu(B_1(D) \cap A_2^*). \end{aligned}$$

Since  $B_1(D) \cap B_2(D) = \emptyset$  implies  $B_1(D) \cap A_2^* \subset A_2^* \setminus B_2(D)$  and Lemma 23 shows  $B_1(D) \subset M_{\rho+\varepsilon_n, \delta_n}$ , we can thus conclude with the help of the previous estimate that

$$\begin{aligned}\mu(B_1(D) \setminus A_1^*) &\leq \mu(B_1(D) \setminus (A_1^* \cup A_2^*)) + \mu(A_2^* \setminus B_2(D)) \\ &\leq \mu(M_{\rho+\varepsilon_n, \delta_n} \setminus (A_1^* \cup A_2^*)) + \mu(A_2^* \setminus B_2(D)) \\ &\leq 2\epsilon,\end{aligned}$$

where in the last step we used (24) and (26). Clearly, we can establish  $\mu(B_2(D) \setminus A_2^*) \leq 2\epsilon$  analogously, and hence we finally obtain  $\mu(B_i(D) \Delta A_i^*) \leq 3\epsilon$  for  $i = 1, 2$ .  $\blacksquare$

STEINWART

# Agnostic KWIK learning and efficient approximate reinforcement learning

István Szita

Csaba Szepesvári

*Department of Computing Science  
University of Alberta, Canada*

SZITA@UALBERTA.CA

SZEPESVA@UALBERTA.CA

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

A popular approach in reinforcement learning is to use a model-based algorithm, i.e., an algorithm that utilizes a model learner to learn an approximate model to the environment. It has been shown that such a model-based learner is efficient if the model learner is efficient in the so-called “knows what it knows” (KWIK) framework. A major limitation of the standard KWIK framework is that, by its very definition, it covers only the case when the (model) learner can represent the actual environment with no errors. In this paper, we study the agnostic KWIK learning model, where we relax this assumption by allowing nonzero approximation errors. We show that with the new definition an efficient model learner still leads to an efficient reinforcement learning algorithm. At the same time, though, we find that learning within the new framework can be substantially slower as compared to the standard framework, even in the case of simple learning problems.

**Keywords:** KWIK learning, agnostic learning, reinforcement learning, PAC-MDP

## 1. Introduction

The *knows what it knows* (KWIK) model of learning (Li et al., 2008) is a framework for online learning against an adversary. Before learning, the KWIK learner chooses a hypothesis class and the adversary selects a function from this hypothesis class, mapping inputs to responses. Then, the learner and the adversary interact in a sequential manner: Given the past interactions, the adversary chooses an input, which is presented to the learner. The learner can either pass, or produce a prediction of the value one would obtain by applying the function selected by the adversary to the selected input. When the learner passed and only in that case, the learner is shown the noise-corrupted true response. *All* predictions produced by the learner must be in a close vicinity to the true response (up to a prespecified tolerance), while the learner’s efficiency is measured by the number of times it passes.

The problem with this framework is that if the hypothesis class is small, it unduly limits the power of the adversary, while with a larger hypothesis class efficient learning becomes

problematic. Hence, in this paper we propose an alternative framework that we call the *agnostic KWIK* framework, where we allow the adversary to select functions outside of the hypothesis class, as long the function remains “close” to the hypothesis class, while simultaneously relaxing the accuracy requirement on the predictions.

New models of learning abound in the learning theory literature, and it is not immediately clear why the KWIK framework makes these specific assumptions on the learning process. For the extension investigated in the paper, the *agnostic KWIK* model, even the name seems paradoxical: “agnostic” means “no knowledge is assumed”, while KWIK is acronym for “knows what it knows”. Therefore, we begin the paper by motivating the framework.

### 1.1. Motivation

The motivation of the KWIK framework is rooted in reinforcement learning (RL). An RL agent makes sequential decisions in an environment to maximize the long-term cumulated reward it incurs during the interaction (Sutton and Barto, 1998). The environment is initially unknown to the agent, so the agent needs to spend some time exploring it. Exploration, however, is costly as an agent exploring its environment may miss some reward collecting opportunities. Therefore, an efficient RL agent must spend as little time with exploration as possible, while ensuring that the best possible policy is still discovered.

Many efficient RL algorithms (Kearns and Singh, 2002; Brafman and Tennenholtz, 2001; Strehl, 2007; Szita and Lörincz, 2008; Szita and Szepesvári, 2010) share a common core idea: (1) they keep track of which parts of the environment are known with high accuracy; (2) they strive to get to unknown areas and collect experience; (3) in the known parts of the environment, they are able to plan the path of the agent to go wherever it wants to go, such as the unknown area or a highly rewarding area. The KWIK learning model of Li et al. (2008) abstracts the first point of this core mechanism. This explains the requirements of the framework:

- Accuracy of predictions: a plan based on an approximate model will be usable only if the approximation is accurate. Specifically, a single large error in the model can fatally mislead the planning procedure.
- Adversarial setting: the state of the RL agent (and therefore, the queries about the model) depend on the (unknown) dynamics of the environment in a complex manner. While the assumption that the environment is fully adversarial gives more power to the adversary, this assumption makes the analysis easier (while not preventing it).
- Noisy feedback: the rewards and next states are determined by a stochastic environment, so feedback is necessarily noisy.

The main result of the KWIK framework states that if an algorithm “KWIK-learns” the parameters of an RL environment then it can be augmented to an efficient reinforcement learning algorithm (Li, 2009, Chapter 7.1) and (Li et al., 2011a). The result is significant because it reduces efficient RL to a conceptually simpler problem and unifies a large body of previous works (Li, 2009; Strehl et al., 2007; Diuk et al., 2008; Strehl and Littman, 2007).

For finite horizon learning problems, it is even possible to construct *model-free* efficient RL algorithms using an appropriate KWIK learner, as shown by Li and Littman (2010).

An important limitation of the KWIK framework is that the environment must be exactly representable by the learner. Therefore, to make learning feasible, we must assume that the environment comes from a small class of models (that is, characterized with a small number of parameters), for example, it is a Markov decision process (MDP) with a small, finite state space.

However, such a model of the environment is often just an approximation, and in such cases, not much is known about efficient learning in a KWIK-like framework. The agnostic KWIK learning framework is aimed to fill this gap. In this new framework the learner tries to find a good approximation to the true model with a restricted model class.<sup>1</sup> Of course, we will not be able to predict the model parameters accurately any more (the expressive power of our hypothesis class is insufficient), so the accuracy requirement needs to be relaxed. Our main result is that with this definition the augmentation result of Li et al. (2011a) still holds: an efficient agnostic KWIK-learning algorithm can be used to construct an efficient reinforcement learning algorithm even when the environment is outside of the hypothesis class of the KWIK learner. To our knowledge, this is the first result for reinforcement learning that allows for a nonzero approximation error.

## 1.2. The organization of the paper

In the next section (Section 2) we introduce the KWIK framework and its agnostic extension. In the two sections following Section 2 we investigate simple agnostic KWIK learning problems. In particular, in Section 3 we investigate learning when the responses are noiseless. Two problems are considered: As a warm-up we consider learning with finite hypothesis classes, followed by the investigation of learning when the hypothesis class contains linear functions with finitely many parameters. In Section 4 we analyze the case when the responses are noisy. Section 5 contains our main result: the connection between agnostic KWIK and efficient approximate RL. Our conclusions are drawn in Section 6. Proofs of technical theorems and lemmas have been moved to the Appendix.

## 2. From KWIK learning to agnostic KWIK learning

A *problem* is a 5-tuple  $G = (\mathcal{X}, \mathcal{Y}, g, Z, \|\cdot\|)$ , where  $\mathcal{X}$  is the set of inputs,  $\mathcal{Y} \subseteq \mathbb{R}^d$  is a measurable set of possible responses,  $Z : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$  is the noise distribution that is assumed to be zero-mean ( $\mathcal{P}(\mathcal{Y})$  denotes the space of probability distributions over  $\mathcal{Y}$ ) and  $\|\cdot\| : \mathbb{R}^d \rightarrow \mathbb{R}_+$  is a semi-norm on  $\mathbb{R}^d$ . A *problem class*  $\mathcal{G}$  is a set of problems. When each problem in a class shares the same domain  $\mathcal{X}$ , response set  $\mathcal{Y}$  and same semi-norm  $\|\cdot\|$ , for brevity, the semi-norm will be omitted from the problem specifications. If the noise distribution underlying every  $G \in \mathcal{G}$  is a Dirac-measure, we say that the problem class is *deterministic*. For such problem classes, we will also omit to mention the distribution.

---

1. The real environment is *not known* to belong to the restricted model class, hence the name “agnostic”.

The *knows what it knows* (KWIK) framework Li et al. (2011a) is a model of online learning where an (online) learner interacts with an environment.<sup>2</sup> In this context, an *online learner*  $L$  is required to be able to perform two operations:

- **predict:** For an input  $x \in \mathcal{X}$ ,  $L$  must return an answer  $\hat{y} \in \mathcal{Y} \cup \{\perp\}$ . The answer  $\hat{y} = \perp$  means that the learner *passes*.
- **update:** Upon receiving an input-response pair  $(x, y) \in \mathcal{X} \times \mathcal{Y}$ ,  $L$  should update its internal representation.

At the beginning of learning, the environment secretly selects a problem  $(\mathcal{X}, \mathcal{Y}, g^*, Z)$  from some class and it also selects the inputs  $x_t$  which are presented to the learner in a sequential manner. Given an input  $x_t$ , the learner has the option to pass (say “I don’t know”), or to make a prediction. An admissible learner is required to make accurate predictions only. When the learner passes (and only in that case), the environment tells it the response (or answer)  $y_t$ , which is randomly chosen so that  $z_t = y_t - g^*(x_t) \sim Z(x_t)$  and  $z_t$  is independent of the past given  $x_t$ . The learner’s goal is to minimize the number of passes, while staying admissible. The environment is assumed to choose the problem and the inputs adversarially and so we sometimes call the environment the *adversary*. In the case of noisy responses, exact, or near-optimal predictions are in general impossible to achieve with certainty. Correspondingly, we introduce two parameters: the required accuracy  $\epsilon \geq 0$  and the maximum permitted failure probability  $\delta$ .

The KWIK protocol, controlling the interaction between the online learner and the adversary, is shown as Algorithm 1. Note that in the standard KWIK-framework, before

---

**Algorithm 1** *The KWIK protocol*  $(\mathcal{G}, \epsilon)$ .

---

- 1:  $N_0 = 0$   $\{N_t$  is the number of times the learner “passed”. $\}$
  - 2: Adversary picks problem  $G^* = (\mathcal{X}, \mathcal{Y}, g^*, Z, \|\cdot\|) \in \mathcal{G}$  and  $(\mathcal{X}, \mathcal{Y}, \|\cdot\|)$  is told learner
  - 3: **for**  $t = 1, 2, \dots$  **do**
  - 4:   Adversary picks query  $x_t \in \mathcal{X}$ , which is announced to learner.
  - 5:   Learner computes answer  $\hat{y}_t \in \mathcal{Y} \cup \{\perp\}$  (**predict** is called), which is announced to adversary.
  - 6:    $N_t = N_{t-1}$
  - 7:   **if**  $\hat{y}_t = \perp$  **then**
  - 8:     Adversary tells learner  $y_t = g^*(x_t) + z_t$ , where  $z_t \sim Z(\cdot|x_t)$
  - 9:     Learner updates itself (**update** is called)
  - 10:     $N_t$  is incremented by 1
  - 11:   **else if**  $\|\hat{y}_t - g^*(x_t)\| > \epsilon$  **then**
  - 12:     **return** FAIL
- 

learning, both adversary and learner are given  $\mathcal{G}$  and  $\epsilon$ . Further, learner might be given a confidence parameter  $0 \leq \delta \leq 1$ , whose role will be explained soon. In particular, this means that the learner can adjust its strategy to  $(\mathcal{G}, \epsilon, \delta)$ . Note also that the environment

---

2. Our definitions are slightly different from the original ones, mostly for the sake of increased rigour and to make them better fit our results. Specifically, we explicitly include the noise as part of the concept.

is allowed to pick  $x_t$  based on any past information available to it up to time  $t$ .<sup>3</sup> We note in passing that if in an application, regardless of the decision of the learner, the response  $y_t$  is generated and is communicated to the learner at every step, the learning problem can only become easier. We call the so-modified protocol, the *relaxed KWIK protocol*. If the learner is reasonable, the extra information will help it, though, this calls for an explicit proof. All the definitions below extend to the relaxed KWIK protocol.

**Definition 2.1** Fix  $\epsilon \geq 0$  and  $0 \leq \delta \leq 1$  and a problem class  $\mathcal{G}$ . A learner  $L$  is an admissible (and bounded)  $(\epsilon, \delta)$  KWIK-learner for  $\mathcal{G}$  if, with probability at least  $1 - \delta$ , it holds that when  $L$  and an arbitrary adversary interact following the KWIK protocol, the protocol does not fail (and the number of passes  $N_t$  stays bounded by a finite deterministic quantity  $B(\mathcal{G}, \epsilon, \delta)$ ). We call the quantity  $B(\mathcal{G}, \epsilon, \delta)$  the learner's KWIK-bound. The problem class  $\mathcal{G}$  is  $(\epsilon, \delta)$  KWIK-learnable, if there exists a bounded, admissible  $(\epsilon, \delta)$  KWIK-learner  $L$  for  $\mathcal{G}$ . Further,  $\mathcal{G}$  is KWIK-learnable, if it is  $(\epsilon, \delta)$  KWIK learnable for any  $\epsilon > 0$ ,  $0 < \delta < 1$ . If  $B(\mathcal{G}, \epsilon, \delta)$  is the learner's KWIK-bound, we say that  $\mathcal{G}$  is KWIK-learnable with KWIK-bound  $B(\mathcal{G}, \epsilon, \delta)$ .

Note that the learners can be specialized to  $\mathcal{G}$ ,  $\epsilon$  and  $\delta$ . However, interesting results concern general KWIK-learners which are operate for any  $\mathcal{G}$  from a meta-class of concept-classes  $\mathcal{C}$ . For example, the memorization learner of Li (2009) is a bounded, admissible KWIK learner for any problem class  $\mathcal{G}$  where the problems in  $\mathcal{G}$  are deterministic and share the same finite input space  $\mathcal{X}$ .

In addition to the above concepts, it is also customary to define the notion of KWIK-learnability:

**Definition 2.2** Let  $c : \mathcal{G} \rightarrow \mathbb{R}_+$  be a real-valued function. The problem class  $\mathcal{G}$  is  $c$ -efficiently KWIK-learnable, if, for any  $\epsilon > 0$ ,  $0 \leq \delta \leq 1$ , it is  $(\epsilon, \delta)$  KWIK-learnable by a learner  $L$  whose KWIK-bound  $B$  satisfies  $B(\mathcal{G}, \epsilon, \delta) \leq \text{poly}(c(\mathcal{G}), 1/\epsilon, \log(1/\delta))$  for some polynomial  $\text{poly}$ . Further,  $\mathcal{G}$  is  $c$ -efficiently deterministically KWIK-learnable if the above polynomial is independent of  $\delta$ . Finally,  $\mathcal{G}$  is  $c$ -exactly KWIK-learnable if  $\text{poly}(c, 1/\epsilon, \log(1/\delta))$  is independent of  $1/\epsilon$ .<sup>4</sup>

Examples of KWIK-learnable classes can be found in the thesis by Li (2009).

## 2.1. Agnostic KWIK learning

From now on, we will assume that  $\mathcal{G}$  is such that all problems in it share the same domain and response spaces. A crucial assumption of the KWIK framework is that learner gets to know  $\mathcal{G}$  at the beginning of learning – the so-called *realizability* assumption.<sup>5</sup> To illustrate

- 
- 3. The choice must be measurable to avoid pathologies.
  - 4. In the definition, contrary to previous work, we intentionally use  $\log(1/\delta)$  instead of  $1/\delta$  because  $\log(1/\delta)$  is more natural in a learning context and a  $1/\delta$ -bound looks unnecessarily weak.
  - 5. If the learner knows  $\mathcal{G}$ , it can “realize” any problem chosen by the adversary, hence the name.

the importance of this assumption take  $\mathcal{X} = \mathbb{N}$ ,  $\mathcal{Y} = \{-1, +1\}$  and let  $\mathcal{G}$  have two disjoint deterministic problems (functions) in it. Then, trivially, there is a KWIK-learner which is bounded and admissible *independently of* how the two deterministic functions are chosen. However, for any KWIK-learner who remains *uninformed about the choice of  $\mathcal{G}$*  there exists a class  $\mathcal{G}$  with two functions that makes the learner fail.

However, in practice the realizability assumption might be restrictive: The user of a learning algorithm might give the learner a problem class (the hypothesis class),  $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ , that may or may not contain the problem to be learned. In this case one still expects performance to degrade in a graceful manner as a function of the “distance” between the problem selected by the adversary and  $\mathcal{H}$ . In particular, it is reasonable to relax the accuracy requirements in proportion to this distance. Therefore, in our *agnostic KWIK learning framework* we propose to allow prediction errors of size  $rD + \epsilon$  (instead of  $\epsilon$ ), where  $D$  is the maximum tolerable approximation error.

The formal definitions are as follows. Let

$$\Delta(\mathcal{G}, \mathcal{H}) \stackrel{\text{def}}{=} \sup_{(\mathcal{X}, \mathcal{Y}, g, Z) \in \mathcal{G}} \inf_{h \in \mathcal{H}} \|h - g\|_{\infty}.$$

denote the error of approximating the functions of  $\mathcal{G}$  by elements of  $\mathcal{H}$ .

**Definition 2.3** Fix a hypothesis class  $\mathcal{H}$  over the domain  $\mathcal{X}$  and response set  $\mathcal{Y}$ , an approximation error bound  $D > 0$ , a competitiveness factor  $r > 0$ , an accuracy-slack  $\epsilon \geq 0$  and a confidence parameter  $0 \leq \delta \leq 1$ . Let  $\mathcal{G}$  be a problem class  $\mathcal{G}$  over  $(\mathcal{X}, \mathcal{Y})$  that satisfies  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$ . Then, a learner  $L$  is a  $(D, r, \epsilon, \delta)$  agnostic KWIK-learner for the pair  $(\mathcal{H}, \mathcal{G})$  if  $L$  is an  $(rD + \epsilon, \delta)$  KWIK-learner for  $\mathcal{G}$ .

The other learnability concepts (e.g.,  $(\epsilon, \delta)$  learnability, learnability, efficiency, etc.) can also be defined analogously.

### 3. Learning deterministic problem classes

The results in this section are used to illustrate the definitions, and the role of the various parameters (such as  $r$ ). The common property of the learning problems studied here is that the responses are noise-free. As a warm-up, learning with finite hypothesis classes is considered. Next, we consider learning with a hypothesis class composed of linear functions. We will see that in this case KWIK-learning is still possible, but can be exponentially slow as a function of the dimension of the input space. Note that our learners are deterministic. Hence, all statements hold either with probability one or probability zero. In particular, a bounded, admissible KWIK-learner is necessarily a bounded and admissible KWIK-learner even with the choice of  $\delta = 0$ .

### 3.1. Learning with a finite hypothesis class

For any  $d > 0$  and  $y \in \mathcal{Y}$ , define the  $d$ -ball around  $y$  as

$$B_d(y) \stackrel{\text{def}}{=} \{y' \in \mathcal{Y} : \|y' - y\| \leq d\}.$$

---

**Algorithm 2** Generic Agnostic Learner for deterministic problem classes.

---

<b>initialize</b> ( $D, \mathcal{H}$ ) $\mathcal{F} := \mathcal{H}$ and store $D$ <b>learn</b> ( $x, y$ ) $\mathcal{F} := \mathcal{F} \setminus \{f \in \mathcal{F} : \ f(x) - y\  > D\}$	<b>predict</b> ( $x$ ) $Y := \bigcap_{f \in \mathcal{F}} B_D(f(x))$ <b>if</b> $Y \neq \emptyset$ <b>then</b> <b>return</b> an arbitrary $\hat{y} \in Y$ <b>else</b> <b>return</b> $\perp$
--	--

---

Consider the Generic Agnostic Learner (Algorithm 2) of Littman (the algorithm is published by Li (2009), without analysis and for  $\mathcal{Y} \subseteq \mathbb{R}$ ). Every time a new query  $x_t$  is received, the algorithm checks whether there exists some value  $\hat{y}_t$  that remaining functions agree at  $x_t$  up to the accuracy  $D$ . If such a value exists, the learner predicts  $\hat{y}_t$ , otherwise it passes. When the learner passes it learns the response  $y_t$ , based on which it can exclude at least one concept. This results in the following statement:

**Theorem 3.1** *Let  $(\mathcal{X}, \mathcal{Y})$  be arbitrary sets,  $r = 2$ ,  $\epsilon = 0$ ,  $D > 0$ ,  $\mathcal{H}$  a finite hypothesis class over  $(\mathcal{X}, \mathcal{Y})$ ,  $\mathcal{G}$  a deterministic problem class over  $(\mathcal{X}, \mathcal{Y})$  with  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$ . Then, the Generic Agnostic Learner is an agnostic  $(D, r, \epsilon)$  KWIK-learner for  $(\mathcal{H}, \mathcal{G})$  with KWIK-bound  $|\mathcal{H}| - 1$ .*

The factor  $r = 2$  in the above theorem is the best possible as long as  $\mathcal{X}$  is infinite:

**Theorem 3.2** *Fix any  $D > 0$  and an infinite domain  $\mathcal{X}$ . Then, there exists a finite response set  $\mathcal{Y} \subset \mathbb{R}$ , a two-element hypothesis class  $\mathcal{H}$  and a deterministic problem class  $\mathcal{G}$ , both over  $(\mathcal{X}, \mathcal{Y})$ , that satisfy  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$  such that there is no bounded agnostic  $(D, r, 0)$  KWIK-learner for  $(\mathcal{H}, \mathcal{G})$  with competitiveness factor  $0 \leq r < 2$ .*

Because of this result, in what follows, we will restrict our attention to  $r = 2$ . Note that the KWIK-bound we got is identical to the (worst-case) bound that is available when  $D = 0$ , that is, seemingly there is no price associated to  $D > 0$  in terms of the KWIK-bound. However, the worst-case approach taken here leads to overly conservative bounds as the structure of a hypothesis space may allow much better bounds (this is discussed in Section 5.5.2 of Li (2009)). In the next section we study linear hypothesis classes for which the above bound would be vacuous.

### 3.2. Learning with linear hypotheses

Sometimes Algorithm 2 is applicable even when the hypothesis class  $\mathcal{H}$  is infinite: First, the set of remaining hypotheses  $\mathcal{F} = \mathcal{H}(x_1, y_1, \dots, x_n, y_n)$  after  $n$  passes can be “implicitly

represented” by storing the list  $(x_1, y_1, \dots, x_n, y_n)$  of pairs received after the  $n$  passes. Then, the algorithm remains applicable as long as there is some procedure for checking whether  $Y = \bigcap_{f \in \mathcal{F}} B_D(f(x))$  is empty (and finding an element of  $Y$ , if it is non-empty). It remains to see then if the procedure is efficient and if the learner stays bounded (that the procedure stays admissible for  $r \geq 2$  follows from its definition).

In this section we consider the case when the functions in the hypothesis class are linear in the inputs. More specifically, let  $\mathcal{X} = [-X_{\max}, X_{\max}]^d$  for some  $d \in \{1, 2, \dots\} = \mathbb{N}$ ,  $\mathcal{Y} = \mathbb{R}$ ,  $\|\cdot\| = |\cdot|$ ,  $M > 0$  and choose  $\mathcal{H}$  to be the set of bounded-parameter linear functions: Denote by  $f_\theta : \mathcal{X} \rightarrow \mathbb{R}$  the linear function  $f_\theta(x) = \theta^\top x$ ,  $x \in \mathcal{X}$ . Then,

$$\mathcal{H}_{\text{lin}(M)} \stackrel{\text{def}}{=} \{f_\theta : \theta \in \mathbb{R}^d, \|\theta\|_\infty \leq M\}.$$

Then,  $\mathcal{F} = \mathcal{H}(x_1, y_1, \dots, x_n, y_n) = \{f_\theta : \theta \in \mathbb{R}^d, -M \leq \theta_i \leq M, y_j - D \leq f_\theta(x_j) \leq y_j + D, 1 \leq i \leq d, 1 \leq j \leq n\}$  since for any  $(x_i, y_i)$  pair the hypotheses not excluded must satisfy  $|f_\theta(x) - y| \leq D$ . Now,  $Y = [y^-, y^+]$ , where  $y^- = \min_{f \in \mathcal{F}} f(x)$  and  $y^+ = \max_{f \in \mathcal{F}} f(x)$  and both  $y^-$  and  $y^+$  can be efficiently computed using linear programming. The resulting algorithm is called the deterministic linear agnostic learner (Algorithm 3). In the algorithm, we also allow for a slack  $\epsilon > 0$ .

---

**Algorithm 3** Deterministic Linear Agnostic Learner

---

```

initialize( $X_{\max}$ ,  $M$ ,  $D$ ,  $\epsilon$ )
   $C := \{\theta : -M \leq \theta_i \leq M, \forall i \in \{1, \dots, d\}\}$ 
  learn( $x, y$ )
     $C := C \cap \{\theta : y - D \leq \theta^\top x \leq y + D\}$ 
predict( $x$ )
   $y^+ := \max_{\theta \in C} \theta^\top x$  {solve LP}
   $y^- := \min_{\theta \in C} \theta^\top x$  {solve LP}
  if  $y^+ - y^- \leq 2(D + \epsilon)$  then
    return  $(y^+ + y^-)/2$ 
  else
    return  $\perp$ 

```

---

The following theorem shows that for  $\epsilon > 0$  this algorithm is a bounded, admissible agnostic KWIK learner:

**Theorem 3.3** *Let  $X_{\max} > 0$ ,  $\mathcal{X} = [-X_{\max}, X_{\max}]^d$ ,  $\mathcal{Y} = \mathbb{R}$ ,  $M, D, \epsilon > 0$ ,  $r = 2$ ,  $\mathcal{H} = \mathcal{H}_{\text{lin}(M)}$ . Then, for any  $\mathcal{G}$  deterministic problem class over  $(\mathcal{X}, \mathcal{Y})$  with  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$ , it holds that the deterministic linear agnostic learner is an agnostic  $(D, r, \epsilon)$  KWIK-learner for  $(\mathcal{H}, \mathcal{G})$  with the KWIK-bound  $2d! \left( \frac{MX_{\max}}{\epsilon} + 1 \right)^d$ .*

The theorem cannot hold with  $\epsilon = 0$ , as shown by the following result:

**Theorem 3.4** *Let  $\mathcal{X}, \mathcal{Y}, D, r, \mathcal{H}$  be as in Theorem 3.3. Then, there exists a problem class  $\mathcal{G}$  that satisfies  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$  such that there is no bounded, agnostic  $(D, r, 0)$  KWIK-learner for  $(\mathcal{H}, \mathcal{G})$ .*

In Theorem 3.3 the KWIK-bound scales exponentially with  $d$ . The next result shows that this is the best possible scaling behavior:

**Theorem 3.5** Fix  $\mathcal{X}, \mathcal{Y}, D, \mathcal{H}, \epsilon$  as in Theorem 3.3 and let  $r \geq 2$ . Then, there exists some problem class  $\mathcal{G}$  so that any algorithm that agnostic  $(D, r, \epsilon)$  KWIK learns  $(\mathcal{H}, \mathcal{G})$  will pass at least  $2^{d-1}$  times.

Whether the scaling behavior of the bound of Theorem 3.3 as a function of  $\epsilon$  can be improved remains for future work. Note that for  $D = 0$  we get an algorithm close to Algorithm 13 of Li (2009), the difference being that Algorithm 13 never predicts unless the new input lies in the span of past training vectors. Nevertheless, for  $D = 0$  and any  $\epsilon \geq 0$  (including  $\epsilon = 0$ ), our algorithm is also an  $\epsilon$  KWIK-learner with KWIK-bound  $d$ . Thus, we see that the price of non-realizability is quite high.

#### 4. Learning in bounded noise

As opposed to the previous section, in this section we consider the case when the responses are noisy. We first consider the case of learning with finite hypothesis classes and then we briefly outline a simple discretization-based approach to the case when the hypothesis class is infinite. We further assume that  $\mathcal{Y} \subseteq \mathbb{R}$  and the range of the noise in the responses is bounded by  $K$ .

##### 4.1. The case of finite hypothesis classes

Let the finite hypothesis class given to the learner be  $\mathcal{H}$  and fix some  $D > 0$ . Let  $g^*$  be the function underlying the problem chosen by the adversary from some class  $\mathcal{G}$  which satisfies  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$ . Assuming that the noise in the responses is bounded to lie in  $[-K, K]$  for some  $K > 0$ , an application of the Hoeffding-Azuma inequality gives that for any  $\epsilon > 0$  and any fixed function  $f \in \mathcal{H}$  such that  $\|f - g^*\| \leq D$  (such functions exists, by our assumption connecting  $\mathcal{G}$  and  $\mathcal{H}$ ),  $0 < \delta \leq 1$ , with probability  $1 - \delta$ , it holds that

$$\left| \frac{1}{m} \sum_{k=1}^m \{f(x_k) - y_k\} \right| \leq D + K \sqrt{\frac{2}{m} \log \left( \frac{2}{\delta} \right)} \leq D + \epsilon, \quad (1)$$

where  $m = m(\epsilon, \delta) > 0$  is chosen large enough so that the second inequality is satisfied and where  $((x_k, y_k); k = 1, 2, \dots)$  is the list of training examples available to the algorithm (the application of Hoeffding-Azuma is not entirely immediate, for the details see Lemma A.1). One idea then is to eliminate those functions  $f$  from  $\mathcal{H}$  which fail to satisfy (1) after  $m$  examples have been seen. The problem is that this rule is based on an average. A clever adversary, who wants to prevent the elimination of some function  $\hat{f} \in \mathcal{H}$  could then provide many examples  $(x_k, y_k)$  such that  $y_k$  is close to  $\hat{f}(x_k)$ , thus shifting the average to a small value. Therefore, we propose an alternate strategy which is based on the pairwise comparison of hypothesis.

The idea is that if  $f, f'$  are far from each other, say at  $x \in \mathcal{X}$  it holds that  $|f(x) - f'(x)| > 2(D + \epsilon)$ , then if the adversary feeds  $x$  enough number of times, we can eliminate at least one of  $f$  and  $f'$ . The following definition will become handy: for two numbers  $y, y' \in \mathbb{R}$ , we define

$y \ll y'$  by  $y + 2(D + \epsilon) \leq y'$ . We index the elements of  $\mathcal{H}$  from 1 to  $N$ :  $\mathcal{H} = \{f_1, \dots, f_N\}$ . The algorithm that we propose is shown as Algorithm 4.

---

**Algorithm 4** Pairwise Elimination-based Agnostic Learner

---

```

initialize( $D, \mathcal{H}, \epsilon$ )
   $N := |\mathcal{H}|, I := \{1, \dots, N\}$ 
   $n_{i,j} := 0, s_{i,j} = 0, \forall i, j \in I$ 
   $m := \lceil \frac{2K^2}{\epsilon^2} \log \frac{2(N-1)}{\delta} \rceil$ 

predict( $x$ )
   $Y := \bigcap_{i \in I} B_{D+\epsilon}(f_i(x))$ 
  if  $Y \neq \emptyset$  then
    return an arbitrary  $\hat{y} \in Y$ 
  else
    return  $\perp$ 

```

---

```

learn( $x, y$ )
  for all  $i, j \in I$  such that  $f_i(x) \ll f_j(x)$  do
     $n_{i,j} := n_{i,j} + 1$ 
     $s_{i,j} := s_{i,j} + (f_i(x) + f_j(x))/2 - y$ 
    if  $n_{i,j} = m$  then
      if  $s_{i,j} < 0$  then
         $I := I \setminus \{i\}$ 
      else
         $I := I \setminus \{j\}$ 

```

---

The following theorem holds true for this algorithm:

**Theorem 4.1** *Let  $\mathcal{H}$  be a finite hypothesis class over  $(\mathcal{X}, \mathbb{R})$ ,  $D, \epsilon > 0$ ,  $0 \leq \delta \leq 1$ ,  $r = 2$ . Then, for any  $\mathcal{G}$  problem class such that the noise in the responses lies in  $[-K, K]$  and  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$  it holds that the pairwise elimination based agnostic learner is an agnostic  $(D, r, \epsilon, \delta)$  KWIK-learner for  $(\mathcal{H}, \mathcal{G})$  with KWIK-bound  $((\lceil \frac{2K^2}{\epsilon^2} \log \frac{2(N-1)}{\delta} \rceil - 1)N + 1)(N - 1) = O\left(\frac{K^2 N^2}{\epsilon^2} \log \frac{N}{\delta}\right)$ .*

## 4.2. The case of infinite hypothesis classes

Note that by introducing an appropriate discretization, the algorithm can also be applied to problems when the hypothesis set is infinite. In particular, given  $\epsilon > 0$ , if there exists  $N > 0$  and  $\mathcal{H}_N \subset \mathcal{H}$  with  $n$  functions such that for any function  $f \in \mathcal{H}$  there exists some function  $f' \in \mathcal{H}_N$  such that  $\|f - f'\|_\infty \leq \epsilon/2$  then if we run the above algorithm with  $\mathcal{H}_N$  and  $\epsilon/2$  (instead of  $\epsilon$ ) then from  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$  it follows that  $\Delta(\mathcal{G}, \mathcal{H}_N) \leq D + \epsilon/2$ . Therefore, by the above theorem, the pairwise elimination based agnostic learner with  $\mathcal{H}_N$  will be  $2D + \epsilon$  accurate and will pass at most  $O\left(\frac{K^2 N^2}{\epsilon^2} \log \frac{N}{\delta}\right)$  times, outside of an event of probability at most  $\delta$ . Therefore, it is a  $(D, r, \epsilon, \delta)$  agnostic KWIK-learner for  $(\mathcal{H}, \mathcal{G})$ . In the case of a linear hypothesis class with a  $d$ -dimensional input,  $N = \Theta((1/\epsilon)^d)$  and thus the number of passes in the bound scales with  $(1/\epsilon)^{2d+2}$ . We see that as compared to the noise-free case, we lose a factor of two in the exponent. The approach just described is general, but the complexity explodes with the dimension. It remains to be seen if there exists alternative, more efficient algorithms.

## 5. Reinforcement learning with KWIK-learning

In reinforcement learning an agent is interacting with its environment by executing actions and observing states and rewards of the environment, the goal being to collect as much reward as possible. Here we consider the case when the state transitions are Markovian. An agent unfamiliar with its environment must spend some time exploring the environment, or it may miss essential information and loose a lot reward in the long run. However, with time the agent must reduce exploration, or it will fail to collect reward. The basic question is how to find the right balance between exploration and exploitation.

The KWIK-Rmax construction of Li et al. (2011a) shows that if efficient KWIK-learning is possible for some environment models then efficient reinforcement learning is possible for the same class. The purpose of this section is to show that this result readily extends to the agnostic case in a sensible manner, justifying the choice of the agnostic learning model proposed.

### 5.1. Markovian Decision Processes and efficient learning agents

In this section we introduce a minimal formal framework for studying the efficiency of reinforcement learning algorithms when they are used to learn to control Markovian Decision Processes (MDPs). For further information on learning in MDPs the reader is referred to the book of Szepesvari (2010) and the references therein.

Technically, an MDP  $M$  is a triple  $(S, A, \mathcal{P})$ , where  $S$  is the set of states,  $A$  is the set of actions, both are non-empty, Borel-spaces; and  $\mathcal{P}$ , determining the evolution of the decision process, is a transition probability map from  $S \times A$  to  $S \times \mathbb{R}$ .<sup>6</sup> In particular, an agent interacting with an environment described by an MDP receives at time step the state  $s_t \in S$  of the environment, decides about the action  $a_t \in A$  to take based on the information available to it and then executes the action in the environment. As a result the environment moves the state and generates the reward associated with the transition:  $(r_t, s_{t+1}) \sim P(\cdot|s_t, a_t)$ . The process then repeats. An algorithm (which may use randomness) for computing an action based on past information is called a (non-stationary) policy. The value of a policy  $\pi$  in state  $s$  is the expected total discounted reward, or expected return when the interaction starts at state  $s$ . Formally,

$$V^\pi(s) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid (r_t, s_{t+1}) \sim P(\cdot|s_t, a_t), a_t \sim \pi(\cdot|s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_t), t \geq 0, s_0 = s \right].$$

An *optimal policy*  $\pi^*$  is such that in every state  $s \in S$ ,  $V^{\pi^*}(s)$  is the best possible value:  $V^{\pi^*}(s) = V^*(s) \stackrel{\text{def}}{=} \sup_\pi V^\pi(s)$ . Here,  $V^*$  is the so-called *optimal value function* and  $V^\pi$  is

---

6. We call  $\mathcal{P}$  a transition probability map between measurable spaces  $(E_1, \mathcal{E}_1)$  and  $(E_2, \mathcal{E}_2)$  if (i)  $\mathcal{P}(\cdot|e_2)$  is a probability measure on  $(E_1, \mathcal{E}_1)$  for any  $e_2 \in E_2$  and (ii) the function  $\mathcal{P}(B|\cdot)$  is measurable on  $E_2$  for any  $B \in \mathcal{E}_1$ . In what follows, to minimize clutter, we omit technical assumptions needed to establish e.g. the existence of measurable optimal stationary policies. See e.g. Theorem 6.11.11 in the book by Puterman (2005) for a compact result and the references in this book. All results presented hold for finite MDPs without any further assumptions.

called the *value function* of policy  $\pi$ . In finite discounted MDPs an optimal policy always exists. We also need the *optimal action-value function*  $Q^* : S \times A \rightarrow \mathbb{R}$ ;  $Q^*(s, a)$  is defined as the total expected discounted reward assuming that the interaction starts at state  $s$ , the first action is  $a$  and in the subsequent timesteps an optimal policy is followed. We will write  $V_M^\pi$ ,  $V_M^*$ ,  $Q_M^*$  when there is a need to emphasize that these objects are specific to the MDP  $M$ .

A learning agent's goal is to act “near-optimally” in every finite MDP while having little *a priori* information about the MDP. In particular, given a finite MDP  $M = (S, A, \mathcal{P})$ , the learning agent  $\mathcal{A}$  is told  $S, A$ , a bound on the rewards and their expected value, a discount factor  $0 < \gamma < 1$ . Then,  $\mathcal{A}$  starts interacting with an environment described by  $M$ . Note that a learning agent is slightly more general than a policy: A policy is MDP-specific (by definition), while a learning agent must be able to act in any (finite) MDP. We can also view learning as the learning agent *choosing* an appropriate (non-stationary) policy to follow based on the *a priori* information received. We also identify an agent  $\mathcal{A}$  with its “learning algorithm” and will also say “algorithm  $\mathcal{A}$ ”.

One possible goal for a learning agent is to minimize the number of timesteps when the future value collected from the state just visited is worse than the optimal value less some value  $\epsilon > 0$ . Following Kakade (2003) and Strehl and Littman (2005), we formalize this as follows:

**Definition 5.1 ( $\epsilon$ -mistake count)** Let  $\epsilon > 0$  be a prescribed accuracy. Assume that a learning agent  $\mathcal{A}$  interacts with an MDP  $M$  and let  $(s_t, a_t, r_t)_{t \geq 0}$  be the resulting  $S \times A \times \mathbb{R}$ -valued stochastic process. Define the expected future return of  $\mathcal{A}$  at time step  $t \geq 0$  as

$$V_{t,M}^{\mathcal{A}} = \mathbb{E} \left[ \sum_{s=0}^{\infty} \gamma^s r_{t+s} \mid s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_t \right].$$

Agent  $\mathcal{A}$  is said to make an  $\epsilon$ -mistake at time step if  $V_{t,M}^{\mathcal{A}} < V_M^*(s_t) - \epsilon$  and we will use  $N_{M,\epsilon}^{\mathcal{A}}$  to denote the number of  $\epsilon$ -mistakes agent  $\mathcal{A}$  makes in  $M$ :

$$N_{M,\epsilon}^{\mathcal{A}} = \sum_{t=0}^{\infty} \mathbb{I}_{\{V_{t,M}^{\mathcal{A}} < V_M^*(s_t) - \epsilon\}}.$$

The competence of an algorithm  $\mathcal{A}$  is measured by  $N_{M,\epsilon}^{\mathcal{A}}$ .<sup>7</sup> Formally, an algorithm  $\mathcal{A}$  is called *PAC-MDP* if for any  $\epsilon > 0$ ,  $0 < \delta \leq 1$ , MDP  $M$ ,  $N_{M,\epsilon}^{\mathcal{A}}$  can be bounded with probability  $1 - \delta$  with a polynomial of the form  $\text{poly}(|S|, |A|, 1/\epsilon, \log(1/\delta), 1/(1 - \gamma))$ , assuming that the rewards belong to  $[0, 1]$  interval. For MDPs with infinite state and action spaces,  $|S|$  and  $|A|$  should be replaced by an appropriate “complexity” measure.

7. An alternative, closely related way for measuring competence is to count the number of timesteps when  $Q_M^*(x_t, a_t) < V_M^*(x_t) - \epsilon$ :  $\hat{N}_{M,\epsilon}^{\mathcal{A}} = \sum_{t=0}^{\infty} \mathbb{I}_{\{Q_M^*(s_t, a_t) < V_M^*(s_t) - \epsilon\}}$ . If the learning agent does not randomize,  $Q_M^*(s_t, a_t) > V_{t,M}^{\mathcal{A}}$  follows from the definitions. It follows then that  $N_{M,\epsilon}^{\mathcal{A}} \leq \hat{N}_{M,\epsilon}^{\mathcal{A}}$  holds almost surely. For further, alternative notions of efficient learning consult Fiechter (1994); Auer et al. (2008).

From a static, non-learning, computational viewpoint, the stochasticity of the rewards does not play a role. Hence, by slightly abusing notation, we will also call a 4-tuple  $(S, A, P, R)$  an MDP, where  $P$  is a transition probability map from  $S \times A$  to  $S$ , and  $R : S \times A \rightarrow \mathbb{R}$  is the immediate expected reward function underlying  $\mathcal{P}$ . A policy  $\pi$  which chooses the actions based on the last state only in a fixed manner is called a *stationary (Markov) policy*.<sup>8</sup> Such a policy can and will be identified with a transition probability map from  $S$  to  $A$ . In particular, we will use  $\pi(\cdot|s)$  to denote the probability distribution over  $A$  designated by  $\pi$  at  $s \in S$ .

## 5.2. A general theorem on efficient RL

In this section we show a general result on constructing efficient reinforcement learning algorithms. The result states that if a policy *i*) keeps track whether state-action pairs are “known”, *ii*) ensures that the model parameters corresponding to “known” pairs are indeed known with reasonable accuracy, *iii*) makes an effort to get to unknown areas, then it will be efficient as soon as the number of times it happens that the currently visited state-action pair is not “known” is small.

Theorems of this style have appeared in Strehl et al. (2006, 2009); Li (2009); Li et al. (2011b,a). These results in fact generalize the proof of Kakade (2003) which shows that RMAX is efficient. The closest in spirit to the result below is Theorem 10 by Strehl et al. 2009 (originally appeared as Proposition 1 of Strehl et al. (2006) and then repeated as Theorem 4 by Li et al. 2011b). The differences between the result below and this theorem are mostly at the cosmetic level, although the particular form below makes our theorem particularly easy to apply to the model-based setting of the next section and it also allows imperfect (even stochastic) planners.<sup>9</sup> However, the main reason we included this result is to give a fully rigorous proof, which avoids the inaccuracies and ambiguities of previous proofs.<sup>10</sup> Also, we slightly improve the previous bounds: we remove a  $1/(1 - \gamma)$ -factor.

Let  $\mathcal{Y} = M(S) \times \mathbb{R}$ , where  $M(S)$  is the space of finite signed-measures. We define the norm  $\|\cdot\|_{\mathcal{Y}} : \mathcal{Y} \rightarrow \mathbb{R}_+$  as follows: For any  $P \in M(S)$ ,  $r \in \mathbb{R}$ , let  $\|(P, r)\|_{\mathcal{Y}} \stackrel{\text{def}}{=} |r| + \gamma \|P\|_{TV} V_{\max}$ , where  $V_{\max} > 0$  will be chosen to be a common upper bound on the value functions in a class of MDPs of interest and  $\|\cdot\|_{TV}$  is the total-variation norm of finite signed measures ( $\|P\|_{TV} = |P|(S)$ ). Define the distance of MDPs  $M_1 = (S, A, P_1, R_1)$ ,  $M_2 = (S, A, P_2, R_2)$  sharing  $S$  and  $A$  by

$$d(M_1, M_2) \stackrel{\text{def}}{=} \sup_{(s,a) \in S \times A} \|(P_1(\cdot|s, a) - P_2(\cdot|s, a), r_1(s, a) - r_2(s, a))\|_{\mathcal{Y}}.$$

- 
- 8. A policy is Markov if the distribution assigned to a history depends only on the last state of the history. In what follows, by a stationary policy we will always mean a stationary Markov policy.
  - 9. Imperfect planners are also allowed in Walsh et al. (2010).
  - 10. Some of the differences in our proof follow from our slightly changed assumptions. However, the proof of Lemma 5.3 is new. In place of our argument, earlier works either did not present a proof or used an argument which did not allow serial correlations and thus, strictly speaking, cannot be applied in the setting considered here.

**Theorem 5.1** Consider agent  $\mathcal{A}$  interacting with an MDP  $M = (S, A, \mathcal{P})$ . Let  $(s_1, a_1, r_1, s_2, a_2, r_2, \dots)$  be the trajectory that results when  $\mathcal{A}$  interacts with  $M$  and let  $\mathcal{F}_t = \sigma(s_1, a_1, r_1, \dots, s_t)$  be the  $\sigma$ -field corresponding to the history at time  $t$ . Let  $G \subset \Omega$  be a measurable event of the probability space that holds the random variables  $(s_t, a_t, r_t)_{t \geq 1}$ . Assume that there exist a sequence of state-action sets  $(K_t)_{t \geq 1}$ ,  $K_t \subseteq S \times A$ , a sequence of models  $(\hat{M}_t)_{t \geq 1}$ , and a sequence of stationary policies  $(\pi_t)_{t \geq 1}$  such that for some  $e_{\text{plan}}, e_{\text{model}}, V_{\max} > 0$  and for any  $t \geq 1$  the following hold:

- (a) the expected immediate rewards underlying  $M$  and  $(\hat{M}_t)_{t \geq 1}$  are bounded by  $(1 - \gamma)V_{\max}$ ;
- (b)  $K_t, \hat{M}_t, \pi_t$  are  $\mathcal{F}_t$ -measurable;
- (c)  $a_t = \pi_t(s_t)$  (the action at time  $t$  is selected by  $\pi_t$ );
- (d)  $V_{\hat{M}_t}^{\pi_t}(s_t) \geq V_{\hat{M}_t}^*(s_t) - e_{\text{plan}}$  (the policy  $\pi_t$  is at least  $e_{\text{plan}}$ -optimal at  $s_t$  in model  $\hat{M}_t$ );
- (e) for any  $(s, a) \in K_t$ ,  $\left\| \hat{M}_t(s, a) - M(s, a) \right\|_{\mathcal{Y}} \leq e_{\text{model}}$  holds true on  $G$  (the model is  $e_{\text{model}}$ -accurate for “known” state-action pairs);
- (f) for any  $(s, a) \notin K_t$ , any stationary policy  $\pi$ ,  $V_{\max} \leq Q_{\hat{M}_t}^\pi(s, a)$  (in “unknown” states, the model is optimistic, but not overly so); and
- (g) if  $(s_t, a_t) \in K_t$  then  $\pi_{t+1}(s_{t+1}) = \pi_t(s_{t+1})$  (old policy used as long as visiting known state-action pairs).

Let  $B$  be a deterministic upper bound on the number of times it happens that  $(s_t, a_t) \notin K_t$  on  $G$ :  $\sum_{t=1}^{\infty} \mathbb{I}_{\{(s_t, a_t) \notin K_t, G\}} \leq B \mathbb{I}_{\{G\}}$ . Then, for any  $0 < \delta \leq 1$  there exists an event  $F = F_\delta$  such that  $\mathbb{P}(F_\delta) \geq 1 - \delta$  and on  $F \cap G$ , the number of  $5e_{\text{model}}/(1 - \gamma) + e_{\text{plan}}$ -mistakes is bounded by  $\frac{2V_{\max}(1-\gamma)L}{e_{\text{model}}} \left\{ B + (\sqrt{2B} + 3)\sqrt{\log(\frac{L}{\delta})} + 6\log(\frac{L}{\delta}) \right\}$ , where  $L = \max(1, \lceil (1 - \gamma)^{-1} \log(V_{\max}(1 - \gamma)/e_{\text{model}}) \rceil)$ .

We will need two lemmas for the proof, which we state first. The first lemma is a standard result which follows from simple contraction arguments:

**Lemma 5.2 (Simulation Lemma)** For any two MDPs  $M_1$  and  $M_2$  sharing the same state-action set  $S \times A$  and any stationary policy  $\pi$  over  $(S, A)$ , it holds that  $\|V_{M_1}^\pi - V_{M_2}^\pi\|_\infty \leq d(M_1, M_2)/(1 - \gamma)$ .

The next lemma compares the number of times the bias underlying an infinite sequence of coin flips can exceed a certain threshold  $\epsilon$  given that the biases are sequentially chosen and that the number of heads in the infinite coin flip sequence assumes some bound  $m$ . The proof is based on the idea underlying Markov’s inequality and a stopping time construction used in conjunction with a Bernstein-like inequality due to Freedman (1975).

**Lemma 5.3** Let  $0 < \epsilon < 1$ ,  $m \in \mathbb{N}$  be deterministic constants,  $(\mathcal{F}_t)_{t \geq 1}$  be some filtration and let  $(A_t)_{t \geq 1}$  be an  $(\mathcal{F}_{t+1})_{t \geq 1}$ -adapted sequence of indicator variables. Let

$$a_t = \mathbb{E}[A_t | \mathcal{F}_t]$$

and let  $G$  be an event such that on  $G$  the inequality  $\sum_{t=1}^{\infty} A_t \leq m$  holds almost surely. Then, for any  $0 < \delta \leq 1$  with probability  $1 - \delta$ , either  $G^c$  holds, or

$$\sum_{t=1}^{\infty} \mathbb{I}_{\{a_t \geq \epsilon\}} \leq \frac{1}{\epsilon} \left\{ m + \sqrt{2m \log(\frac{1}{\delta})} + 3\sqrt{\log(\frac{1}{\delta})} + 6\log(\frac{1}{\delta}) \right\}.$$

With this, we are ready to give the proof of Theorem 5.1:

**Proof** We need to show that with an appropriate choice of  $F$ , and for  $\epsilon \stackrel{\text{def}}{=} 5e_{\text{model}}/(1 - \gamma) + e_{\text{plan}}$ , the inequality  $N_{M,\epsilon}^{\mathcal{A}} \leq \frac{2V_{\max}(1-\gamma)L}{e_{\text{model}}} \left\{ B + (\sqrt{2B} + 3)\sqrt{\log(\frac{L}{\delta})} + 6\log(\frac{L}{\delta}) \right\}$  holds on  $F \cap G$ .

Let  $e_{\text{trunc}} \stackrel{\text{def}}{=} e_{\text{model}}/(1 - \gamma)$  be the allowed truncation-error. Note that with this notation,  $L = \max(1, \lceil \frac{1}{1-\gamma} \log \frac{V_{\max}}{e_{\text{trunc}}} \rceil)$ . The quantity  $L$  is known as the so-called  $e_{\text{trunc}}$ -horizon: if  $V_M^{\pi}(s; L)$  denotes the expected  $L$ -step return of  $\pi$  in  $M$  when the decision process starts at  $s$  then  $|V_M^{\pi}(s) - V_M^{\pi}(s; L)| \leq e_{\text{trunc}}$ .

Fix  $t \geq 1$  and let  $E_t$  be the event that in the next  $L$  steps, the agent “escapes” the known set:

$$E_t = \bigcup_{i=0}^{L-1} \{(s_{t+i}, a_{t+i}) \notin K_{t+i}\}. \quad (2)$$

The plan for the proof is as follows: We first show that whenever  $p_t = \mathbb{P}(E_t | \mathcal{F}_t)$  is small, in particular, when

$$p_t \leq \frac{e_{\text{trunc}}}{2V_{\max}} \quad (3)$$

then, on  $G$ , the agent does not make an  $\epsilon$ -mistake at time  $t$ . Then, based on Lemma 5.3 we will give a high-probability bound on the number of timesteps when (3) fails to hold.

Turning to the first step, assume that (3) holds. We want to show that the agent does not make an  $\epsilon$ -mistake on  $G$ , i.e., on this event,  $V_{t,M}^{\mathcal{A}} \geq V_M^*(s_t) - \epsilon$ . Let  $\tilde{\pi}_t$  be the non-stationary policy of  $M$  induced by  $\mathcal{A}$  at time step  $t$ . Then,  $V_{t,M}^{\mathcal{A}} = V_M^{\tilde{\pi}_t}(s_t) \geq V_M^{\tilde{\pi}_t}(s_t; L) - e_{\text{trunc}}$  by the choice of  $L$ . Let  $\hat{M}_t$  be the model which agrees with  $M$  on  $K_t$ , while it agrees with  $\hat{M}_t$  outside of  $K_t$ .

**Claim 5.4** We have  $V_M^{\tilde{\pi}_t}(s_t; L) \geq V_{\hat{M}_t}^{\pi_t}(s_t; L) - 2V_{\max}p_t$ , with probability one.

The proof, which is given in the appendix, uses that the immediate rewards for both  $M$  and  $\hat{M}_t$  are bounded by  $(1 - \gamma)V_{\max}$  (i.e., condition (a)), the measurability condition (b) and the condition that there is no policy update while visiting known states (i.e., condition (g)).

Now, by the definition of  $L$ ,  $V_{\bar{M}_t}^{\pi_t}(s_t; L) \geq V_{\bar{M}_t}^{\pi_t}(s_t) - e_{\text{trunc}}$ , while by the Simulation Lemma (Lemma 5.2) and (e), on  $G$ , it holds that  $V_{\bar{M}_t}^{\pi_t}(s_t) \geq V_{\hat{M}_t}^{\pi_t}(s_t) - e_{\text{model}}/(1-\gamma)$ . By (d),  $V_{\hat{M}_t}^{\pi_t}(s_t) \geq V_{\hat{M}_t}^*(s_t) - e_{\text{plan}}$ . Let  $\pi^*$  be an optimal stationary policy in  $M$ . We have  $V_{\hat{M}_t}^*(s_t) \geq V_{\hat{M}_t}^{\pi^*}(s_t)$ .

Now, let  $\tilde{M}_t$  be the MDP which is identical to  $\hat{M}_t$  for state-action pairs in  $K_t$ , while outside of  $K_t$  it is identical to  $M$ . We claim that the following holds true:

**Claim 5.5** *On  $G$ , it holds that  $V_{\tilde{M}_t}^{\pi^*}(s_t) \geq V_{\hat{M}_t}^{\pi^*}(s_t)$ .*

The proof, which is again given in the appendix, uses the optimism condition (condition (f)), condition (a).

By the Simulation Lemma and (e), on  $G$ , it holds that  $V_{\tilde{M}_t}^{\pi^*}(s_t) \geq V_M^{\pi^*}(s_t) - e_{\text{model}}/(1-\gamma)$ . Chaining the inequalities obtained, we get that, on  $G$ , the inequality  $V_{t,M}^A \geq V_M^{\pi^*}(s_t) - (2p_t V_{\max} + 4e_{\text{trunc}} + e_{\text{plan}})$  holds. Thus, when (3) holds, on  $G$ , we also have  $V_{t,M}^A \geq V_M^{\pi^*}(s_t) - (5e_{\text{trunc}} + e_{\text{plan}}) = V_M^{\pi^*}(s_t) - \epsilon$ , which concludes the proof of the first step.

Let us now turn to the second step of the proof. By the first step, on  $G$ ,  $\sum_{t=1}^{\infty} \mathbb{I}_{\{V_{t,M}^A < V_M^* - \epsilon\}} \leq \sum_{t=1}^{\infty} \mathbb{I}_{\{p_t < \frac{e_{\text{trunc}}}{1-\gamma}\}}$ . Let  $T_{\text{non-opt}} = \sum_{t=1}^{\infty} \mathbb{I}_{\{p_t > e_{\text{trunc}}/(2V_{\max})\}}$ . In order to bound  $T_{\text{non-opt}}$ , we write it as the sum of  $L$  terms as follows

$$T_{\text{non-opt}} = \underbrace{\sum_{i=0}^{L-1} \sum_{j=0}^{\infty} \mathbb{I}_{\{p_{jL+i+1} > e_{\text{trunc}}/(2V_{\max})\}}}_{T_{\text{non-opt}}^{(i)}}. \quad (4)$$

We will apply Lemma 5.3 to each of the resulting  $L$  terms separately. To do so, fix  $0 \leq i \leq L-1$  and choose the sequence of random variables  $(A_t)_{t \geq 0}$  to be  $(\mathbb{I}_{\{E_{tL+i+1}\}})_{t \geq 0}$ , while let the corresponding sequence of  $\sigma$ -fields be  $(\mathcal{F}_{tL+i+1})_{t \geq 0}$ . Further, choose  $\epsilon$  of Lemma 5.3 as  $\epsilon = e_{\text{trunc}}/(2V_{\max})$ . By condition (b),  $A_t$  is  $\mathcal{F}_{(t+1)L+i+1}$ -measurable, since  $E_{tL+i+1} \in \mathcal{F}_{(t+1)L+i+1}$ . The upper bound  $m$  on the sum  $\sum_t A_t$  is obtained from

$$\sum_{j=0}^{\infty} \mathbb{I}_{\{E_{jL+i+1}\}} \leq \sum_{t=1}^{\infty} \mathbb{I}_{\{E_t\}} \leq \sum_{t=1}^{\infty} \mathbb{I}_{\{(x_t, a_t) \notin K_t\}},$$

where the last inequality follows from the definition of  $E_t$  (cf. (2)). By assumption, on  $G$ , the last expression is bounded by  $B$ . Therefore, on  $G$ ,  $\sum_{j=0}^{\infty} \mathbb{I}_{\{E_{jL+i+1}\}} \leq B$  also holds and Lemma 5.3 gives that with probability  $1 - \delta/L$ , either  $G^c$  holds or

$$T_{\text{non-opt}}^{(i)} \leq \frac{2V_{\max}}{e_{\text{trunc}}} \left\{ B + \sqrt{2B \log \left(\frac{L}{\delta}\right)} + 3\sqrt{\log \left(\frac{L}{\delta}\right)} + 6\log \left(\frac{L}{\delta}\right) \right\}.$$

Combining this with (4) gives that, with probability  $1 - \delta$ , either  $G^c$  holds or

$$T_{\text{non-opt}} \leq \frac{2V_{\max}L}{e_{\text{trunc}}} \left\{ B + (\sqrt{2B} + 3)\sqrt{\log \left(\frac{L}{\delta}\right)} + 6\log \left(\frac{L}{\delta}\right) \right\},$$

■

thus, finishing the proof.

### 5.3. The KWIK-Rmax construction

In this section we consider the KWIK-RMAX algorithm of Li et al. (2011a) (see, also Li et al. (2011b)). This algorithm is identical to the RMAX algorithm of Brafman and Tennenholtz (2000), except that the model learning and planning components of RMAX are replaced by general components. This way one gets a whole family of algorithms, depending on what model learner and planner is used. In addition to unifying a large number of previous works which considered different model learners and planners (for a list of these, see the introduction of this article), this allowed Li et al. to improve the previously known efficiency bounds, too (see, Chapter 7 of Li (2009) and Li et al. (2011a)).

Here, the exact same algorithm is considered, but we derive a more general result: We show that if the model learning algorithm is an *agnostic* KWIK-learner enjoying some KWIK-bound and a “good” planner is used then the resulting instance of KWIK-RMAX will be efficient even when the MDP considered is outside of the hypothesis class that the KWIK-learner uses. In essence, our analysis shows how approximation errors propagate in a reinforcement learning context. In the special case of realizable learning, our result reproduces the result of Li et al. (2011a) (our bound is slightly better in terms of its dependence on the discount factor  $\gamma$ ).

The KWIK-RMAX algorithm is shown as Algorithm 5. The algorithm takes as input two “objects”, a learner (MDPLearner) and a planner (Planner). The learner’s job is to learn an approximation to the MDP that KWIK-RMAX interacts with. The learned model is fed to the planner. The planner is assumed to interact with models by querying next state distributions and immediate rewards at select certain state-action pairs. The **predict** method of a model is assumed to return the returned values for the planner. The planner itself could use these in many ways – the details of the planning mechanism are not of our concern here (our result allows for both deterministic and stochastic planners). An important aspect of the algorithm is that the model learned is not actually fed directly to the planner, but it is fed to a wrapper. In fact, since a KWIK learner might pass in any round, it is the job of the wrapper to produce a next-state distribution, reward pair in all cases. This can be done in many different ways. However, the main idea here is to return a next-state distribution and a reward, which makes an “unknown” state-action pair highly desirable. One implementation of this is shown in the right-hand side of algorithm listing 5. The KWIK-RMAX algorithm itself repeatedly calls the planner (with the optimistically wrapped model learned by the KWIK learner), executes the returned action and upon observing the next state and reward, if the KWIK learner passed, it feeds the learner with the observed values. Note that for the analysis it is critical that the learner is not fed with information when it did not pass, contradicting one’s intuition.

Our main result, stated below, says that if MDPLearner is an agnostic KWIK-learner and the planner is near-optimal then KWIK-RMAX will be an efficient RL algorithm. In

**Algorithm 5** The KWIK-Rmax Algorithm and the Optimistic Wrapper

---

<b>KWIK-Rmax</b> (MDPLearner, Planner)	{Optimistic Wrapper}
MDPLearner.initialize(...)	
Planner.initialize(...)	<b>Opt</b> (MDPLearner). <b>predict</b> ( $s, a$ )
Observe $s_1$	<b>if</b> MDPLearner.predict( $s, a$ ) = $\perp$ <b>then</b>
<b>for</b> $t := 1, 2, \dots$ <b>do</b>	<b>return</b> $(\delta_s(\cdot), (1 - \gamma)V_{\max})$
$a_t = \text{Planner.plan(OPT(MDPLearner), } s_t)$	<b>else</b>
Execute $a_t$ and observe $s_{t+1}, r_t$	<b>return</b> MDPLearner.predict( $s, a$ )
<b>if</b> MDPLearner.predict( $s_t, a_t$ ) = $\perp$ <b>then</b>	
MDPLearner.learn(( $s_t, a_t$ ), ( $\delta_{s_{t+1}}, r_t$ ))	

---

order to state this result formally, first we need to define what we mean by KWIK-learning in the context of MDPs.

As before, we fix the set of states ( $S$ ) and actions ( $A$ ) and  $V_{\max} > 0$ , an upper bound on the value functions for the MDPs of interest. Remember that  $\mathcal{Y} = M(S) \times \mathbb{R}$  and the norm  $\|\cdot\|_{\mathcal{Y}}$  defined by  $\|(P, r)\|_{\mathcal{Y}} = |r| + \gamma \|P\|_{TV} V_{\max}$  ( $P \in M(S)$ ,  $r \in \mathbb{R}$ ). The space  $\mathcal{Y}$  will be the output space for the predictors. Learning an MDP model means learning the immediate expected rewards and transition probabilities when the inputs are state-action pairs. That is, we let  $\mathcal{X} = S \times A$  and encode an MDP as a mapping  $g : \mathcal{X} \rightarrow \mathcal{Y}$ , where for  $x = (s, a) \in \mathcal{X}$ ,  $g(s, a) = (P, r)$ , where  $P$  is the next-state distribution over  $S$  and  $r \in \mathbb{R}$  is the associated expected immediate reward. What is left in specifying an *MDP problem instance*  $(\mathcal{X}, \mathcal{Y}, g, Z, \|\cdot\|)$  is the noise component  $Z$ . In the  $\mathbb{R}^S$  component of  $\mathcal{Y}$ , the noise is completely determined by  $g$ , while, for the reward component the noise distribution is arbitrary, except that the noisy reward is restricted to be bounded by  $(1 - \gamma)V_{\max}$ . A subset of all  $g : \mathcal{X} \rightarrow \mathcal{Y}$  functions will be called an MDP hypothesis space. Now, we are ready to state our main result.

**Theorem 5.6** Fix a state space  $S$  and an action space  $A$ , which are assumed to be non-empty Borel spaces. Let  $\mathcal{X}, \mathcal{Y}$  be as described above,  $\mathcal{H}$  be an MDP hypothesis set,  $\mathcal{G}$  be a set of MDP problem instances, both over  $\mathcal{X}, \mathcal{Y}$ . Assume that  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$ . Assume that  $V_{\max} > 0$  is such that  $(1 - \gamma)V_{\max}$  is an upper bound on the immediate rewards of the MDPs determined by members of  $\mathcal{H}$  and  $\mathcal{G}$ . Fix  $\epsilon > 0, r \geq 1, 0 < \delta \leq 1/2$ . Assume that MDPLearner is an agnostic  $(D, r, \epsilon)$  KWIK-learner for  $(\mathcal{H}, \mathcal{G})$  with KWIK-bound  $B(\delta)$ . Assume further that we are given a Planner which is  $e_{\text{planner}}$ -accurate. Consider the instance of the KWIK-RMAX algorithm which uses MDPLearner and Planner, interacting with some MDP  $M$  from  $\mathcal{G}$ . Let  $\epsilon' = \frac{5(rD+\epsilon)}{1-\gamma} + e_{\text{planner}}$ . Then, with probability  $1 - 2\delta$ , the number of  $\epsilon'$ -mistakes,  $N_{M, \epsilon'}$ , made by KWIK-RMAX is bounded by  $\frac{2V_{\max}(1-\gamma)L}{rD+\epsilon} \left\{ B(\delta) + (\sqrt{2B(\delta)} + 3) \sqrt{\log\left(\frac{L}{\delta}\right)} + 6 \log\left(\frac{L}{\delta}\right) \right\}$ , where  $L = \max(1, \lceil (1-\gamma)^{-1} \log(V_{\max}(1-\gamma)/(rD+\epsilon)) \rceil)$ .

## 6. Discussion

In the first part of the paper we formalized and explored the agnostic KWIK framework (first mentioned in Li (2009)), and presented several simple agnostic KWIK learning algorithms for finite hypothesis classes with and without noise, and for deterministic linear hypothesis classes. In the second part of the paper we showed that an agnostic KWIK-learner leads to an efficient reinforcement learning, even when the environment is outside of the hypothesis class that the KWIK-learner uses. To our knowledge, this is the first result that proves any kind of efficiency for an RL algorithm in the agnostic setting. Our bound also saves a factor of  $1/(1 - \gamma)$  compared to previous bounds. Unfortunately, our (limited) exploration of agnostic KWIK-learning indicated that efficient agnostic KWIK-learning might be impossible for some of the most interesting (simple) hypothesis classes. These negative results do not imply that efficient agnostic reinforcement learning is impossible, but indicate that the problem itself requires further work.

## Acknowledgments

This work was supported in part by AICML, AITF (formerly iCore and AIF), NSERC and the PASCAL2 Network of Excellence under EC grant no. 216886. We thank Yaoliang Yu for his careful reading of some parts of this paper, in particular for his suggestions that led to a simpler proof of Lemma A.1.

## Appendix A. Proofs

### A.1. Proofs for Section 3

**Theorem 3.1** *Let  $(\mathcal{X}, \mathcal{Y})$  be arbitrary sets,  $r = 2$ ,  $\epsilon = 0$ ,  $D > 0$ ,  $\mathcal{H}$  a finite hypothesis class over  $(\mathcal{X}, \mathcal{Y})$ ,  $\mathcal{G}$  a deterministic problem class over  $(\mathcal{X}, \mathcal{Y})$  with  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$ . Then, the Generic Agnostic Learner is an agnostic  $(D, r, \epsilon)$  KWIK-learner for  $(\mathcal{H}, \mathcal{G})$  with KWIK-bound  $|\mathcal{H}| - 1$ .*

**Proof** Suppose that the adversary chose problem  $G = (\mathcal{X}, \mathcal{Y}, g, 0)$ . Since  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$  and  $\mathcal{H}$  is finite, there exists a function  $f^* \in \mathcal{H}$  such that  $\|f^* - g\|_\infty = \Delta(G, \mathcal{H}) \leq D$ . Consequently,  $f^*$  will be never excluded from  $\mathcal{F}$ : for any  $x$  and  $y = g(x)$ ,  $\|f^*(x) - y\| \leq D$ .

Let us now show that every time the learner makes a prediction, the prediction is  $2D$ -accurate. Indeed, if  $\hat{y}$  is the prediction of the learner,  $\|\hat{y} - y\| \leq \|\hat{y} - f^*(x)\| + \|f^*(x) - y\|$ . Now, by the definition of  $\hat{y}$  and because  $f^*$  is never excluded,  $\|\hat{y} - f^*(x)\| \leq D$ . Further, by the choice of  $f^*$  and because  $y = g(x)$ , we also have  $\|f^*(x) - y\| \leq D$ , altogether showing that the prediction error is upper bounded by  $2D$ .

It remains to show that the learner cannot pass more than  $|\mathcal{H}| - 1$  times. This follows, because after each pass, the learner excludes *at least one* hypothesis. To see why note that  $Y = \emptyset$  means that for every  $y' \in \mathcal{Y}$  there is some  $f \in \mathcal{F}$  such that  $y' \notin B_D(f(x))$ .

Specifically, this also holds for  $y = g(x)$  and thus there is a function  $f \in \mathcal{F}$  such that  $\|f(x) - y\| > D$ . By definition, this function will be eliminated in the update. As the learner eliminates at least one hypothesis from  $\mathcal{F}$  when passing, and  $f^*$  is never eliminated from  $\mathcal{F}$ , the number of passes is at most  $|\mathcal{H}| - 1$ . ■

**Theorem 3.2** Fix any  $D > 0$  and an infinite domain  $\mathcal{X}$ . Then, there exists a finite response set  $\mathcal{Y} \subset \mathbb{R}$ , a two-element hypothesis class  $\mathcal{H}$  and a deterministic problem class  $\mathcal{G}$ , both over  $(\mathcal{X}, \mathcal{Y})$ , that satisfy  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$  such that there is no bounded agnostic  $(D, r, 0)$  KWIK-learner for  $(\mathcal{H}, \mathcal{G})$  with competitiveness factor  $0 \leq r < 2$ .

**Proof** Fix  $D > 0$ . Let  $\mathcal{X}$  be an infinite set,  $x_1, x_2, \dots$  be a sequence of distinct elements in  $\mathcal{X}$ ,  $\mathcal{Y} = \{-2D, -D, 0, D, 2D\}$  and  $\mathcal{H}$  be a two-element set containing  $f_{+D} \equiv D$  and  $f_{-D} \equiv -D$ . For  $n \in \mathbb{N}$ , define the functions

$$g_{n, \pm D}(x) = \begin{cases} \pm 2D, & \text{if } x = x_n; \\ 0, & \text{otherwise} \end{cases}$$

and let  $\mathcal{G}$  be the set of these functions. Clearly,  $\Delta(\mathcal{G}, \mathcal{H}) = D$ . Before picking a hypothesis, the adversary simulates its interaction with the learner. At step  $t$  of the simulation, the adversary asks query  $x_t$ . If the learner passes with probability 1, then  $A$  answers 0. Suppose now that there is some  $t$  when the learner makes some prediction  $\hat{y}_t$  with probability  $p > 0$ . Without loss of generality we may assume that  $\mathbb{P}(\hat{y}_t \geq 0) \geq p/2$ . At this point, the adversary stops the simulation and picks  $g_{t, -D}$ . During the learning process,  $L$  passes to any  $x_i$  with  $i < t$  and gets feedback 0. At time step  $t$ , however, with probability  $p/2$ ,

$$\hat{y}_t - g_{t, -D}(x_t) \geq 0 + 2D,$$

so the learner fails. On the other hand, if the learner always passes then it is not bounded. ■

**Theorem 3.3** Let  $X_{\max} > 0$ ,  $\mathcal{X} = [-X_{\max}, X_{\max}]^d$ ,  $\mathcal{Y} = \mathbb{R}$ ,  $M, D, \epsilon > 0$ ,  $r = 2$ ,  $\mathcal{H} = \mathcal{H}_{lin(M)}$ . Then, for any  $\mathcal{G}$  deterministic problem class over  $(\mathcal{X}, \mathcal{Y})$  with  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$ , it holds that the deterministic linear agnostic learner is an agnostic  $(D, r, \epsilon)$  KWIK-learner for  $(\mathcal{H}, \mathcal{G})$  with the KWIK-bound  $2d! \left( \frac{MX_{\max}}{\epsilon} + 1 \right)^d$ .

**Proof** First of all, we need to show that the calculations of the algorithm are meaningful. Specifically, solutions  $y^+$  and  $y^-$  to the linear programs need to be finite. This will hold because  $\Theta(C)$  is bounded and non-empty during any point of the learning. Boundedness holds because  $\Theta(C) \subseteq \{\theta : \|\theta\|_\infty \leq M\}$ . We assumed that there is a hypothesis  $f_{\theta^*}$  so that  $\|f_{\theta^*} - g^*\|_\infty \leq D$ , and the KWIK protocol sends training samples  $(x, y)$  such that  $y = g^*(x)$ . Therefore,  $D \geq |f_{\theta^*}(x) - g^*(x)| = |(\theta^*)^T x - y|$ , so  $\theta^*$  satisfies all constraints in  $C$ , making  $\Theta(C)$  nonempty.

Secondly, the following calculation shows that if a prediction is made, it is correct:

$$\begin{aligned}\hat{y}_t - y_t &= y_t^+ / 2 + y_t^- / 2 - g(x_t) \\ &\leq D + \epsilon + y_t^- / 2 + y_t^- / 2 - g(x_t) \\ &= D + \epsilon + (y_t^- - f^*(x_t)) + (f^*(x_t) - g(x_t)) \\ &\leq D + \epsilon + 0 + D,\end{aligned}$$

and similarly,

$$\begin{aligned}\hat{y}_t - y_t &= y_t^+ / 2 + y_t^- / 2 - g(x_t) \\ &\geq y_t^+ / 2 + y_t^+ / 2 - (D + \epsilon) - g(x_t) \\ &= -D - \epsilon + (y_t^+ - f^*(x_t)) + (f^*(x_t) - g(x_t)) \\ &\geq -D - \epsilon + 0 - D,\end{aligned}$$

so  $|\hat{y}_t - y_t| \leq 2D + \epsilon$ , as required.

Finally, we prove the upper-bound on the number of  $\perp$ s. Let  $\mathcal{X}' \stackrel{\text{def}}{=} [-X_{\max} - \epsilon/M, X_{\max} + \epsilon/M]^d$  be a slightly increased version of  $\mathcal{X}$ . Let  $H_k$  be the set of  $x \in \mathcal{X}'$  for which the learner makes a prediction, that is,  $y^+(x) - y^-(x) \leq 2D(1 + \epsilon)$ . The set  $H_k$  is convex, following from the convexity of the constraints and the convexity of the max operator (and the concavity of  $\min$ ). If the adversary asks some  $x \notin H_k$ , then  $x$  gets added to the known set, together with some of its neighborhood  $B(x) \stackrel{\text{def}}{=} \{x' \in \mathcal{X}' : \|x' - x\|_1 \leq \epsilon/M\}$ . This follows from the calculation below: let  $x' \in B(x)$ , then for any  $\theta \in \Theta(C_{k+1})$ ,

$$\begin{aligned}\theta^T x' &\leq \theta^T x + \|\theta\|_\infty \|x' - x\|_1 \leq \theta^T x + \epsilon, \text{ so} \\ \max_{\theta \in \Theta(C_{k+1})} \theta^T x' &\leq \max_{\theta \in \Theta(C_{k+1})} \theta^T x + \epsilon, \text{ that is,} \\ y^+(x') &\leq y^+(x) + \epsilon.\end{aligned}$$

With similar reasoning,  $y^-(x') \geq y^-(x) - \epsilon$ . In step  $k+1$ , the constraint  $y - D \leq \theta^T x \leq y + D$  was added for  $x$ , so  $y^+(x) - y^-(x) \leq 2D$ . Consequently,  $y^+(x') - y^-(x') \leq 2(D + \epsilon)$ , so the learner knows  $x'$ .

The set  $H_k$  is convex,  $x \notin H_k$ , and  $B(x)$  is symmetric to  $x$ . So for any  $x' \in B(x)$ , at most one of  $x'$  and  $x + (x - x')$  can be  $\in B(x)$ , that is, at least half of the volume of  $B(x)$  is outside  $H_k$ .  $H_k \cup B(x) \subseteq H_{k+1}$ , so  $\text{Vol}(H_{k+1}) \geq \text{Vol}(H_k) + \text{Vol}(B(x))/2$ . The volume of  $B(x)$  is  $(2\epsilon/M)^d/d!$ , so  $\text{Vol}(H_k) \geq k \frac{(2\epsilon/M)^d}{2d!}$ . On the other hand,  $H_k \subseteq \mathcal{X}'$ , so  $\text{Vol}(H_k) \leq (2X_{\max} + 2\epsilon/M)^d$ , which yields an upper bound on  $k$ :

$$k \leq 2d! \left( \frac{MX_{\max}}{\epsilon} + 1 \right)^d.$$

■

**Theorem 3.4** *Let  $\mathcal{X}, \mathcal{Y}, D, r, \mathcal{H}$  be as in Theorem 3.3. Then, there exists a problem class  $\mathcal{G}$  that satisfies  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$  such that there is no bounded, agnostic  $(D, r, 0)$  KWIK-learner for  $(\mathcal{H}, \mathcal{G})$ .*

**Proof** For any learner  $L$ , we will construct a problem class  $\mathcal{G}$ , a strategy for the adversary so that  $L$  will either make a mistake larger than  $2D$  with nonzero probability, or passes infinitely many times. The construction will be similar to the previous one.

Without loss of generality, let  $\mathcal{X} = [0, 2]$  and let  $\mathcal{Y} = \mathbb{R}$  with the absolute loss norm. Let  $(x_1, x_2, \dots)$  be a strictly increasing sequence of numbers with  $1 < x_t < 2$ ,  $t \in \mathbb{N}$ . For convenience, define  $x_0 = 1$ . For  $n \in \mathbb{N}$ , let

$$g_{n,\pm}(x) \stackrel{\text{def}}{=} \begin{cases} \pm D(1 + \frac{x}{x_{n-1}}) & \text{if } x > x_{n-1}, \\ 0 & \text{if } x \leq x_{n-1} \end{cases}$$

and let  $\mathcal{G}$  be the set of these functions. For this set of problems,  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$ : for any concept, there exists a function  $f \in \mathcal{H}_{\text{lin}(M)}$  that is at most at distance  $D$  from the chosen concept. Specifically, for  $g_{n,+}$ ,  $f_\theta$  with  $\theta = D/x_{n-1}$  is satisfactory:

$$\begin{aligned} \|g_{n,+} - f_{D/x_{n-1}}\|_\infty &= \max \left\{ \max_{x \leq x_{n-1}} |g_{n,+}(x) - f_{D/x_{n-1}}(x)|; \max_{x > x_{n-1}} |g_{n,+}(x) - f_{D/x_{n-1}}(x)| \right\} \\ &= \max \left\{ \max_{x \leq x_{n-1}} |0 - \frac{D}{x_{n-1}}x|; \max_{x > x_{n-1}} |D(1 + \frac{x}{x_{n-1}}) - \frac{D}{x_{n-1}}x| \right\} = D. \end{aligned}$$

We prove the statement by contradiction. Assume that there exists a bounded, agnostic learner for the above problem. The adversary proceeds similarly to the adversary of Theorem 3.2: it finds out the first index  $t$  where the learner would make a prediction with nonzero probability (provided that it receives feedback 0 only). Unless the learner passes infinitely many times, such a  $t$  exists. If the first prediction is nonnegative with at least  $1/2$  chance, the adversary picks  $g_{t,-}$ , otherwise it picks  $g_{t,+}$ . In the first case,

$$\hat{y}_t - g_{t,-}(x_t) \geq 0 + D(1 + \frac{x_n}{x_{n-1}}) > 2D,$$

with nonzero probability, and similarly,  $\hat{y}_t - g_{t,-}(x_t) < -2D$  for the second case, showing that a bounded learner will make a mistake larger than  $2D$  with positive probability. ■

**Theorem 3.5** Fix  $\mathcal{X}, \mathcal{Y}, D, \mathcal{H}, \epsilon$  as in Theorem 3.3 and let  $r \geq 2$ . Then, there exists some problem class  $\mathcal{G}$  so that any algorithm that agnostic  $(D, r, \epsilon)$  KWIK learns  $(\mathcal{H}, \mathcal{G})$  will pass at least  $2^{d-1}$  times.

**Proof**  $\mathcal{X} = [-2, 2]$ ,  $\mathcal{Y} = \mathbb{R}$  with the absolute loss norm, fix some  $1 > D > 0$ . The adversary asks the  $2^{d-1}$  vertices of the hypercube  $\{-1, +1\}^d$  that have positive first coordinates. It is easy to see that the adversary can pick the values on the vertices independently, and if the learner predicts anything with nonzero probability then the adversary can make the protocol fail. The proof is completely analogous to the previous one. ■

## A.2. Proofs for Section 4

The following lemma, which follows from an application of the Hoeffding-Azuma inequality and a careful argumentation with “skipping processes”, will be our basic tool. The novelty of the lemma is that we allow for the possibility of unbounded stopping times, otherwise the lemma would directly follow from Theorem 2.3 in Chapter VII of Doob (1953) and the Hoeffding-Azuma inequality. (It is very well possible that the lemma exists in the literature, however, we could not find it.)

**Lemma A.1** *Let  $\mathcal{F} = (\mathcal{F}_t)_{t \geq 1}$  be a filtration and let  $(\epsilon_t, Z_t)_{t \geq 1}$  be a sequence of  $\{0, 1\} \times \mathbb{R}$ -valued random variables such that  $\epsilon_t$  is  $\mathcal{F}_{t-1}$ -measurable,  $Z_t$  is  $\mathcal{F}_t$ -measurable,  $\mathbb{E}[Z_t | \mathcal{F}_{t-1}] = 0$  and  $Z_t \in [A, A + K]$  for some deterministic quantities  $A, K \in \mathbb{R}$ . Let  $m > 0$  and let  $\tau = \min\{t \geq 1 : \sum_{s=1}^t \epsilon_s = m\}$ , where we take  $\tau = \infty$  when  $\sum_{s=1}^\infty \epsilon_s < m$ . Then, for any  $0 < \delta \leq 1$ , with probability  $1 - \delta$ ,*

$$\sum_{t=1}^{\tau} \epsilon_t Z_t \leq K \sqrt{\frac{m}{2} \log\left(\frac{1}{\delta}\right)}. \quad (5)$$

**Remark A.1 (Analysis of the sum in (5))** *Let  $(\Omega, \mathcal{A})$  be the probability space holding the random variables and the filtration  $\mathcal{F}$ , and let  $S_n = \sum_{t=1}^n \epsilon_t Z_t$  ( $n = 0, 1, \dots$ , the empty sum being zero). The sum  $S$  on the left-hand side of (5) is well-defined almost everywhere on  $\Omega$  as it has at most  $m$  terms no matter whether  $\tau(\omega) < \infty$  or  $\tau(\omega) = \infty$ . It also holds true that the sum  $S$  is an integrable random variable. To see why, consider  $S'_\infty \stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} S_{\tau \wedge n}$ . We claim that the random variable  $S'_\infty$  is well-defined, integrable and  $S'_\infty = S$  holds almost surely. First, notice that  $(S_{\tau \wedge n})_{n \geq 1}$  is a martingale (this follows, e.g., from the Corollary on p.341 to Theorem 9.3.4 of Chung 2001). Next, note that  $(S_{\tau \wedge n})_{n \geq 1}$  is  $L^1$ -bounded (by the condition on  $Z_t$  and because  $S_{\tau \wedge n}$  has at most  $m$  terms, we have  $\mathbb{E}[|S_{\tau \wedge n}|] \leq m(|A| + K)$ ) and therefore it is also uniformly integrable. Hence, Theorem 9.4.6 of Chung 2001 gives that  $S'_\infty$  is a well-defined, integrable random variable. Finally, a simple case analysis shows that  $S'_\infty = S$  holds almost surely.*

**Proof** Let  $S'_n = S_{\tau \wedge n}$ , where  $S_n = \sum_{t=1}^n \epsilon_t Z_t$ ,  $n \geq 0$ . Define  $\mathbb{N}_\infty = \mathbb{N} \cup \{\infty\}$ . By Remark A.1,  $S = S'_\infty$  a.s., where  $S$  is the sum on the right-hand side of (5) and  $S'_\infty = \lim_{n \rightarrow \infty} S_{\tau \wedge n}$ , Theorem 9.4.6 of Chung 2001 mentioned in the remark not only gives that  $S'_\infty$  is integrable, but it also gives that  $(S'_n, \mathcal{F}_n)_{n \in \mathbb{N}_\infty}$  is a martingale, i.e.,  $S'_\infty$  is a “closure” of  $(S'_n, \mathcal{F}_n)_{n \in \mathbb{N}}$ . Let  $\tau_i = \min\{k \geq 1 : \sum_{t=1}^k \epsilon_t = i\}$ ,  $i = 1, \dots, m$  (as before,  $\min \emptyset = \infty$ ). Note that  $\tau_m = \tau$  and  $\tau_i \leq \tau_{i+1}$ ,  $i = 1, \dots, m-1$ . Note that just like  $S_\tau$ ,  $S_{\tau_i}$  is also well-defined by Remark A.1. Consider the process  $(S'_{\tau_i}, \mathcal{F}_{\tau_i})_{i=1, \dots, m}$ , where  $S'_{\tau_i}(\omega) \stackrel{\text{def}}{=} S'_{\tau_i(\omega)}(\omega)$  and  $\mathcal{F}_{\tau_i}$  is the  $\sigma$ -algebra of pre- $\tau_i$  events. By the optional sampling theorem of Doob (see, e.g., Theorem 9.3.5 of Chung 2001),  $(S'_{\tau_i}, \mathcal{F}_{\tau_i})_{i=1, \dots, m}$  is a martingale. Let us now apply the Hoeffding-Azuma inequality to this martingale. In order to be able to do this, we need to show that the increments,  $X_i = S'_{\tau_{i+1}} - S'_{\tau_i}$  lie in some bounded set for  $i = 0, \dots, m-1$ , where we define  $S'_0 = 0$ . When  $\tau_{i+1} = \infty$ ,  $S'_{\tau_{i+1}} = S'_\infty = S$ . Now, if  $\tau_i = \infty$ , we also

have  $S'_{\tau_i} = S'_\infty = S$ , while if  $\tau_i < \infty$ , we have  $S = S_{\tau_i} = S'_{\tau_i}$ . Thus, in both cases,  $X_i = 0 \in [A, A + K]$  (that zero is in this interval follows because  $(Z_t)$  is a martingale increment). When  $\tau_{i+1} < \infty$ , we also have  $\tau_i < \infty$  and thus  $S'_{\tau_{i+1}} = S_{\tau_{i+1}}$  and also  $S'_{\tau_i} = S_{\tau_i}$  and so  $S'_{\tau_{i+1}} - S'_{\tau_i} = \epsilon_{\tau_{i+1}} Z_{\tau_{i+1}}$  and so  $X_i \in [A, A + K]$  by our assumption on  $(Z_t)$ . Thus, we have shown that  $X_i \in [A, A + K]$  almost surely. The application of the Hoeffding-Azuma inequality to  $(S'_{\tau_i}, \mathcal{F}_{\tau_i})_{i=1,\dots,m}$  gives then the desired result. ■

**Remark A.2** Note that the proof does not carry through for the case when we replace the assumption on the range of  $Z_t$  by the assumptions that  $|Z_t| \leq B$  a.s. and  $Z_t \in [A_t, A_t + K]$  for some  $A_t$ ,  $\mathcal{F}_{t-1}$ -measurable random variable and a deterministic constant  $K > 0$ . The problem is twofold:  $(A_{\tau_i})_{i=1,\dots,m}$  might not be well-defined and even if it is, all we can say is that  $\epsilon_{\tau_{i+1}} Z_{\tau_{i+1}} \in [A_{\tau_{i+1}}, A_{\tau_{i+1}} + K]$ , but  $A_{\tau_{i+1}}$  is not necessarily  $\mathcal{F}_{\tau_i}$ -measurable.

In the proof of Theorem 4.1 we will need the one-dimensional version of Helly's theorem, which we nevertheless state for the  $d$ -dimensional Euclidean spaces:

**Theorem A.2 (Helly's Theorem)** Let  $d, N \in \mathbb{N}$ ,  $N > d$ ,  $C_1, \dots, C_N \subseteq \mathbb{R}^d$  be convex subsets of  $\mathbb{R}^d$ . If the intersection of any  $d + 1$  of these sets is nonempty, then  $\cap_{i=1}^N C_i \neq \emptyset$ .

With these preparations, we are ready to prove Theorem 4.1.

**Theorem 4.1** Let  $\mathcal{H}$  be a finite hypothesis class over  $(\mathcal{X}, \mathbb{R})$ ,  $D, \epsilon > 0$ ,  $0 \leq \delta \leq 1$ ,  $r = 2$ . Then, for any  $\mathcal{G}$  problem class such that the noise in the responses lies in  $[-K, K]$  and  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$  it holds that the pairwise elimination based agnostic learner is an agnostic  $(D, r, \epsilon, \delta)$  KWIK-learner for  $(\mathcal{H}, \mathcal{G})$  with KWIK-bound  $((\lceil \frac{2K^2}{\epsilon^2} \log \frac{2(N-1)}{\delta} \rceil - 1)N + 1)(N - 1) = O\left(\frac{K^2 N^2}{\epsilon^2} \log \frac{N}{\delta}\right)$ .

**Proof** Fix  $\mathcal{H}$ ,  $\mathcal{G}$ ,  $D$ ,  $\epsilon$ ,  $\delta$  as in the theorem statement. Let  $g^* : \mathcal{X} \rightarrow \mathbb{R}$  be the function underlying the problem chosen by the adversary. Let  $i^*$  be the index of a function  $f_i \in \mathcal{H}$  that satisfies  $\|f_{i^*} - g^*\|_\infty \leq D$ . By our assumption on  $\mathcal{G}$  and  $\mathcal{H}$ ,  $i^*$  is well-defined.

Let  $E$  be the (error) event when  $i^*$  is eliminated by the algorithm. We will show that  $\mathbb{P}(E) \leq \delta$  as from this, the rest follows easily: Indeed, on the complements of  $E$ , i.e., on  $E^c$ , by the definition of **predict**, the algorithm makes  $2D + \epsilon$ -accurate predictions since if for some input  $x \in \mathcal{X}$ ,  $Y = \bigcap_{i \in I} B_{D+\epsilon}(f_i(x)) \neq \emptyset$  then for any  $\hat{y} \in Y$ ,  $|\hat{y} - f_{i^*}(x)| \leq D + \epsilon$  (since  $i^* \in I$ ) and thus

$$|\hat{y} - g(x)| \leq |\hat{y} - f_{i^*}(x)| + |f_{i^*}(x) - g(x)| \leq 2D + \epsilon.$$

Also, every time the algorithm passes, at least one of the counters  $n_{i,j}$  is incremented. Indeed, if upon receiving input  $x$  the algorithm did not pass then  $\bigcap_{i \in I} B_{D+\epsilon}(f_i(x)) = \emptyset$ . Therefore,  $|I| > 1$  and it follows from Helly's theorem (Theorem A.2) that there exists two distinct indices  $i, j \in I$  such that  $B_{D+\epsilon}(f_i(x)) \cap B_{D+\epsilon}(f_j(x)) = \emptyset$ . For this pair  $(i, j)$ , either

$n_{i,j}$  or  $n_{j,i}$  is incremented. When some counter  $n_{i,j}$  reaches the value  $m$ , at least one of  $i$  and  $j$  is excluded. Therefore, there can be at most  $(m-1)N(N-1)$  calls to **learn** with no exclusions. Further, there can only be  $N-1$  exclusions (since on  $E^c$  the index  $i^*$  does not get excluded). Thus, on  $E^c$ , there can be no more than  $(m-1)N(N-1) + (N-1)$  calls to **learn**. Plugging in the value of  $m$  gives the KWIK-bound.

Thus, it remains to show that the probability of  $E$  is small, i.e., that  $\mathbb{P}(E) \leq \delta$ . To prove this we need some more notation. Let  $G = (\mathcal{X}, \mathcal{Y}, g^*, Z)$  be the problem and let  $(x_t)_{t \geq 1}$  be the sequence of covariates ( $x_t \in \mathcal{X}$ ) chosen by the adversary. To simplify the presentation, we introduce for each  $t \geq 1$  a response,  $y_t = g(x_t) + z_t$ . Let  $\mathcal{F}_t = \sigma(x_1, y_1, \dots, x_t, y_t)$ . By assumption, the noise satisfies  $z_t \sim Z(x_t)$  and, in particular,  $z_t$  lies in  $[-K, K]$ , and  $\mathbb{E}[z_t | \mathcal{F}_{t-1}, x_t] = 0$ . Let  $\pi_t$  be the indicator of whether the learner has passed when presented with the input  $x_t$ :  $\pi_t = 1$  if the learner passed (and thus  $y_t$  is available for learning) and  $\pi_t = 0$ , otherwise. For  $i, j \in \{1, \dots, N\}$ ,  $t \geq 1$  let  $\epsilon_t^{(i,j)} = \mathbb{I}_{\{f_i(x_t) \ll f_j(x_t)\}}$  and let  $\tau^{(i,j)} = \min\{n \geq 1 : \sum_{t=1}^n \pi_t \epsilon_t^{(i,j)} = m\}$  be the time when the counter  $n_{i,j}$  reaches  $m$  and thus either  $i$  or  $j$  gets eliminated by the algorithm (if it was not eliminated before). Here we let  $\tau^{(i,j)} = \infty$  when  $\sum_{t=1}^\infty \pi_t \epsilon_t^{(i,j)} < m$ . Note that  $i^*$  gets eliminated only if one of  $\tau^{(i^*,j)}$  or  $\tau^{(j,i^*)}$  is finite for some  $j \neq i^*$ ,  $1 \leq j \leq N$ . Thus,

$$E = \bigcup_{j \neq i^*}^*(E \cap \{\tau^{(i^*,j)} < \infty\}) \cup^* \bigcup_{j \neq i^*}^*(E \cap \{\tau^{(j,i^*)} < \infty\}). \quad (6)$$

We show that for  $j \neq i^*$ , both  $E \cap \{\tau^{(i^*,j)} < \infty\}$  and  $E \cap \{\tau^{(j,i^*)} < \infty\}$  happen with small probability.

Fix  $j \neq i^*$  and consider  $E \cap \{\tau^{(i^*,j)} < \infty\}$ . To simplify the notation introduce  $\tau = \tau^{(i^*,j)}$  and  $\epsilon_t = \epsilon_t^{(i^*,j)}$ . Let  $F$  be the event  $F = \{\sum_{t=1}^\tau \pi_t \epsilon_t (f_{i^*}(x_t) + f_j(x_t)) < 2 \sum_{t=1}^\tau \pi_t \epsilon_t y_t\}$ . Then  $E \cap \{\tau^{(i^*,j)} < \infty\} = E \cap \{\tau < \infty\} \subset F \cap \{\tau < \infty\}$  holds because by the definition of the **learn** procedure, if  $i^*$  gets eliminated at time  $\tau$  due to  $n_{i^*,j}$  reaching  $m$  then it must hold that

$$\sum_{t=1}^\tau \pi_t \epsilon_t (f_{i^*}(x_t) + f_j(x_t)) < 2 \sum_{t=1}^\tau \pi_t \epsilon_t y_t. \quad (7)$$

Define  $G = \{\sum_{t=1}^\tau \pi_t \epsilon_t z_t > K \sqrt{2m \log(2(N-1)/\delta)}\}$ . We claim that

$$F \cap \{\tau < \infty\} \subset G \cap \{\tau < \infty\}. \quad (8)$$

To prove this, assume that (7) holds and  $\tau < \infty$ . Then,

$$\begin{aligned}
\sum_{t=1}^{\tau} \pi_t \epsilon_t z_t &= \sum_{t=1}^{\tau} \pi_t \epsilon_t (y_t - g^*(x_t)) \\
&> \sum_{t=1}^{\tau} \pi_t \epsilon_t \left\{ \frac{f_{i^*}(x_t) + f_j(x_t)}{2} - g^*(x_t) \right\} && (\text{because of (7)}) \\
&\geq \sum_{t=1}^{\tau} \pi_t \epsilon_t \{ f_{i^*}(x_t) + (D + \epsilon) - g^*(x_t) \} && (\text{definition of } \epsilon_t = \epsilon_t^{(i^*, j)}) \\
&\geq \left\{ \sum_{t=1}^{\tau} \pi_t \epsilon_t \right\} \epsilon && (\text{because } f_{i^*}(x_t) \geq g^*(x_t) - D) \\
&= m\epsilon && (\text{definition of } \tau \text{ and } \tau < \infty) \\
&\geq K \sqrt{2m \log \left( \frac{2(N-1)}{\delta} \right)} && (\text{definition of } m),
\end{aligned}$$

finishing the proof of (8). By Lemma A.1,  $\mathbb{P}(G) \leq \delta/(2(N-1))$  and thus we also have

$$\mathbb{P}\left(E \cap \{\tau^{(i^*, j)} < \infty\}\right) \leq \frac{\delta}{2(N-1)}.$$

With an entirely similar argument we can show that  $\mathbb{P}(E \cap \{\tau^{(j, i^*)} < \infty\}) \leq \delta/(2(N-1))$  holds, too. Therefore, by the decomposition (6),  $\mathbb{P}(E) \leq \delta$ , finishing the proof. ■

## Appendix B. Proofs for Section 5

Before turning to the proof of Lemma 5.3, we state Freedman's version of Bernstein's inequality (see, Freedman 1975, Theorem 1.6).

**Theorem B.1** *Let  $\mathcal{F} = (\mathcal{F}_k)_{k \geq 0}$  be a filtration and consider a sequence of  $\mathcal{F}$ -adapted random variables  $(X_k)_{k \geq 1}$ . Assume that  $\mathbb{E}[X_k | \mathcal{F}_{k-1}] \leq 0$  and  $X_k \leq R$  a.s. for  $k = 1, 2, 3, \dots$ . Let the  $k$ th partial sum of  $(X_k)_{k \geq 1}$  be  $S_k$  and let  $V_k^2$  be the total conditional variance up to time  $k$ :  $S_k = \sum_{i=1}^k X_i$ ,  $V_k^2 = \sum_{i=1}^k \text{Var}[X_i^2 | \mathcal{F}_{i-1}]$ . Let  $\tau$  be a (not necessarily finite) stopping time w.r.t.  $\mathcal{F}$ . Then, for all  $t \geq 0$ ,  $v \in \mathbb{R}$ ,*

$$\mathbb{P}(S_\tau \geq t, V_\tau^2 \leq v^2 \text{ and } \tau < \infty) \leq \exp \left\{ -\frac{t^2/2}{v^2 + Rt/3} \right\}.$$

In the literature the above theorem is sometimes stated for *finite* stopping times only (or for the specific case when  $\tau = k$  for some  $k$ ). In fact, inequality 1.5(a) in Freedman's paper, from which the above theorem follows, is presented for finite stopping times only. However, the form of Theorem 1.6 of Freedman (1975) is actually equivalent to Theorem B.1. The next result follows from Theorem B.1 by a simple “inversion” argument and is given here because it will suit our needs better:

**Corollary B.2** Let  $\mathcal{F}$ ,  $(X_k)_{k \geq 1}$ ,  $(S_k)_{k \geq 1}$ ,  $(V_k^2)_{k \geq 1}$ ,  $R$  and  $\tau$  be as in Theorem B.1. Then, for any  $v \in \mathbb{R}$ ,  $0 < \delta \leq 1$ , it holds that

$$\mathbb{P} \left( S_\tau \geq \sqrt{2v^2 \ln \left( \frac{1}{\delta} \right)} + \frac{2R}{3} \ln \left( \frac{1}{\delta} \right), V_\tau^2 \leq v^2 \text{ and } \tau < \infty \right) \leq \delta.$$

Let us now turn to the proof of Lemma 5.3:

**Lemma 5.3** Let  $0 < \epsilon < 1$ ,  $m \in \mathbb{N}$  be deterministic constants,  $(\mathcal{F}_t)_{t \geq 1}$  be some filtration and let  $(A_t)_{t \geq 1}$  be an  $(\mathcal{F}_{t+1})_{t \geq 1}$ -adapted sequence of indicator variables. Let

$$a_t = \mathbb{E}[A_t | \mathcal{F}_t]$$

and let  $G$  be an event such that on  $G$  the inequality  $\sum_{t=1}^{\infty} A_t \leq m$  holds almost surely. Then, for any  $0 < \delta \leq 1$  with probability  $1 - \delta$ , either  $G^c$  holds, or

$$\sum_{t=1}^{\infty} \mathbb{I}_{\{a_t \geq \epsilon\}} \leq \frac{1}{\epsilon} \left\{ m + \sqrt{2m \log \left( \frac{1}{\delta} \right)} + 3\sqrt{\log \left( \frac{1}{\delta} \right)} + 6\log \left( \frac{1}{\delta} \right) \right\}.$$

It is interesting to compare the result of this lemma to what happens when  $(a_t)_{t \geq 1}$  is a deterministic sequence, and  $A_t$  is Bernoulli with parameter  $a_t$ , independently chosen of all the other random variables. Clearly, in this case if  $\sum_t A_t \leq m$  holds almost surely, we will also have  $\sum_{t=1}^{\infty} a_t \leq m$ . In contrast to this, in the sequential setting of the above lemma there exists  $(A_t, a_t)$  satisfying the conditions of the lemma such that for any  $B > 0$ , with positive probability,  $\sum_{t=1}^{\infty} a_t > B$  (note that in both cases,  $\mathbb{E}[\sum_{t=1}^{\infty} a_t] = \mathbb{E}[\sum_{t=1}^{\infty} A_t]$ ). Since  $\sum_t a_t \geq \sum_t \mathbb{I}_{\{a_t \geq \epsilon\}} a_t \geq \epsilon \sum_t \mathbb{I}_{\{a_t \geq \epsilon\}}$ , in the setting of independent Bernoulli trials, we get that almost surely,  $\sum_t \mathbb{I}_{\{a_t \geq \epsilon\}} \leq \frac{1}{\epsilon} \sum_t A_t \leq m/\epsilon$ . Thus, we see that the dependent and independent cases are quite different and the above lemma can be seen as quantifying the price of choosing  $a_t$  in a sequential manner.

**Proof** Define  $S_n = \sum_{t=1}^n (a_t - A_t)$ ,  $s_n = \sum_{t=1}^n a_t$ ,  $V_n^2 = \sum_{t=1}^n \text{Var}[a_t - A_t | \mathcal{F}_t] = \sum_{t=1}^n a_t(1 - a_t)$ ,  $\hat{s}_n = \sum_{t=1}^n \mathbb{I}_{\{a_t \geq \epsilon\}}$ , for  $n = 1, 2, \dots, \infty$ . Note that  $V_n^2 \leq s_n$  holds for any  $n$  and  $V_n^2$  is the total conditional variance associated with the martingale  $(S_n, \mathcal{F}_{n+1})_{n \geq 0}$  (the empty sum is defined to be zero).

Fix  $0 < \delta \leq 1$ . Let  $f = f(m, \delta)$  be a real number to be chosen later. We will define this number such that on some event  $F_\delta$ , whose probability is at least  $1 - \delta$ , we will have that

$$s_\infty \geq f \text{ implies that } \sum_{t=1}^{\infty} A_t > m. \quad (9)$$

Once we prove this, it follows by our assumption on  $\sum_{t=1}^{\infty} A_t$  that on  $F_\delta \cap G$ , we must have  $s_\infty < f$ . Now, using  $\mathbb{I}_{\{a_t \geq \epsilon\}} \epsilon \leq \mathbb{I}_{\{a_t \geq \epsilon\}} a_t$ , we get that  $\epsilon \hat{s}_n \leq \sum_{t=1}^n \mathbb{I}_{\{a_t \geq \epsilon\}} a_t \leq s_n$ . Therefore, on  $F_\delta \cap G$ ,  $\hat{s}_\infty \leq \epsilon^{-1} s_\infty \leq \epsilon^{-1} f$ . Plugging in the value of  $f$  will then finish the proof.

The event  $F_\delta$  is chosen as follows: Let  $\tau$  be the first index  $n$  when  $s_n \geq f$  holds and let  $\tau = \infty$  when there is no such index. Using Corollary B.2, we get that the probability of the event

$$E = \left\{ S_\tau \geq \sqrt{2(f+1) \log(\frac{1}{\delta})} + \frac{2}{3} \log(\frac{1}{\delta}), V_\tau^2 \leq f+1, \tau < \infty \right\}$$

is at most  $\delta$ :  $\mathbb{P}(E) \leq \delta$ . Note that  $V_\tau^2 \leq s_\tau \leq f+1$  holds almost surely, where the last inequality follows from the definition of  $\tau$  and because  $a_t \in [0, 1]$ . Therefore, the second condition can be dropped in the definition of  $E$  without changing it:  $E = \left\{ S_\tau \geq \sqrt{2(f+1) \log(\frac{1}{\delta})} + \frac{2}{3} \log(\frac{1}{\delta}), \tau < \infty \right\}$ . Take  $F_\delta = E^c$ . Thus,  $\mathbb{P}(F_\delta) \geq 1 - \delta$  and on  $F_\delta$ , we have

$$S_\tau < \sqrt{2(f+1) \log(\frac{1}{\delta})} + \frac{2}{3} \log(\frac{1}{\delta}) \text{ or } \tau = \infty$$

which is equivalent to

$$\sum_{t=1}^{\tau} A_t > s_\tau - \sqrt{2(f+1) \log(\frac{1}{\delta})} - \frac{2}{3} \log(\frac{1}{\delta}) \text{ or } \tau = \infty. \quad (10)$$

Let us now show that on  $F_\delta$ , (9) holds. Consider an outcome  $\omega \in F_\delta$  and assume that we also have  $s_\infty(\omega) \geq f$  (to avoid clutter, we suppress  $\omega$  in what follows). Because of  $s_\infty \geq f$ , it follows that  $\tau < \infty$  and  $s_\tau \geq f$ . Therefore, from (10) and from  $\sum_{t=1}^{\tau} A_t \leq \sum_{t=1}^{\infty} A_t$ , we get that

$$\sum_{t=1}^{\infty} A_t > (f+1) - \sqrt{2(f+1) \log(\frac{1}{\delta})} - \frac{2}{3} \log(\frac{1}{\delta}) - 1. \quad (11)$$

Now, define  $f = f(m, \delta)$  to be a number such that

$$(f+1) - \sqrt{2(f+1) \log(\frac{1}{\delta})} - \frac{2}{3} \log(\frac{1}{\delta}) - 1 \geq m. \quad (12)$$

Such a number exists because the left hand side, as a function of  $f$ , is unbounded. In fact, a simple calculation shows that, with the definitions  $c = \sqrt{2 \log(1/\delta)}$  and  $L = 2/3 \log(1/\delta) + 1$ , choosing  $f$  so that  $(f+1)^{1/2}$  is larger than  $(c + \sqrt{c + 4(m+L)})/2$  makes (12) hold true. Some calculation shows that  $f \leq m + \sqrt{2m \log(1/\delta)} + 3\sqrt{\log(1/\delta)} + 6 \log(1/\delta)$ . Chaining the inequality (11) with the inequality (12), we get that on  $F_\delta$ , (9) indeed holds, thus, finishing the proof. ■

**Claim 5.4** *We have  $V_M^{\tilde{\pi}_t}(s_t; L) \geq V_{M_t}^{\pi_t}(s_t; L) - 2V_{\max} p_t$ , with probability one.*

**Proof** Let  $\mathcal{Z} = (S \times A)^L$  be the space of trajectories of length  $L$  which is viewed as a measurable space with the product  $\sigma$ -algebra (for  $S, A$  finite, this is just the discrete  $\sigma$ -algebra). Let  $\pi_t^\circ$  be an arbitrary  $\mathcal{F}_t$ -measurable policy,  $M_t = (S, A, P_{M_t}, R_{M_t})$  be an MDP,

where  $P_{M_t}$  and  $R_{M_t}$  are  $\mathcal{F}_t$ -measurable. Let  $F_{t,M_t,\pi_t^\circ}$  be the measure induced by  $M_t$  and  $\pi_t^\circ$  on the space of  $L$ -step trajectories  $\mathcal{Z}$ , and the initial state distribution that is concentrated at the single state  $s_t$  (i.e., the initial state distribution used in the definition of  $F_{t,M_t,\pi_t^\circ}$  is the Dirac-measure  $\delta_{s_t}(\cdot)$ ). Note that  $F_{t,M_t,\pi_t^\circ}$  is a random measure, which is itself  $\mathcal{F}_t$ -measurable. Let  $\mathcal{R}_{M_t} : \mathcal{Z} \rightarrow \mathbb{R}$  be the mapping that assigns  $M_t$ -returns to the trajectories in  $\mathcal{Z}$ :

$$\mathcal{R}_{M_t}(s_0, a_0, \dots, s_{L-1}, a_{L-1}) = \sum_{i=0}^{L-1} \gamma^i R_{M_t}(s_i, a_i). \quad (13)$$

Now, consider the measures  $F_{t,M,\tilde{\pi}_t}$  and  $F_{t,\bar{M}_t,\tilde{\pi}_t}$  (these are  $\mathcal{F}_t$ -measurable thanks to condition (b)). An important property of these measures is that they agree when restricted to  $Z_{K_t}$ :

$$F_{t,M,\tilde{\pi}_t}|_{Z_{K_t}} = F_{t,\bar{M}_t,\tilde{\pi}_t}|_{Z_{K_t}}. \quad (14)$$

This property will play a crucial role in proving the desired inequality. Two further identities that we will need are the following: Let  $Z_{K_t} = K_t^L$  be the set of  $L$ -step trajectories that stay within  $K_t$ . Then, we have

$$p_t = \int \mathbb{I}_{\{z \notin Z_{K_t}\}} dF_{t,M,\tilde{\pi}_t}(z) \quad (15)$$

$$= \int \mathbb{I}_{\{z \notin Z_{K_t}\}} dF_{t,M,\pi_t}(z). \quad (16)$$

Clearly, these two equations are equivalent to the following ones:

$$1 - p_t = \int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,M,\tilde{\pi}_t}(z) \quad (17)$$

$$= \int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,M,\pi_t}(z). \quad (18)$$

Therefore, it will suffice to prove that these latter equations hold true.

To show (17), first notice that  $1 - p_t = \mathbb{E} [\mathbb{I}_{\{E_t^c\}} | \mathcal{F}_t]$  and  $\mathbb{I}_{\{E_t^c\}} = \prod_{i=0}^{L-1} \mathbb{I}_{\{(s_{t+i}, a_{t+i}) \in K_t\}}$ . Therefore,  $\mathbb{E} [\mathbb{I}_{\{E_t^c\}} | \mathcal{F}_t] = \int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,M,\tilde{\pi}_t}(z)$ , which shows that (17) indeed holds.

Let us now turn to the proof of (18). By (14),  $\int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,M,\tilde{\pi}_t}(z) = \int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,\bar{M}_t,\tilde{\pi}_t}(z)$ . Thanks to condition (g), along those trajectories that stay in  $K_t$ , the policy followed does not change. This implies that

$$dF_{t,\bar{M}_t,\tilde{\pi}_t}|_{Z_{K_t}} = dF_{t,\bar{M}_t,\pi_t}|_{Z_{K_t}}. \quad (19)$$

Therefore, we also have  $\int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,\bar{M}_t,\tilde{\pi}_t}(z) = \int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,\bar{M}_t,\pi_t}(z)$ , finishing the proof of (18).

Let us continue with the lower bound on  $V^{\tilde{\pi}_t}(s_t; L)$ . Then,  $V_M^{\tilde{\pi}_t}(s_t; L) = \int \mathcal{R}_M(z) dF_{t,M,\tilde{\pi}_t}(z)$ , where  $\mathcal{R}_M(z)$  is the return assigned by model  $M$  to trajectory  $z \in \mathcal{Z}$  (cf. (13)). Now, break

the integral into two parts using the decomposition  $\mathcal{Z} = Z_{K_t} \cup^* Z_{K_t}^c$ . For the integral over  $Z_{K_t}^c$  use that  $|\mathcal{R}_M(z)| \leq V_{\max}$  (which holds by condition (a)) and then (15) to get  $V_M^{\tilde{\pi}_t}(s_t; L) \geq \int_{Z_{K_t}} \mathcal{R}_M(z) dF_{t,M,\tilde{\pi}_t}(z) - V_{\max} p_t$ . By (14) and (19),

$$\int_{Z_{K_t}} \mathcal{R}_M(z) dF_{t,M,\tilde{\pi}_t}(z) = \int_{Z_{K_t}} \mathcal{R}_{\bar{M}_t}(z) dF_{t,\bar{M}_t,\pi_t}(z) = V_{\bar{M}_t}^{\pi_t}(s_t; L) - \int_{Z_{K_t}^c} \mathcal{R}_{\bar{M}_t}(z) dF_{t,\bar{M}_t,\pi_t}(z).$$

Using again  $\mathcal{R}_{\bar{M}_t}(z) \leq V_{\max}$  (which follows from condition (a)) and then (16) and chaining the previous equalities and inequalities, we get  $V_M^{\tilde{\pi}_t}(s_t; L) \geq V_{\bar{M}_t}^{\pi_t}(s_t; L) - 2p_t V_{\max}$ , which is the inequality that was to be proven.  $\blacksquare$

We need some preparations before we give the proof of Claim 5.5. The next lemma also follows from a simple contraction argument (the proof is omitted). The lemma uses the partial ordering of functions:  $f_1 \leq f_2$  if  $f_1(x) \leq f_2(x)$  holds for all  $x \in \text{Dom}(f_1) = \text{Dom}(f_2)$ . Also, an operator is *isotone* if it preserves the ordering of its arguments.

**Lemma B.3 (Comparison Lemma)** *Let  $B$  be a Banach-space of real-valued functions over some domain  $D$ . Let  $T_1, T_2 : B \rightarrow B$  be contractions and let  $f_1^*, f_2^* \in B$  be their respective (unique) fixed-points. Assume that  $T_1$  is isotone. Then, if  $T_1 f_2^* \leq T_2 f_2^* = f_2^*$  then  $f_1^* \leq f_2^*$ .*

In the proof below, we also need the concept of Bellman operators. Let  $M = (S, A, P, R)$  be an MDP and let  $\pi$  be a stationary policy over  $(S, A)$ . The Bellman operator  $T_M^\pi : \mathbb{R}^{S \times A} \rightarrow \mathbb{R}^{S \times A}$  underlying  $M$  and  $\pi$  is defined by

$$(T_M^\pi Q)(s, a) = R(s, a) + \gamma \int Q(s', a) d\pi(a|s') dP(s'|s, a), \quad (s, a) \in S \times A.$$

As it is well known,  $T_M^\pi$  is a contraction with respect to the maximum norm and if  $Q_M^\pi$  denotes the unique fixed point of  $T_M^\pi$ ,  $\int Q_M^\pi(s, a) d\pi(a|s) = V_M^\pi(s)$  holds for all  $s \in S$  (in fact,  $Q_M^\pi(s, a)$  is the so-called action-value function underlying  $\pi$ , i.e.,  $Q_M^\pi(s, a)$  is the expected total discounted return if the decision process is started at state  $s$ , the first action is  $a$  and the subsequent actions are chosen by  $\pi$ ).

**Claim 5.5** *On  $G$ , it holds that  $V_{\bar{M}_t}^{\pi^*}(s_t) \geq V_{\hat{M}_t}^{\pi^*}(s_t)$ .*

**Proof** Instead of the claimed inequality, we prove the stronger inequality  $Q_{\bar{M}_t}^{\pi^*} \geq Q_{\hat{M}_t}^{\pi^*}$ . To prove this, we apply the Comparison Lemma (Lemma B.3). Choose  $B$  as the Banach-space of bounded, real-valued functions over  $S \times A$  with the supremum norm  $\|\cdot\|_\infty$ , and consider two operators  $T_{\bar{M}_t}^{\pi^*}, T_{\hat{M}_t}^{\pi^*} : B \rightarrow B$ . Operator  $T_{\bar{M}_t}^{\pi^*}$  is the policy evaluation operator corresponding to  $\pi^*$  on  $\bar{M}_t$ :  $T_{\bar{M}_t}^{\pi^*} Q(s, a) = R_{\bar{M}_t}(s, a) + \gamma \int Q(s', a) d\pi^*(a|s') dP_{\bar{M}_t}(s'|s, a)$ , while  $T_{\hat{M}_t}^{\pi^*}$  is the  $V_{\max}$ -truncated policy evaluation operator corresponding to  $\pi^*$  on  $\hat{M}_t$ :

$T_{\hat{M}_t}^{\pi^*} Q(s, a) = \Pi \left[ R_{\tilde{M}_t}(s, a) + \gamma \int Q(s', a) d\pi(a|s') dP_{\tilde{M}_t}(s'|s, a) \right]$ , where  $\Pi : \mathbb{R} \rightarrow \mathbb{R}$  is the projection to  $[-V_{\max}, V_{\max}]$ , i.e.,  $\Pi(x) = \max(\min(x, V_{\max}), -V_{\max})$ . Clearly,  $Q_{\hat{M}_t}^{\pi^*}$  is the fixed point of  $Q_{\hat{M}_t}^{\pi^*}$ , while  $Q_{\tilde{M}_t}^{\pi^*}$  is the fixed point of  $Q_{\tilde{M}_t}^{\pi^*}$ , the latter of which follows because  $\|Q_{\tilde{M}_t}^{\pi^*}\|_{\infty} \leq V_{\max}$ , thanks to condition (a). It is also clear that both operators are contractions. Take any  $(s, a) \in S \times A$ . We claim that  $T_{\hat{M}_t}^{\pi^*} Q_{\hat{M}_t}^{\pi^*}(s, a) \geq T_{\tilde{M}_t}^{\pi^*} Q_{\hat{M}_t}^{\pi^*}(s, a)$ . Let us first show that this inequality holds when  $(s, a) \in K_t$ . In this case,  $|T_{\hat{M}_t}^{\pi^*} Q_{\hat{M}_t}^{\pi^*}(s, a)| \leq |r_{\tilde{M}_t}(s, a)| + \gamma V_{\max} \leq V_{\max}$ , because, by construction  $r_{\tilde{M}_t}(s, a) = r_{\hat{M}_t}(s, a)$  and by condition (a)  $|r_{\hat{M}_t}(s, a)| \leq (1 - \gamma)V_{\max}$ . Therefore, the projection has no effect. Using that on the set  $K_t$  the models  $\hat{M}_t$  and  $\tilde{M}_t$  coincide, we get that  $T_{\hat{M}_t}^{\pi^*} Q_{\hat{M}_t}^{\pi^*}(s, a) = T_{\tilde{M}_t}^{\pi^*} Q_{\hat{M}_t}^{\pi^*}(s, a)$ . Now, consider the case when  $(s, a) \notin K_t$ . In this case,  $T_{\hat{M}_t}^{\pi^*} Q_{\hat{M}_t}^{\pi^*}(s, a) = Q_{\hat{M}_t}^{\pi^*}(s, a) \geq V_{\max} \geq T_{\tilde{M}_t}^{\pi^*} Q_{\hat{M}_t}^{\pi^*}(s, a)$ , where the first inequality follows from condition (f), while the second follows because  $T_{\tilde{M}_t}^{\pi^*} Q(s, a)$  is restricted to the interval  $[-V_{\max}, V_{\max}]$ . This finishes the verification of the conditions of the Comparison Lemma. Therefore, the lemma gives that  $Q_{\hat{M}_t}^{\pi^*} \geq Q_{\tilde{M}_t}^{\pi^*}$ , which is the inequality that we wished to prove. ■

**Theorem 5.6** Fix a state space  $S$  and an action space  $A$ , which are assumed to be non-empty Borel spaces. Let  $\mathcal{X}, \mathcal{Y}$  be as described above,  $\mathcal{H}$  be an MDP hypothesis set,  $\mathcal{G}$  be a set of MDP problem instances, both over  $\mathcal{X}, \mathcal{Y}$ . Assume that  $\Delta(\mathcal{G}, \mathcal{H}) \leq D$ . Assume that  $V_{\max} > 0$  is such that  $(1 - \gamma)V_{\max}$  is an upper bound on the immediate rewards of the MDPs determined by members of  $\mathcal{H}$  and  $\mathcal{G}$ . Fix  $\epsilon > 0, r \geq 1, 0 < \delta \leq 1/2$ . Assume that MDPLearner is an agnostic  $(D, r, \epsilon)$  KWIK-learner for  $(\mathcal{H}, \mathcal{G})$  with KWIK-bound  $B(\delta)$ . Assume further that we are given a Planner which is  $e_{\text{planner}}$ -accurate. Consider the instance of the KWIK-RMAX algorithm which uses MDPLearner and Planner, interacting with some MDP  $M$  from  $\mathcal{G}$ . Let  $\epsilon' = \frac{5(rD+\epsilon)}{1-\gamma} + e_{\text{planner}}$ . Then, with probability  $1 - 2\delta$ , the number of  $\epsilon'$ -mistakes,  $N_{M, \epsilon'}$ , made by KWIK-RMAX is bounded by  $\frac{2V_{\max}(1-\gamma)L}{rD+\epsilon} \left\{ B(\delta) + (\sqrt{2B(\delta)} + 3) \sqrt{\log\left(\frac{L}{\delta}\right)} + 6 \log\left(\frac{L}{\delta}\right) \right\}$ , where  $L = \max(1, \lceil (1-\gamma)^{-1} \log(V_{\max}(1-\gamma)/(rD+\epsilon)) \rceil)$ .

**Proof** We apply Theorem 5.1 to the agent KWIK-RMAX that uses MDPLearner and Planner. Fix  $0 < \delta \leq 1$ . The event  $G$  is constructed as follows: MDPLearner interacts with an “environment” according to the KWIK protocol. Consider the event on which it holds that the number of timesteps when MDPLearner passes is bounded by  $B(\delta)$ , while the learner’s prediction errors (on the same event) is always below  $rD + \epsilon$ . By assumption, this event has probability at least  $1 - \delta$ . Call this event  $G$ . Now, the sequence  $(K_t)_{t \geq 1}$  is simply determined as follows. Let  $g_t : S \times A \rightarrow \mathcal{Y} \cup \{\perp\}$  be the function underlying the predictions made by MDPLearner in step  $t$ . Then,  $K_t = \{(s, a) \in S \times A : g_t(s, a) \neq \perp\}$ . Further, let  $M_t$  be the model returned by the optimistic wrapper and let the policy  $\pi_t$  be the policy that Planner would “compute” at time  $t$  (i.e.,  $\pi_t(\cdot|s)$  is the distribution of actions returned by Planner if state  $s$  is fed to it). Let us verify the conditions of Theorem 5.1. The

bound on the expected immediate rewards (condition (a)) holds by assumption, just like the measurability condition (b) and that the action selected at time  $t$  is sampled from  $\pi_t(\cdot|s_t)$  (condition (c)). The condition on the accuracy of the planner (condition (d)) was assumed as a condition of this theorem. The accuracy condition (e) holds with  $e_{\text{model}} = rD + \epsilon$  on  $G$  by the choice of  $G$ , while the optimism condition (f) is met because of the use of the optimistic wrapper (in fact, because of this wrapper,  $Q_{\hat{M}_t}^\pi(s, a) = V_{\max}$  holds for any  $(s, a) \notin K_t$ ). Also, condition (g) is met because the learn method of MDP Learner is not called when  $(s_t, a_t) \in K_t$ , hence in that case  $g_{t+1} = g_t$  and thus  $\pi_{t+1} = \pi_t$ . Finally, on  $G$ ,  $B = B(\delta)$  bounds the number of times  $(s_t, a_t) \notin K_t$  happens. Therefore, by the conclusion of Theorem 5.1, with probability at least  $1 - 2\delta$ , the number of  $5e_{\text{model}}/(1 - \gamma) + e_{\text{plan}}$  mistakes is bounded by  $\frac{2V_{\max}(1-\gamma)L}{e_{\text{model}}} \left\{ B + (\sqrt{2B} + 3)\sqrt{\log\left(\frac{L}{\delta}\right)} + 6\log\left(\frac{L}{\delta}\right) \right\}$ , where  $L = \max(1, \lceil(1 - \gamma)^{-1} \log(V_{\max}(1 - \gamma)/e_{\text{model}})\rceil)$ . Plugging in the value  $e_{\text{model}} = rD + \epsilon$  gives the final bound. ■

## References

- Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *NIPS*, pages 89–96, 2008.
- Ronen I. Brafman and Moshe Tennenholtz. A near-optimal polynomial time algorithm for learning in certain classes of stochastic games. *Artificial Intelligence*, 121(1-2):31–47, 2000.
- Ronen I. Brafman and Moshe Tennenholtz. R-MAX - a general polynomial time algorithm for near-optimal reinforcement learning. In *IJCAI*, pages 953–958, 2001.
- Kai Lai Chung. *A course in probability theory*. Academic Press, 3 edition, 2001.
- Carlos Diuk, Andre Cohen, and Michael L. Littman. An object-oriented representation for efficient reinforcement learning. In *ICML*, pages 240–247, 2008.
- Joseph L. Doob. *Stochastic processes*. John Wiley & Sons, 1953.
- Claude-Nicolas Fiechter. Efficient reinforcement learning. In *COLT*, pages 88–97, 1994.
- David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, 1975.
- Sham M. Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University College London, 2003.
- Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

Lihong Li. *A Unifying Framework for Computational Reinforcement Learning Theory*. PhD thesis, Department of Computer Science, Rutgers University, New Brunswick, NJ, USA, 2009.

Lihong Li and Michael L. Littman. Reducing reinforcement learning to kwik online regression. *Annals of Mathematics and Artificial Intelligence*, 58(3-4):217–237, 2010.

Lihong Li, Michael L. Littman, and Thomas J. Walsh. Knows what it knows: A framework for self-aware learning. In *ICML*, pages 568–575, 2008.

Lihong Li, Michael L. Littman, Thomas J. Walsh, and Alexander L. Strehl. Knows what it knows: A framework for self-aware learning. *Machine Learning*, 82(3):399–443, 2011a.

Lihong Li, Michael L. Littman, Thomas J. Walsh, and Alexander L. Strehl. Knows what it knows: a framework for self-aware learning. *Machine learning*, 82:399–443, 2011b.

Martin .L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 2005.

Alexander L. Strehl. Model-based reinforcement learning in factored-state MDPs. In *IEEE ADPRL*, pages 103–110, 2007.

Alexander L. Strehl and Michael L. Littman. A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd international conference on Machine learning*, pages 856–863, 2005.

Alexander L. Strehl and Michael L. Littman. Online linear regression and its application to model-based reinforcement learning. In *NIPS*, 2007.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Incremental model-based learners with formal learning-time guarantees. In *UAI*, pages 485–493, 2006.

Alexander L. Strehl, Carlos Diuk, and Michael L. Littman. Efficient structure learning in factored-state MDPs. In *AAAI*, pages 645–650, 2007.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *The Journal of Machine Learning Research*, 10:2413–2444, 2009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 1998.

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

István Szita and András Lőrincz. The many faces of optimism: a unifying approach. In *ICML*, pages 1048–1055, 2008.

István Szita and Csaba Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML*, pages 1031–1038, June 2010.

SZITA SZEPESVÁRI

Thomas J. Walsh, Sergiu Goschin, and Michael Littman. Integrating sample-based planning and model-based reinforcement learning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, 2010.

# The Sample Complexity of Dictionary Learning

**Daniel Vainsencher**

**Shie Mannor**

*Department of Electrical Engineering  
Technion, Israel Institute of Technology  
Haifa 32000, Israel*

DANIELV@TX.TECHNION.AC.IL

SHIE@EE.TECHNION.AC.IL

**Alfred M. Bruckstein**

*Department of Computer Science  
Technion, Israel Institute of Technology  
Haifa 32000, Israel*

FREDDY@CS.TECHNION.AC.IL

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

A large set of signals can sometimes be described sparsely using a dictionary, that is, every element can be represented as a linear combination of few elements from the dictionary. Algorithms for various signal processing applications, including classification, denoising and signal separation, learn a dictionary from a given set of signals to be represented. Can we expect that the error in representing by such a dictionary a previously unseen signal from the same source will be of similar magnitude as those for the given examples? We assume signals are generated from a fixed distribution, and study these questions from a statistical learning theory perspective.

We develop generalization bounds on the quality of the learned dictionary for two types of constraints on the coefficient selection, as measured by the expected  $L_2$  error in representation when the dictionary is used. For the case of  $l_1$  regularized coefficient selection we provide a generalization bound of the order of  $O(\sqrt{np \ln(m\lambda)/m})$ , where  $n$  is the dimension,  $p$  is the number of elements in the dictionary,  $\lambda$  is a bound on the  $l_1$  norm of the coefficient vector and  $m$  is the number of samples, which complements existing results. For the case of representing a new signal as a combination of at most  $k$  dictionary elements, we provide a bound of the order  $O(\sqrt{np \ln(mk)/m})$  under an assumption on the closeness to orthogonality of the dictionary (low Babel function). We further show that this assumption holds for *most* dictionaries in high dimensions in a strong probabilistic sense. Our results also include bounds that converge as  $1/m$ , not previously known for this problem. We provide similar results in a general setting using kernels with weak smoothness requirements.

**Keywords:** statistical machine learning, dictionary learning, generalization bounds, signal processing, kernel methods

## 1. Introduction

A common technique in processing signals from  $\mathcal{X} = \mathbb{R}^n$  is to use sparse representations; that is, to approximate each signal  $x$  by a “small” linear combination  $a$  of elements  $d_i$  from a dictionary  $D \in \mathcal{X}^p$ , so that  $x \approx Da = \sum_{i=1}^p a_i d_i$ . This has various uses detailed in Section 1.1. The smallness of  $a$  is often measured using either  $\|a\|_1$ , or the number of non

zero elements in  $a$ , often denoted  $\|a\|_0$ . The approximation error is measured here using a Euclidean norm appropriate to the vector space. We denote the approximation error of  $x$  using dictionary  $D$  and coefficients from a set  $A$  by

$$h_{A,D}(x) = \min_{a \in A} \|Da - x\|, \quad (1.1)$$

where  $A$  is one of the following sets determining the sparsity required of the representation:

$$H_k = \{a : \|a\|_0 \leq k\}$$

induces a “hard” sparsity constraint, which we also call  $k$  sparse representation, while

$$R_\lambda = \{a : \|a\|_1 \leq \lambda\}$$

induces a convex constraint that is considered a “relaxation” of the previous constraint.

The dictionary learning problem is to find a dictionary  $D$  minimizing

$$E(D) = \mathbb{E}_{x \sim \nu} h_{A,D}(x), \quad (1.2)$$

where  $\nu$  is a distribution over signals that is known to us only through samples from it. The problem addressed in this paper is the “generalization” (in the statistical learning sense) of dictionary learning: to what extent does the performance of a dictionary chosen based on a finite set of samples indicate its expected error in (1.2)? This clearly depends on the number of samples and other parameters of the problem such as the dictionary size. In particular, an obvious algorithm is to represent each sample using itself, if the dictionary is allowed to be as large as the sample, but the performance on unseen signals is likely to disappoint.

To state our goal more quantitatively, assume that an algorithm finds a dictionary  $D$  suited to  $k$  sparse representation, in the sense that the average representation error  $E_m(D)$  on the  $m$  examples given to the algorithm is low. Our goal is to bound the generalization error  $\varepsilon$ , which is the additional expected error that might be incurred:

$$E(D) \leq (1 + \eta)E_m(D) + \varepsilon, \quad (1.3)$$

where  $\eta \geq 0$  is sometimes zero, and the bound  $\varepsilon$  depends on the number of samples and problem parameters. Since efficient algorithms that find the optimal dictionary for a given set of samples (also known as empirical risk minimization, or ERM, algorithms) are not known for dictionary learning, we prove uniform convergence bounds that apply simultaneously over all admissible dictionaries  $D$ , thus bounding from above the sample complexity of the dictionary learning problem. In particular, such a result means that every procedure for approximate minimization of empirical error (empirical dictionary learning) is also a procedure for approximate dictionary learning (as defined above) in a similar sense.

Many analytic and algorithmic methods relying on the properties of finite dimensional Euclidean geometry can be applied in more general settings by applying kernel methods. These consist of treating objects that are not naturally represented in  $\mathbb{R}^n$  as having their similarity described by an inner product in an abstract *feature space* that is Euclidean. This allows the application of algorithms depending on the data only through a computation of inner products to such diverse objects as graphs, DNA sequences and text documents (?). Is it possible to extend the usefulness of dictionary learning techniques to this setting? We address sample complexity aspects of this question as well.

### 1.1. Background and related work

Sparse representations are by now standard practice in diverse fields such as signal processing, natural language processing, etc. Typically, the dictionary is assumed to be known. The motivation for sparse representations is indicated by the following results, in which we assume the signals come from  $\mathcal{X} = \mathbb{R}^n$  are normalized to have length 1, and the representation coefficients are constrained to  $A = H_k$  where  $k < n, p$  and typically  $h_{A,D}(x) \ll 1$ .

- **Compression:** If a signal  $x$  has an approximate sparse representation in some commonly known dictionary  $D$ , it can be stored or transmitted more economically with reasonable precision. Finding a good sparse representation can be computationally hard but if  $D$  fulfills certain geometric conditions, then its sparse representation is unique and can be found efficiently (see, e.g., [?](#)).
- **Denoising:** If a signal  $x$  has a sparse representation in some known dictionary  $D$ , and  $\tilde{x} = x + \nu$ , where the random noise  $\nu$  is Gaussian, then the sparse representation found for  $\tilde{x}$  will likely be very close to  $x$  (for example [?](#)).
- **Compressed sensing:** Assuming that a signal  $x$  has a sparse representation in some known dictionary  $D$  that fulfills certain geometric conditions, this representation can be approximately retrieved with high probability from a small number of random linear measurements of  $x$ . The number of measurements needed depends on the sparsity of  $x$  in  $D$  ([?](#)).

The implications of these results are significant when a dictionary  $D$  is known that sparsely represents simultaneously many signals. In some applications the dictionary is chosen based on prior knowledge, but in many applications the dictionary is learned based on a finite set of examples. To motivate dictionary learning, consider an image representation used for compression or denoising. Different types of images may have different properties (MRI images are not similar to scenery images), so that learning a dictionary specific to each type of images may lead to improved performance. The benefits of dictionary learning have been demonstrated in many applications ([??](#)).

Two extensively used techniques related to dictionary learning are Principal Component Analysis (PCA) and  $K$ -means clustering. The former finds a single subspace minimizing the sum of squared representation errors which is very similar to dictionary learning with  $A = H_k$  and  $p = k$ . The latter finds a set of locations minimizing the sum of squared distances between each signal and the location closest to it which is very similar to dictionary learning with  $A = H_1$  where  $p$  is the number of locations. Thus we could see dictionary learning as PCA with multiple subspaces, or as clustering where multiple locations are used to represent each signal. The sample complexities of both algorithms are well studied ([????](#)).

This paper does not address questions of computational cost, though they are very relevant. Finding optimal coefficients for  $k$  sparse representation (that is, minimizing (1.1) with  $A = H_k$ ) is NP-hard in general ([?](#)). Dictionary learning as the optimization problem of minimizing (1.2) is less well understood, even for empirical  $\nu$  (consisting of a finite number of samples), despite over a decade of work on related algorithms with good empirical results ([??????](#)).

The only prior work we are aware of that addresses generalization in dictionary learning, by ?, addresses the convex representation constraint  $A = R_\lambda$ ; we discuss the relation of our work to theirs in Section 2.

## 2. Results

Except where we state otherwise, we assume signals are generated in the unit sphere  $\mathbb{S}^{n-1}$ . Our results are:

**A new approach to dictionary learning generalization.** Our first main contribution is an approach to generalization bounds in dictionary learning that is complementary to the approach used by ?. The previous result, given below in Theorem 6 has generalization error bounds of order

$$O\left(\sqrt{p \min(p, n) \left(\lambda + \sqrt{\ln(m\lambda)}\right)^2 / m}\right)$$

on the squared representation error. A notable feature of this result is the weak dependence on the signal dimension  $n$ . In Theorem 1 we quantify the complexity of the class of functions  $h_{A,D}$  over all dictionaries whose columns have unit length, where  $A \subset R_\lambda$ . Combined with standard methods of uniform convergence this results in generalization error bounds  $\varepsilon$  of order  $O\left(\sqrt{np \ln(m\lambda)/m}\right)$  when  $\eta = 0$ . While our bound does depend strongly on  $n$ , this is acceptable in the case  $n < p$ , also known in the literature as the “over-complete” case (??). Note that our generalization bound applies with different constants to the representation error itself and many variants including the squared representation error, and has a weak dependence on  $\lambda$ . The dependence on  $\lambda$  is significant, for example, when  $\|a\|_1$  is used as a weighted penalty term by solving  $\min_a \|Da - X\| + \gamma \cdot \|a\|_1$ ; in this case  $\lambda = O(\gamma^{-1})$  may be quite large.

**Fast rates.** For the case  $\eta > 0$  our methods allow bounds of order  $O(np \ln(\lambda m)/m)$ . The main significance of this is in that the general statistical behavior they imply occurs in dictionary learning. For example, generalization error has a “proportional” component which is reduced when the empirical error is low. Whether fast rates results can be proved in the dimension free regime is an interesting question we leave open. Note that due to lower bounds by ? of order  $\sqrt{m^{-1}}$  on the  $k$ -means clustering problem, which corresponds to dictionary learning for 1-sparse representation, fast rates may be expected only with  $\eta > 0$ , as presented here.

We now describe the relevant function class and the bounds on its complexity, which are proved in Section 3. The resulting generalization bounds are given explicitly at the end of this section.

**Theorem 1** *For every  $\varepsilon > 0$ , the function class*

$$\mathcal{G}_\lambda = \left\{ h_{R_\lambda, D} : \mathbb{S}^{n-1} \rightarrow \mathbb{R} : D \in \mathbb{R}^{n \times p}, \|d_i\| \leq 1 \right\},$$

*taken as a metric space with the distance induced by  $\|\cdot\|_\infty$ , has a subset of cardinality at most  $(4\lambda/\varepsilon)^{np}$ , such that every element from the class is at distance at most  $\varepsilon$  from the subset.*

While we give formal definitions in Section 3, such a subset is called an  $\varepsilon$  cover, and such a bound on its cardinality is called a covering number bounds.

**Extension to  $k$  sparse representation.** Our second main contribution is to extend both our approach and that of ? to provide generalization bounds for dictionaries for  $k$  sparse representations, by using a bound  $\lambda$  on the  $l_1$  norm of the representation coefficients when the dictionaries are close to orthogonal. Distance from orthogonality is measured by the Babel function (which, for example, upper bounds the magnitude of the maximal inner product between distinct dictionary elements) defined below and discussed in more detail in Section 4.

**Definition 2 (Babel function, ?)** For any  $k \in \mathbb{N}$ , the Babel function  $\mu_k : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^+$  is defined by:

$$\mu_k(D) = \max_{i \in \{1, \dots, p\}} \max_{\Lambda \subset \{1, \dots, p\} \setminus \{i\}; |\Lambda|=k} \sum_{j \in \Lambda} |\langle d_j, d_i \rangle|.$$

The following proposition, which is proved in Section 3, bounds the 1-norm of the dictionary coefficients for a  $k$  sparse representation and also follows from analysis previously done by ??.

**Proposition 3** Let each column  $d_i$  of  $D$  fulfill  $\|d_i\| \in [1, \gamma]$  and  $\mu_{k-1}(D) \leq \delta < 1$ , then a coefficient vector  $a \in \mathbb{R}^p$  minimizing the  $k$ -sparse representation error  $h_{H_k, D}(x)$  exists which has  $\|a\|_1 \leq \gamma k / (1 - \delta)$ .

We now consider the class of all  $k$  sparse representation error functions. We prove in Section 3 the following bound on the complexity of this class.

**Corollary 4** The function class

$$\mathcal{F}_{\delta, k} = \left\{ h_{H_k, D} : \mathbb{S}^{n-1} \rightarrow \mathbb{R} : \mu_{k-1}(D) < \delta, d_i \in \mathbb{S}^{n-1} \right\},$$

taken as a metric space with the metric induced by  $\|\cdot\|_\infty$ , has a covering number bound of  $(4k / (\varepsilon(1 - \delta)))^{np}$ .

The dependence of the last two results on  $\mu_{k-1}(D)$  means that the resulting bounds will be meaningful only for algorithms which explicitly or implicitly prefer near orthogonal dictionaries. Contrast this to Theorem 1 which has no significant conditions on the dictionary.

**Asymptotically almost all dictionaries are near orthogonal.** A question that arises is what values of  $\mu_{k-1}$  can be expected for parameters  $n, p, k$ ? We shed some light on this question through the following probabilistic result, which we discuss in Section 4 and prove in the full version.

**Theorem 5** Suppose that  $D$  consist of  $p$  vectors chosen uniformly and independently from  $\mathbb{S}^{n-1}$ . Then we have

$$P \left( \mu_k > \frac{1}{2} \right) \leq \frac{1}{\left( e^{(n-2)/(10k \ln p)^2} - 1 \right)}.$$

Since low values of the Babel function have implications to representation finding algorithms, this result is of interest also outside the context of dictionary learning. Essentially it means that random dictionaries of size sub-exponential in  $(n - 2)/k^2$  have low Babel function.

**New generalization bounds for  $l_1$  case.** The covering number bound of Theorem 1 implies several generalization bounds for the problem of dictionary learning for  $l_1$  regularized representation which we give here. These differ from those by ? in depending more strongly on the dimension of the space, but less strongly on the particular regularization term. We first give the relevant specialization of the result by ? for comparison and for reference as we will later build on it. This result is independent of the dimension  $n$  of the underlying space, thus the Euclidean unit ball  $B$  may be that of a general Hilbert space, and the errors measured by  $h_{A,D}$  are in the same norm.

**Theorem 6 (?)** *Let  $A \subset R_\lambda$ , and let  $\nu$  be any distribution on the unit sphere  $B$ . Then with probability at least  $1 - e^{-x}$  over the  $m$  samples in  $E_m$  drawn according to  $\nu$ , for all dictionaries  $D \subset B$  with cardinality  $p$ :*

$$Eh_{A,D}^2 \leq E_m h_{A,D}^2 + \sqrt{\frac{p^2 \left(14\lambda + 1/2\sqrt{\ln(16m\lambda^2)}\right)^2}{m}} + \sqrt{\frac{x}{2m}}.$$

Using the covering number bound of Theorem 1 and a bounded differences concentration inequality (see Lemma 21), we obtain the following result. The details are given in Section 3.

**Theorem 7** *Let  $\lambda > e/4$ , with  $\nu$  a distribution on  $\mathbb{S}^{n-1}$ . Then with probability at least  $1 - e^{-x}$  over the  $m$  samples in  $E_m$  drawn according to  $\nu$ , for all  $D$  with unit length columns:*

$$Eh_{R_\lambda,D} \leq E_m h_{R_\lambda,D} + \sqrt{\frac{np \ln(4\sqrt{m}\lambda)}{2m}} + \sqrt{\frac{x}{2m}} + \sqrt{\frac{4}{m}}.$$

Using the same covering number bound and the general result Corollary 23 (given in Section 3), we obtain the following fast rates result. A slightly more general result is easily derived by using Proposition 22 instead.

**Theorem 8** *Let  $\lambda > e/4$ ,  $np \geq 20$  and  $m \geq 5000$  with  $\nu$  a distribution on  $\mathbb{S}^{n-1}$ . Then with probability at least  $1 - e^{-x}$  over the  $m$  samples in  $E_m$  drawn according to  $\nu$ , for all  $D$  with unit length columns:*

$$Eh_{R_\lambda,D} \leq 1.1E_m h_{R_\lambda,D} + 9\frac{np \ln(4\lambda m) + x}{m}.$$

**Generalization bounds for  $k$  sparse representation.** Proposition 3 and Corollary 4 imply certain generalization bounds for the problem of dictionary learning for  $k$  sparse representation, which we give here.

A straight forward combination of Theorem 2 of ? (given here as Theorem 6) and Proposition 3 results in the following theorem.

**Theorem 9** Let  $\delta < 1$  with  $\nu$  a distribution on  $\mathbb{S}^{n-1}$ . Then with probability at least  $1 - e^{-x}$  over the  $m$  samples in  $E_m$  drawn according to  $\nu$ , for all  $D$  s.t.  $\mu_{k-1}(D) \leq \delta$  and with unit length columns:

$$Eh_{H_k,D}^2 \leq E_m h_{H_k,D}^2 + \frac{p}{\sqrt{m}} \left( \frac{14k}{1-\delta} + \frac{1}{2} \sqrt{\ln \left( 16m \left( \frac{k}{1-\delta} \right)^2 \right)} \right) + \sqrt{\frac{x}{2m}}.$$

In the case of clustering we have  $k = 1$  and  $\delta = 0$  and this result approaches the rates of ?.

The following theorems follow from the covering number bound of Corollary 4 and applying the general results of Section 3 as for the  $l_1$  sparsity results.

**Theorem 10** Let  $\delta < 1$  with  $\nu$  a distribution on  $\mathbb{S}^{n-1}$ . Then with probability at least  $1 - e^{-x}$  over the  $m$  samples in  $E_m$  drawn according to  $\nu$ , for all  $D$  s.t.  $\mu_{k-1}(D) \leq \delta$  and with unit length columns:

$$Eh_{H_k,D} \leq E_m h_{H_k,D} + \sqrt{\frac{np \ln \frac{4\sqrt{mk}}{1-\delta}}{2m}} + \sqrt{\frac{x}{2m}} + \sqrt{\frac{4}{m}}.$$

**Theorem 11** Let  $\delta < 1$ ,  $np \geq 20$  and  $m \geq 5000$  with  $\nu$  a distribution on  $\mathbb{S}^{n-1}$ . Then with probability at least  $1 - e^{-x}$  over the  $m$  samples in  $E_m$  drawn according to  $\nu$ , for all  $D$  s.t.  $\mu_{k-1}(D) \leq \delta$  and with unit length columns:

$$Eh_{H_k,D} \leq 1.1 E_m h_{H_k,D} + 9 \frac{np \ln \left( \frac{4\sqrt{mk}}{1-\delta} \right) + x}{m}.$$

**Generalization bounds for dictionary learning in feature spaces.** We further consider applications of dictionary learning to signals that are not represented as elements in a vector space, or that have a very high (possibly infinite) dimension.

In addition to providing an approximate reconstruction of signals, sparse representation can also be considered as a form of analysis, if we treat the choice of non zero coefficients and their magnitude as features of the signal. In the domain of images, this has been used to perform classification (in particular, face recognition) by ?. Such analysis does not require that the data itself be represented in  $\mathbb{R}^n$  (or in any vector space); it is enough that the similarity between data elements is induced from an inner product in a feature space. This requirement is fulfilled by using an appropriate kernel function.

**Definition 12** Let  $\mathcal{R}$  be a set of data representations, and let the kernel function  $\kappa : \mathcal{R}^2 \rightarrow \mathbb{R}$  and the feature mapping  $\phi : \mathcal{R} \rightarrow \mathcal{H}$  be such that:

$$\kappa(x, y) = \langle \phi(x), \phi(y) \rangle$$

where  $\mathcal{H}$  is some Hilbert space.

As a concrete example, choose a sequence of  $n$  words, and let  $\phi$  map a document to the vector of counts of appearances of each word in it (also called bag of words). Treating  $\kappa(a, b) = \langle \phi(a), \phi(b) \rangle$  as the similarity between documents  $a$  and  $b$ , is the well known “bag of words” approach, applicable to many document related tasks (?). Then the statement  $\phi(a) + \phi(b) \approx \phi(c)$  does not imply that  $c$  can be reconstructed from  $a$  and  $b$ , but we might consider it indicative of the content of  $c$ . The dictionary of elements used for representation could be decided via dictionary learning, and it is natural to choose the dictionary so that the bags of words of documents are approximated well by small linear combinations of those in the dictionary.

As the example above suggests, the kernel dictionary learning problem is to find a dictionary  $D$  minimizing

$$\mathbb{E}_{x \sim \nu} h_{\phi, A, D}(x),$$

where we consider the representation error function

$$h_{\phi, A, D}(x) = \min_{a \in A} \|(\Phi D) a - \phi(x)\|_{\mathcal{H}},$$

in which  $\Phi$  acts as  $\phi$  on the elements of  $D$ ,  $A \in \{R_\lambda, H_k\}$ , and the norm  $\|\cdot\|_{\mathcal{H}}$  is that induced by the kernel on the feature space  $\mathcal{H}$ .

Analogues of all the generalization bounds mentioned so far can be replicated in the kernel setting. The dimension free results of ? apply most naturally in this setting, and may be combined with our results to cover also dictionaries for  $k$  sparse representation, under reasonable assumptions on the kernel.

**Proposition 13** *Let  $\nu$  be any distribution on  $\mathcal{R}$  such that  $x \sim \nu$  implies that  $\phi(x)$  is in the unit ball  $B_{\mathcal{H}}$  of  $\mathcal{H}$  with probability 1. Then with probability at least  $1 - e^{-x}$  over the  $m$  samples in  $E_m$  drawn according to  $\nu$ , for all  $D \subset \mathcal{R}$  with cardinality  $p$  such that  $\Phi D \subset B_{\mathcal{H}}$  and  $\mu_{k-1}(\Phi D) \leq \delta < 1$ :*

$$Eh_{\phi, H_k, D}^2 \leq E_m h_{\phi, H_k, D}^2 + \sqrt{\frac{p^2 \left( 14k/(1-\delta) + 1/2 \sqrt{\ln \left( 16m \left( \frac{k}{1-\delta} \right)^2 \right)} \right)^2}{m}} + \sqrt{\frac{x}{2m}}.$$

Note that in  $\mu_{k-1}(\Phi D)$  the Babel function is defined in terms of inner products in  $\mathcal{H}$ , and can therefore be computed efficiently by applications of the kernel.

In Section 5 we prove the above result and also cover number bounds as in the linear case considered before. In the current setting, these bounds depend on the Hölder smoothness order  $\alpha$  of the feature mapping  $\phi$ . Formal definitions are given in Section 5 but as an example, the well known Gaussian kernel has  $\alpha = 1$ . We give now one of the generalization bounds using this method.

**Theorem 14** *Let  $\mathcal{R}$  have  $\varepsilon$  covers of order  $(C/\varepsilon)^n$ . Let  $\kappa : \mathcal{R}^2 \rightarrow \mathbb{R}^+$  be a kernel function s.t.  $\kappa(x, y) = \langle \phi(X), \phi(Y) \rangle$ , for  $\phi$  which is uniformly  $L$ -Hölder of order  $\alpha > 0$  over  $\mathcal{R}$ , and let  $\gamma = \max_{x \in \mathcal{R}} \|\phi(x)\|_{\mathcal{H}}$ . Let  $\delta < 1$ , and  $\nu$  any distribution on  $\mathcal{R}$ , then with probability at*

least  $1 - e^{-x}$  over the  $m$  samples in  $E_m$  drawn according to  $\nu$ , for all dictionaries  $D \subset \mathcal{R}$  of cardinality  $p$  s.t.  $\mu_{k-1}(\Phi D) \leq \delta < 1$  (where  $\Phi$  acts like  $\phi$  on columns):

$$Eh_{H_k, D} \leq E_m h_{H_k, D} + \gamma \left( \sqrt{\frac{np \ln \left( \sqrt{m} C^\alpha \frac{k\gamma^2 L}{1-\delta} \right)}{2\alpha m}} + \sqrt{\frac{x}{2m}} \right) + \sqrt{\frac{4}{m}}.$$

The covering number bounds needed to prove this theorem and analogs for the other generalization bounds are proved in Section 5.

### 3. Covering numbers of $\mathcal{G}_\lambda$ and $\mathcal{F}_{\delta, k}$

The main content of this section is the proof of Theorem 1 and Corollary 4. We also show that in the  $k$  sparse representation setting a finite bound on  $\lambda$  does not occur generally thus an additional restriction, such as the near-orthogonality on the set of dictionaries on which we rely in this setting, is necessary. Lastly, we recall known results from statistical learning theory that link covering numbers to generalization bounds.

We recall the definition of the covering numbers we wish to bound. ? give a textbook introduction to covering numbers and their application to generalization bounds.

**Definition 15 (Covering number)** Let  $(M, d)$  be a metric space and  $S \subset M$ . Then the  $\varepsilon$  covering number of  $S$  defined as  $N(\varepsilon, S, d) = \min \{ |A| \mid A \subset M \text{ and } S \subset (\bigcup_{a \in A} B_d(a, \varepsilon)) \}$  is the size of the minimal  $\varepsilon$  cover of  $S$  using  $d$ .

To prove Theorem 1 and Corollary 4 we first note that the space of all possible dictionaries is a subset of a unit ball in a Banach space of dimension  $np$  (with a norm specified below). Thus (see formalization in Proposition 5 of ?) the space of dictionaries has an  $\varepsilon$  cover of size  $(4/\varepsilon)^{np}$ . We also note that a uniformly  $L$  Lipschitz mapping between metric spaces converts  $\varepsilon/L$  covers into  $\varepsilon$  covers. Then it is enough to show that  $\Psi_\lambda$  defined as  $D \mapsto h_{R_\lambda, D}$  and  $\Phi_k$  defined as  $D \mapsto h_{H_k, D}$  are uniformly Lipschitz (when  $\Phi_k$  is restricted to the dictionaries with  $\mu_{k-1}(D) \leq c < 1$ ). The proof of these Lipschitz properties is our next goal, in the form of Lemmas 18 and 19.

The first step is to be clear about the metrics we consider over the spaces of dictionaries and of error functions.

**Definition 16 (Induced matrix norm)** Let  $p, q \geq 1$ , then a matrix  $A \in \mathbb{R}^{n \times m}$  can be considered as an operator  $A : (\mathbb{R}^m, \|\cdot\|_p) \rightarrow (\mathbb{R}^n, \|\cdot\|_q)$ . The  $p, q$  induced norm is  $\|A\|_{p,q} \triangleq \sup_{x \in \mathbb{R}^m, \|x\|_p=1} \|Ax\|_q$ .

**Lemma 17** For any matrix  $D$ ,  $\|D\|_{1,2}$  is equal to the maximal Euclidean norm of any column in  $D$ .

**Proof** That the maximal norm of a column bounds  $\|D\|_{1,2}$  can be seen geometrically;  $Da / \|a\|_1$  is a convex combination of column vectors, then  $\|Da\|_2 \leq \max_{d_i} \|d_i\|_2 \|a\|_1$  because a norm is convex. Equality is achieved for  $a = e_i$ , where  $d_i$  is the column of maximal

■

norm.

The images of  $\Psi_\lambda$  and  $\Phi_k$  are sets of representation error functions—each dictionary induces a set of precisely representable signals, and a representation error function is simply a map of distances from this set. Representation error functions are clearly continuous, 1-Lipschitz, and into  $[0, 1]$ . In this setting, a natural norm over the images is the supremum norm  $\|\cdot\|_\infty$ .

**Lemma 18** *The function  $\Psi_\lambda$  is  $\lambda$ -Lipschitz from  $(\mathbb{R}^{n \times m}, \|\cdot\|_{1,2})$  to  $C(\mathbb{S}^{n-1})$ .*

**Proof** Let  $D$  and  $D'$  be two dictionaries whose corresponding elements are at most  $\varepsilon > 0$  far from one another. Let  $x$  be a unit signal and  $Da$  an optimal representation for it. Then  $\|(D - D')a\|_2 \leq \|D - D'\|_{1,2} \|a\|_1 \leq \varepsilon\lambda$ . If  $D'a$  is very close to  $Da$  in particular it is not a much worse representation of  $x$ , and replacing it with the optimal representation under  $D'$ , we have  $h_{R_\lambda, D'}(x) \leq h_{R_\lambda, D}(x) + \varepsilon\lambda$ . By symmetry we have  $|\Psi_\lambda(D)(x) - \Psi_\lambda(D')(x)| \leq \lambda\varepsilon$ . This holds for all unit signals, then  $\|\Psi_\lambda(D) - \Psi_\lambda(D')\|_\infty \leq \lambda\varepsilon$ . ■

We now provide a proof for Proposition 3 which is used in the corresponding treatment for covering numbers under  $k$  sparsity.

**Proof** (Of Proposition 3) Let  $D^k$  be a submatrix of  $D$  whose  $k$  columns from  $D$  achieve the minimum on  $h_{H_k, D}(x)$  for  $x \in \mathbb{S}^{n-1}$ . We now consider the Gram matrix  $G = (D^k)^\top D^k$  whose diagonal entries are the norms of the elements of  $D^k$ , therefore at least 1. By the Gersgorin theorem (?), each eigenvalue of a square matrix is “close” to a diagonal entry of the matrix; the absolute difference between an eigenvalue and its diagonal entry is upper bounded by the sum of the absolute values of the remaining entries of the same row. Since a row in  $G$  corresponds to the inner products of an element from  $D^k$  with every element from  $D^k$ , this sum is upper bounded by  $\delta$  for all rows. Then we conclude the eigenvalues of the Gram matrix are lower bounded by  $1 - \delta > 0$ . Then in particular  $G$  has a symmetric inverse  $G^{-1}$  whose eigenvalues are positive and bounded from above by  $1/(1 - \delta)$ . The maximal magnitude of an eigenvalue of a symmetric matrix coincides with its induced norm  $\|\cdot\|_{2,2}$ , therefore  $\|G^{-1}\|_{2,2} \leq 1/(1 - \delta)$ .

Linear dependence of elements of  $D^k$  would imply a non-trivial nullspace for the invertible  $G$ . Then the elements of  $D^k$  are linearly independent, which implies that the unique optimal representation of  $x$  as a linear combination of the columns of  $D^k$  is  $D^k a$  with

$$a = \left( (D^k)^\top D^k \right)^{-1} (D^k)^\top x.$$

Using the above and the definition of induced matrix norms, we have

$$\|a\|_2 \leq \left\| \left( (D^k)^\top D^k \right)^{-1} \right\|_{2,2} \left\| (D^k)^\top x \right\|_2 \leq 1/(1 - \delta) \left\| (D^k)^\top x \right\|_2.$$

The vector  $(D^k)^\top x$  is in  $\mathbb{R}^k$  and by the Cauchy Schwartz inequality  $\langle d_i, x \rangle \leq \gamma$ , then  $\left\| (D^k)^\top x \right\|_2 \leq \sqrt{k} \left\| (D^k)^\top x \right\|_\infty \leq \sqrt{k}\gamma$ . Since only  $k$  entries of  $a$  are non zero,  $\|a\|_1 \leq$

$$\sqrt{k} \|a\|_2 \leq k\gamma/(1-\delta). \quad \blacksquare$$

**Lemma 19** *The function  $\Phi_k$  is a  $k/(1-\delta)$ -Lipschitz mapping from the set of normalized dictionaries with  $\mu_{k-1}(D) < \delta$  with the metric induced by  $\|\cdot\|_{1,2}$  to  $C(\mathbb{S}^{n-1})$ .*

The proof of this lemma is the same as that of Lemma 18, except that  $a$  is taken to be an optimal representation that fulfills  $\|a\|_1 \leq \lambda = k/(1 - \mu_{k-1}(D))$ , whose existence is guaranteed by Proposition 3.

This concludes the proof of Theorem 1 and Corollary 4.

The next theorem shows that unfortunately,  $\Phi$  is *not* uniformly  $L$ -Lipschitz for any constant  $L$ , requiring its restriction to an appropriate subset of the dictionaries.

**Theorem 20** *For any  $1 < k < n, p$ , there exists  $c > 0$  and  $q$ , such that for every  $\varepsilon > 0$ , there exist  $D, D'$  such that  $\|D - D'\|_{1,2} < \varepsilon$  but  $|h_{H_k, D}(q) - h_{H_k, D'}(q)| > c$ .*

**Proof** First we show that for any dictionary  $D$  there exist  $c > 0$  and  $x \in \mathbb{S}^{n-1}$  such that  $h_{H_k, D}(x) > c$ . Let  $\nu_{\mathbb{S}^{n-1}}$  be the uniform probability measure on the sphere, and  $A_c$  the probability assigned by it to the set within  $c$  of a  $k$  dimensional subspace. As  $c \searrow 0$ ,  $A_c$  also tends to zero, then there exists  $c > 0$  s.t.  $\binom{p}{k} A_c < 1$ . Then for that  $c$  and any dictionary  $D$  there exists a set of positive measure on which  $h_{H_k, D} > c$ , let  $q$  be a point in this set. Since  $h_{H_k, D}(x) = h_{H_k, D}(-x)$ , we may assume without loss of generality that  $\langle e_1, q \rangle \geq 0$ .

We now fix the dictionary  $D$ ; its first  $k-1$  elements are the standard basis  $\{e_1, \dots, e_{k-1}\}$ , its  $k$ th element is  $D_k = \sqrt{1-\varepsilon^2/4}e_1 + \varepsilon e_k/2$ , and the remaining elements are chosen arbitrarily. Now construct  $D'$  to be identical to  $D$  except its  $k$ th element is  $v = \sqrt{1-\varepsilon^2/4}e_1 + lq$  choosing  $l$  so that  $\|v\|_2 = 1$ . Then there exist  $a, b \in \mathbb{R}$  such that  $q = aD'_1 + bD'_k$  and we have  $h_{H_k, D'}(q) = 0$ , fulfilling the second part of the theorem. On the other hand, since  $\langle e_1, q \rangle \geq 0$ , we have  $l \leq \varepsilon/2$ , and then we find  $\|D - D'\|_{1,2} = \|\varepsilon e_k/2 - lq\|_2 \leq \|\varepsilon e_k/2\| + \|lq\| = \varepsilon/2 + l \leq \varepsilon$ .  $\blacksquare$

To conclude the generalization bounds of Theorems 7, 8, 10, 11 and 14 from the covering number bounds we have provided, we use the following results. The first result is a straight forward application of Hoeffding's inequality, a union bound and the  $l_\infty$  cover number bounds. The second result<sup>1</sup> (along with its corollary) gives fast rate bounds and uses the  $\|\cdot\|_\infty$  cover number bounds to achieve better constants for this problem than the more general results by ? and ?.

**Lemma 21** *Let  $\mathcal{F}$  be a class of  $[0, B]$  functions with covering number bound  $(C/\varepsilon)^d > e/B^2$  under the supremum norm. Then for every  $x > 0$ , with probability of at least  $1 - e^{-x}$  over the  $m$  samples in  $E_m$  chosen according to  $\nu$ , for all  $f \in \mathcal{F}$ :*

$$Ef \leq E_m f + B \left( \sqrt{\frac{d \ln(C\sqrt{m})}{2m}} + \sqrt{\frac{x}{2m}} \right) + \sqrt{\frac{4}{m}}.$$

---

1. We thank Andreas Maurer for suggesting this result and a proof elaborated in the full version.

**Proposition 22** Let  $\mathcal{F}$  be a class of  $[0, 1]$  functions that can be covered for any  $\varepsilon > 0$  by at most  $(C/\varepsilon)^d$  balls of radius  $\varepsilon$  in the  $L_\infty$  metric where  $C \geq e$  and  $\beta > 0$ . Then with probability at least  $1 - \exp(-x)$ , we have for all  $f \in \mathcal{F}$ :

$$Ef \leq (1 + \beta) E_m f + K(d, m, \beta) \frac{d \ln(Cm) + x}{m},$$

$$\text{where } K(d, m, \beta) = \sqrt{2 \left( \frac{9}{\sqrt{m}} + 2 \right) \left( \frac{d+3}{3d} \right) + 1} + \left( \frac{9}{\sqrt{m}} + 2 \right) \left( \frac{d+3}{3d} \right) + 1 + \frac{1}{2\beta}.$$

The corollary we use to obtain Theorems 8 and 11 follows because  $K(d, m, \beta)$  is non-increasing in  $d, m$ .

**Corollary 23** Let  $\mathcal{F}, x$  be as above. For  $d \geq 20$ ,  $m \geq 5000$  and  $\beta = 0.1$  we have with probability at least  $1 - \exp(-x)$  for all  $f \in \mathcal{F}$ :

$$Ef \leq 1.1 E_m f + 9 \frac{d \ln(Cm) + x}{m}.$$

#### 4. On the Babel function

The Babel function is one of several metrics defined in the sparse representations literature to quantify an "almost orthogonality" property that dictionaries may enjoy. Such properties have been shown to imply theoretical properties such as uniqueness of the optimal  $k$  sparse representation. In the algorithmic context, ? and ? use the Babel function to show that particular efficient algorithms for finding sparse representations fulfill certain quality guarantees when applied to such dictionaries. This reinforces the practical importance of the learnability of this class of dictionary. We proceed to discuss some elementary properties of the Babel function, and then state a bound on the proportion of dictionaries having sufficiently good Babel function.

Measures of orthogonality are typically defined in terms of inner products between the elements of the dictionary. Perhaps the simplest of these measures of orthogonality is the following special case of the Babel function.

**Definition 24** The coherence of a dictionary  $D$  is  $\mu_1(D) = \max_{i \neq j} |\langle d_i, d_j \rangle|$ .

The Babel function considers sums of  $k$  inner products at a time rather than the maximum over all inner products, and thus better quantifies the effects of non orthogonality on representing a signal with particular level  $k + 1$  of sparsity. As a particular example of the finer grained control  $\mu_k$  when compared to  $\mu_1$ , consider the following example. Let  $D$  consist of  $k$  pairs of elements, so that the subspace spanned by each pair is orthogonal to all other elements, and such that the inner product between the elements of any single pair is half. In this case  $\mu_k(D) = \mu_1(D) = 1/2$ . However note that to ensure  $\mu_k < 1$  only restricting  $\mu_1$  requires the constraint  $\mu_1(D) < 1/k$ , which is not fulfilled in our example.

To better understand  $\mu_k(D)$ , we consider first its extreme values. When  $\mu_k(D) = 0$ , for any  $k > 1$ , this means that  $D$  is an orthogonal set (therefore  $p \leq n$ ). The maximal value of  $\mu_k(D)$  is  $k$ , and occurs only if some dictionary element is repeated (up to sign) at least  $k + 1$  times.

A well known generic class of dictionaries with more elements than a basis is that of *frames* (see ?), which include many wavelet systems and filter banks. Some frames can be trivially seen to fulfill our condition on the Babel function.

**Proposition 25** *Let  $D \in \mathbb{R}^{n \times p}$  be a frame of  $\mathbb{R}^n$ , so that for every  $v \in \mathbb{S}^{n-1}$  we have that  $\sum_{i=1}^n |\langle v, d_i \rangle|^2 \leq B$ , with  $\|d_i\|_2 = 1$  for all  $i$ , and  $B < 1 + 1/k$ . Then  $\mu_{k-1}(D) < 1$ .*

This may be easily verified by considering the inner products of any dictionary element with any other  $k$  elements as a vector in  $\mathbb{R}^k$ ; the frame condition bounds its squared Euclidean norm by  $B - 1$  (we remove the inner product of the element with itself in the frame expression). Then use the equivalence of  $l_1$  and  $l_2$  norms.

#### 4.1. Proportion of dictionaries with $\mu_{k-1}(D) < \delta$

We return to the question of the prevalence of dictionaries having  $\mu_{k-1} < \delta$ . Are almost all dictionaries such? If the answer is affirmative, it implies that Theorem 11 is quite strong, and representation finding algorithms such as basis pursuit are almost always exact, which might help prove properties of dictionary learning algorithms. If the opposite is true and few dictionaries have low Babel function, the results of this paper are weak. While there might be better probability measures on the space of dictionaries, we consider one that seems natural: suppose that a dictionary  $D$  is constructed by choosing  $p$  unit vectors uniformly from  $\mathbb{S}^{n-1}$ ; what is the probability that  $\mu_{k-1}(D) < \delta$ ?

Theorem 5 gives us the following answer to this question. Under the assumption that the sparsity parameter  $k$  grows slowly, if at all, as  $n \nearrow \infty$  (specifically, that  $k \ln p = o(\sqrt{n})$ ), this theorem implies that asymptotically *almost all dictionaries under the Lebesgue measure are learnable*.

### 5. Dictionary learning in feature spaces

We propose in Section 2 a scenario in which dictionary learning is performed in a feature space corresponding to a kernel function. Here we show how to adapt the different generalization bounds discussed in this paper for the particular case of  $\mathbb{R}^n$  to more general feature spaces, and the dependence of the sample complexities on the properties of the kernel function or the corresponding feature mapping. We begin with the relevant specialization of the results of ? which have the simplest dependence on the kernel, and then discuss the extensions to  $k$  sparse representation and to the cover number techniques presented in the current work.

Theorem 6 applies as is to the feature space, under the simple assumption that the dictionary elements and signals are in its unit ball which is guaranteed by some kernels such as the Gaussian kernel. Then we take  $\nu$  on the unit ball of  $\mathcal{H}$  to be induced by some distribution  $\nu'$  on the domain of the kernel, and the theorem applies to any such  $\nu'$  on  $\mathcal{R}$ . Nothing more is required if the representation is chosen from  $R_\lambda$ . The corresponding generalization bound for  $k$  sparse representations when the dictionary elements are near orthogonal in the feature space is given in Proposition 13.

**Proof** (Of Proposition 13) Proposition 3 applies with the Euclidean norm of  $\mathcal{H}$ , and  $\gamma = 1$ . We apply Theorem 6 with  $\lambda = k/(1 - \delta)$ . ■

The results so far show that generalization in dictionary learning can occur despite the potentially infinite dimension of the feature space, without considering practical issues of representation and computation. We now make the domain and applications of the kernel explicit in order to address a basic computational question, and allow the use of cover number based generalization bounds to prove Theorem 14. We now consider signals represented in a metric space  $(\mathcal{R}, d)$ , in which similarity is measured by the kernel  $\kappa$  corresponding to the feature map  $\phi : \mathcal{R} \rightarrow \mathcal{H}$ . The elements of a dictionary  $D$  are now from  $\mathcal{R}$ , and we denote  $\Phi D$  their mapping by  $\phi$  to  $\mathcal{H}$ . The representation error function used is  $h_{\phi, A, D}$ .

We now show that the approximation error in the feature space is a quadratic function of the coefficient vector; the quadratic function for particular  $D$  and  $x$  may be found by applications of the kernel.

**Proposition 26** *Computing the representation error at a given  $x, a, D$  requires  $O(p^2)$  kernel applications in general, and only  $O(k^2 + p)$  when  $a$  is  $k$  sparse.*

The squared error expands to

$$\sum_{i=1}^p a_i \sum_{j=1}^p a_j \kappa(d_i, d_j) + \kappa(x, x) - 2 \sum_{i=1}^p a_i \kappa(x, d_i).$$

We note that the  $k$  sparsity constraint on  $a$  poses algorithmic difficulties beyond those addressed here. Some of the common approaches to these, such as orthogonal matching pursuit (?), also depend on the data only through their inner products, and may therefore be adapted to the kernel setting.

The cover number bounds depend strongly on the dimension of the space of dictionary elements. Taking  $\mathcal{H}$  as the space of dictionary elements is the simplest approach, but may lead to vacuous or weak bounds, for example in the case of the Gaussian kernel whose feature space is infinite dimensional. Instead we propose to use the space of data representations  $\mathcal{R}$ , whose dimensions are generally bounded by practical considerations. In addition, we will assume that the kernel is not “too wild” in the following sense.

**Definition 27** *Let  $L, \alpha > 0$ , and let  $(A, d')$  and  $(B, d)$  be metric spaces. We say a mapping  $f : A \rightarrow B$  is uniformly  $L$  Hölder of order  $\alpha$  on a set  $S \subset A$  if  $\forall x, y \in S$ , the following bound holds:*

$$d(f(x), f(y)) \leq L \cdot d'(x, y)^\alpha.$$

The relevance of this smoothness condition is as follows.

**Lemma 28** *A Hölder function maps an  $\varepsilon$  cover of  $S$  to an  $L\varepsilon^\alpha$  cover of its image  $f(S)$ . Thus, to obtain an  $\varepsilon$  cover of the image of  $S$ , it is enough to begin with an  $(\varepsilon/L)^{1/\alpha}$  cover of  $S$ .*

A Hölder feature map  $\phi$  allows us to bound the cover numbers of the dictionary elements in  $\mathcal{H}$  using their cover number bounds in  $\mathcal{R}$ . Note that not every kernel corresponds to a Hölder feature map (the Dirac  $\delta$  kernel is a counter example: any two distinct elements are

mapped to elements at a mutual distance of 1), and sometimes analyzing the feature map is harder than analyzing the kernel. The following lemma bounds the geometry of the feature map using that of the kernel.

**Lemma 29** *Let  $\kappa(x, y) = \langle \phi(x), \phi(y) \rangle$ , and assume further that  $\kappa$  fulfills a Hölder condition of order  $\alpha$  uniformly in each parameter, that is,  $|\kappa(x, y) - \kappa(x + h, y)| \leq L \|h\|^\alpha$ . Then  $\phi$  uniformly fulfills a Hölder condition of order  $\alpha/2$  with constant  $\sqrt{2L}$ .*

This result is not sharp. For example, for the Gaussian case, both kernel and the feature map are Hölder order 1.

**Proof** Using the Hölder condition, we have that  $\|\phi(x) - \phi(y)\|_{\mathcal{H}}^2 = \kappa(x, x) - \kappa(x, y) + \kappa(y, y) - \kappa(x, y) \leq 2L \|x - y\|^\alpha$ . All that remains is to take the square root of both sides. ■

For a given feature mapping  $\phi$ , set of representations  $\mathcal{R}$ , we define two families of function classes so:

$$\begin{aligned}\mathcal{W}_{\phi, \lambda} &= \{h_{\phi, R_\lambda, D} : D \in \mathcal{D}^p\} \text{ and} \\ \mathcal{Q}_{\phi, k, \delta} &= \{h_{\phi, H_k, D} : D \in \mathcal{D}^p \wedge \mu_{k-1}(\Phi D) \leq \delta\}.\end{aligned}$$

The next proposition completes this section by giving the cover number bounds for the representation error function classes induced by appropriate kernels, from which various generalization bounds easily follow, such as Theorem 14.

**Proposition 30** *Let  $\mathcal{R}$  be a set of representations with a cover number bound of  $(C/\varepsilon)^n$ , and let either  $\phi$  be uniformly  $L$  Hölder condition of order  $\alpha$  on  $\mathcal{R}$ , or  $\kappa$  be uniformly  $L$  Hölder of order  $2\alpha$  on  $\mathcal{R}$  in each parameter, and let  $\gamma = \sup_{d \in \mathcal{D}} \|\phi(d)\|_{\mathcal{H}}$ . Then the function classes  $\mathcal{W}_{\phi, \lambda}$  and  $\mathcal{Q}_{\phi, k, \delta}$  taken as metric spaces with the supremum norm, have  $\varepsilon$  covers of cardinalities at most  $(C(\lambda\gamma L/\varepsilon)^{1/\alpha})^{np}$  and  $(C(k\gamma^2 L/(\varepsilon(1-\delta)))^{1/\alpha})^{np}$ , respectively.*

**Proof** We first consider the case of  $l_1$  constrained coefficients. If  $\|a\|_1 \leq \lambda$  and also  $\max_{d \in \mathcal{D}} \|\phi(d)\|_{\mathcal{H}} \leq \gamma$  then by considerations applied in Section 3, to obtain an  $\varepsilon$  cover of the set  $\{\min_a \|\Phi D a - \phi(x)\|_{\mathcal{H}} : D \in \mathcal{D}\}$ , it is enough to obtain an  $\varepsilon/(\lambda\gamma)$  cover of  $\{\Phi D : D \in \mathcal{D}\}$ . If also  $\phi$  is uniformly  $L$  Hölder of order  $\alpha$  over  $\mathcal{R}$  then an  $(\lambda\gamma L/\varepsilon)^{-1/\alpha}$  cover of the set of dictionaries is sufficient, which as we have seen requires at most  $(C(\lambda\gamma L/\varepsilon)^{1/\alpha})^{np}$  elements.

In the case of  $l_0$  constrained representation, the bound on  $\lambda$  due to Proposition 3 is  $\gamma k(1-\delta)$ , and the result follows from the above by substitution. ■

## 6. Conclusions

Our work has several implications on the design of dictionary learning algorithms as used in signal, image, and natural language processing. First, the fact that generalization is

only logarithmically dependent on the  $l_1$  norm of the coefficient vector widens the set of applicable approaches to penalization. Second, in the particular case of  $k$  sparse representation, we have shown that the Babel function is a key property for the generalization of dictionaries. It might thus be useful to modify dictionary learning algorithms so that they obtain dictionaries with low Babel functions, possibly through regularization or through certain convex relaxations. Third, mistake bounds (e.g., ?) on the quality of the solution to the coefficient finding optimization problem may lead to generalization bounds for practical algorithms, by tying such algorithms to  $k$  sparse representation.

The upper bounds presented here invite complementary lower bounds. The existing lower bounds for  $k = 1$  (vector quantization) and for  $k = p$  (representation using PCA directions) are applicable, but do not capture the geometry of general  $k$  sparse representation, and in particular do not clarify the effective dimension of the unrestricted class of dictionaries for it. We have not excluded the possibility that the class of unrestricted dictionaries has the same dimension as that of those with a small Babel function. The best upper bound we know for the larger class, being the trivial one of order  $O\left(\binom{p}{k}n^2/m\right)$ , leaves a significant gap for future exploration.

We view the dependence on  $\mu_{k-1}$  from an “algorithmic luckiness” perspective (?): if the data is described by a dictionary with low Babel function the generalization bounds are encouraging.

### Acknowledgments

We thank Shahar Mendelson for helpful discussions. A.B. was partly supported by the European Communitys FP7-FET program, SMALL project, under grant agreement no. 225913. S.M and D.V. were partially supported by the ISF under contract 890015.

### References

# Identifiability of Priors from Bounded Sample Sizes with Applications to Transfer Learning

**Liu Yang**

*Machine Learning Department, Carnegie Mellon University*

LIUY@CS.CMU.EDU

**Steve Hanneke**

*Department of Statistics, Carnegie Mellon University*

SHANNEKE@STAT.CMU.EDU

**Jaime Carbonell**

*Language Technologies Institute, Carnegie Mellon University*

JGC@CS.CMU.EDU

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We explore a transfer learning setting, in which a finite sequence of target concepts are sampled independently with an unknown distribution from a known family. We study the total number of labeled examples required to learn all targets to an arbitrary specified expected accuracy, focusing on the asymptotics in the number of tasks and the desired accuracy. Our primary interest is formally understanding the fundamental benefits of transfer learning, compared to learning each target independently from the others. Our approach to the transfer problem is general, in the sense that it can be used with a variety of learning protocols. The key insight driving our approach is that the distribution of the target concepts is identifiable from the joint distribution over a number of random labeled data points equal the Vapnik-Chervonenkis dimension of the concept space. This is not necessarily the case for the joint distribution over any smaller number of points. This work has particularly interesting implications when applied to active learning methods.

**Keywords:** Statistical Learning Theory, Transfer Learning, Bayesian Learning, Active Learning

## 1. Introduction

Transfer learning reuses knowledge from past related tasks to ease the process of learning to perform a new task. The goal of transfer learning is to leverage previous learning and experience to more efficiently learn novel, but related, concepts, compared to what would be possible without this prior experience. The utility of transfer learning is typically measured by a reduction in the number of training examples required to achieve a target performance on a sequence of related learning problems, compared to the number required for unrelated problems: i.e., reduced sample complexity. In many real-life scenarios, just a few training examples of a new concept or process is often sufficient for a human learner to grasp the new concept given knowledge of related ones. For example, learning to drive a van becomes much easier a task if we have already learned how to drive a car. Learning French is somewhat easier if we have already learned English (vs Chinese), and learning Spanish is easier if we know Portuguese (vs German). We are therefore interested in understanding the conditions that enable a learning machine to leverage abstract knowledge obtained as a by-product of learning past concepts, to improve its performance on future learning problems. Furthermore, we are interested in how the magnitude of these improvements grows as the learning system gains more experience from learning multiple related concepts.

The ability to transfer knowledge gained from previous tasks to make it easier to learn a new task can potentially benefit a wide range of real-world applications, including computer vision, natural language processing, cognitive science (e.g., fMRI brain state classification), and speech recognition, to name a few. As an example, consider training a speech recognizer. After training on a number of individuals, a learning system can identify common patterns of speech, such as accents or dialects, each of which requires a slightly different speech recognizer; then, given a new person to train a recognizer for, it can quickly determine the particular dialect from only a few well-chosen examples, and use the previously-learned recognizer for that particular dialect. In this case, we can think of the transferred knowledge as consisting of the common aspects of each recognizer variant and more generally the *distribution* of speech patterns existing in the population these subjects are from. This same type of distribution-related knowledge transfer can be helpful in a host of applications, including all those mentioned above.

Supposing these target concepts (e.g., speech patterns) are sampled independently from a fixed population, having knowledge of the distribution of concepts in the population may often be quite valuable. More generally, we may consider a general scenario in which the target concepts are sampled i.i.d. according to a fixed distribution. As we show below, the number of labeled examples required to learn a target concept sampled according to this distribution may be dramatically reduced if we have direct knowledge of the distribution. However, since in many real-world learning scenarios, we do not have direct access to this distribution, it is desirable to be able to somehow *learn* the distribution, based on observations from a sequence of learning problems with target concepts sampled according to that distribution. The hope is that an estimate of the distribution so-obtained might be almost as useful as direct access to the true distribution in reducing the number of labeled examples required to learn subsequent target concepts. The focus of this paper is an approach to transfer learning based on estimating the distribution of the target concepts. Whereas we acknowledge that there are other important challenges in transfer learning, such as exploring improvements obtainable from transfer under various alternative notions of task relatedness (Evgeniou and Pontil, 2004; Ben-David and Schuller, 2003), or alternative reuses of knowledge obtained from previous tasks (Thrun, 1996), we believe that learning the distribution of target concepts is a central and crucial component in many transfer learning scenarios, and can reduce the total sample complexity across tasks.

Note that it is not immediately obvious that the distribution of targets can even be learned in this context, since we do not have direct access to the target concepts sampled according to it, but rather have only indirect access via a finite number of labeled examples for each task; a significant part of the present work focuses on establishing that as long as these finite labeled samples are larger than a certain size, they hold sufficient information about the distribution over concepts for estimation to be possible. In particular, in contrast to standard results on consistent density estimation, our estimators are not directly based on the target concepts, but rather are only indirectly dependent on these via the labels of a finite number of data points from each task. One desideratum we pay particular attention to is minimizing the number of *extra* labeled examples needed for each task, beyond what is needed for learning that particular target, so that the benefits of transfer learning are obtained almost as a *by-product* of learning the targets. Our technique is general, in that it applies to any concept space with finite VC dimension; also, the process of learning the target concepts is (in some sense) decoupled from the mechanism of learning the concept distribution, so that we may apply our technique to a variety of learning protocols, including passive supervised learning, active supervised learning, semi-supervised learning, and learning with certain general

data-dependent forms of interaction (Hanneke, 2009). For simplicity, we choose to formulate our transfer learning algorithms in the language of active learning; as we explain below, this problem can benefit significantly from transfer. Formulations for other learning protocols would follow along similar lines, with analogous theorems; only the results in Section 4.1 are specific to active learning.

Transfer learning is related at least in spirit to much earlier work on case-based and analogical learning (Carbonell, 1983, 1986; Veloso and Carbonell, 1993; Kolodner (Ed), 1993; Thrun, 1996), although that body of work predated modern machine learning, and focused on symbolic reuse of past problem solving solutions rather than on current machine learning problems such as classification, regression or structured learning. More recently, transfer learning (and the closely related problem of *multitask* learning) has been studied in specific cases with interesting (though sometimes heuristic) approaches (Caruana, 1997; Silver, 2000; Micchelli and Pontil, 2004; Baxter, 1997; Ben-David and Schuller, 2003). This paper considers a general theoretical framework for transfer learning, based on an Empirical Bayes perspective, and derives rigorous theoretical results on the benefits of transfer. We discuss the relation of this analysis to existing theoretical work on transfer learning below.

### 1.1. Outline of the paper

The remainder of the paper is organized as follows. In Section 2 we introduce basic notation used throughout, and survey some related work from the existing literature. In Section 3, we describe and analyze our proposed method for estimating the distribution of target concepts, the key ingredient in our approach to transfer learning, which we then present in Section 4. Finally, in Section 4.1, we describe the particularly strong implications of these results for active learning.

## 2. Definitions and Related Work

First, we state a few basic notational conventions. We denote  $\mathbb{N} = \{1, 2, \dots\}$  and  $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ . For any random variable  $X$ , we generally denote by  $\mathbb{P}_X$  the distribution of  $X$  (the induced probability measure on the range of  $X$ ), and by  $\mathbb{P}_{X|Y}$  the regular conditional distribution of  $X$  given  $Y$ . For any pair of probability measures  $\mu_1, \mu_2$  on a measurable space  $(\Omega, \mathcal{F})$ , we define

$$\|\mu_1 - \mu_2\| = \sup_{A \in \mathcal{F}} |\mu_1(A) - \mu_2(A)|.$$

Next we define the particular objects of interest to our present discussion. Let  $\Theta$  be an arbitrary set (called the *parameter space*),  $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$  be a Borel space (Schervish, 1995) (where  $\mathcal{X}$  is called the *instance space*), and  $\mathcal{D}$  be a fixed distribution on  $\mathcal{X}$  (called the *data distribution*). For instance,  $\Theta$  could be  $\mathbb{R}^n$  and  $\mathcal{X}$  could be  $\mathbb{R}^m$ , for some  $n, m \in \mathbb{N}$ , though more general scenarios are certainly possible as well, including infinite-dimensional parameter spaces. Let  $\mathbb{C}$  be a set of measurable classifiers  $h : \mathcal{X} \rightarrow \{-1, +1\}$  (called the *concept space*), and suppose  $\mathbb{C}$  has VC dimension  $d < \infty$  (Vapnik, 1982) (such a space is called a *VC class*).  $\mathbb{C}$  is equipped with its Borel  $\sigma$ -algebra  $\mathcal{B}$ , induced by the pseudo-metric  $\rho(h, g) = \mathcal{D}(\{x \in \mathcal{X} : h(x) \neq g(x)\})$ . Though all of our results can be formulated for general  $\mathcal{D}$  in slightly more complex terms, for simplicity throughout the discussion below we suppose  $\rho$  is actually a *metric*, in that any  $h, g \in \mathbb{C}$  with  $h \neq g$  have  $\rho(h, g) > 0$ ; this amounts to a topological assumption on  $\mathbb{C}$  relative to  $\mathcal{D}$ .

For each  $\theta \in \Theta$ ,  $\pi_{\theta}$  is a distribution on  $\mathbb{C}$  (called a *prior*). Our only (rather mild) assumption on this family of prior distributions is that  $\{\pi_{\theta} : \theta \in \Theta\}$  be totally bounded, in the sense that  $\forall \varepsilon > 0$ ,

$\exists$  finite  $\Theta_\varepsilon \subseteq \Theta$  s.t.  $\forall \theta \in \Theta, \exists \theta_\varepsilon \in \Theta_\varepsilon$  with  $\|\pi_\theta - \pi_{\theta_\varepsilon}\| < \varepsilon$ . See (Devroye and Lugosi, 2001) for examples of categories of classes that satisfy this.

The general setup for the learning problem is that we have a *true* parameter value  $\theta_* \in \Theta$ , and a collection of  $\mathbb{C}$ -valued random variables  $\{h_{t\theta}^*\}_{t \in \mathbb{N}, \theta \in \Theta}$ , where for a fixed  $\theta \in \Theta$  the  $\{h_{t\theta}^*\}_{t \in \mathbb{N}}$  variables are i.i.d. with distribution  $\pi_\theta$ .

The learning problem is the following. For each  $\theta \in \Theta$ , there is a sequence

$$\mathcal{Z}_t(\theta) = \{(X_{t1}, Y_{t1}(\theta)), (X_{t2}, Y_{t2}(\theta)), \dots\},$$

where  $\{X_{ti}\}_{t,i \in \mathbb{N}}$  are i.i.d.  $\mathcal{D}$ , and for each  $t, i \in \mathbb{N}$ ,  $Y_{ti}(\theta) = h_{t\theta}^*(X_{ti})$ . For  $k \in \mathbb{N}$  we denote by  $\mathcal{Z}_{tk}(\theta) = \{(X_{t1}, Y_{t1}(\theta)), \dots, (X_{tk}, Y_{tk}(\theta))\}$ .

The algorithm receives values  $\varepsilon$  and  $T$  as input, and for each  $t \in \{1, 2, \dots, T\}$  in increasing order, it observes the sequence  $X_{t1}, X_{t2}, \dots$ , and may then select an index  $i_1$ , receive label  $Y_{ti_1}(\theta_*)$ , select another index  $i_2$ , receive label  $Y_{ti_2}(\theta_*)$ , etc. The algorithm proceeds in this fashion, sequentially requesting labels, until eventually it produces a classifier  $\hat{h}_t$ . It then increments  $t$  and repeats this process until it produces a sequence  $\hat{h}_1, \hat{h}_2, \dots, \hat{h}_T$ , at which time it halts. To be called *correct*, the algorithm must have a guarantee that  $\forall \theta_* \in \Theta, \forall t \leq T, \mathbb{E} [\rho(\hat{h}_t, h_{t\theta_*}^*)] \leq \varepsilon$ . We will be interested in the expected number of label requests necessary for a correct learning algorithm, averaged over the  $T$  tasks, and in particular in how shared information between tasks can help to reduce this quantity when direct access to  $\theta_*$  is not available to the algorithm.

## 2.1. Relation to Existing Theoretical Work on Transfer Learning

Although we know of no existing work on the theoretical advantages of transfer learning for active learning, the existing literature contains several analyses of the advantages of transfer learning for passive learning. In his classic work, Baxter (1997) explores a similar setup for a general form of passive learning, except in a *full* Bayesian setting (in contrast to our setting, often referred to as “empirical Bayes,” which includes a constant parameter  $\theta_*$  to be estimated from data). Essentially, Baxter (1997) sets up a hierarchical Bayesian model, in which (in our notation)  $\theta_*$  is a random variable with known distribution (hyper-prior), but otherwise the specialization of Baxter’s setting to the pattern recognition problem is essentially identical to our setup above. This hyper-prior does make the problem slightly easier, but generally the results of Baxter (1997) are of a different nature than our objectives here. Specifically, Baxter’s results on learning from labeled examples can be interpreted as indicating that transfer learning can improve certain *constant factors* in the asymptotic rate of convergence of the average of expected error rates across the learning problems. That is, certain constant complexity terms (for instance, related to the concept space) can be reduced to (potentially much smaller) values related to  $\pi_{\theta_*}$  by transfer learning. Baxter argues that, as the number of tasks grows large, this effectively achieves close to the known results on the sample complexity of passive learning with direct access to  $\theta_*$ . A similar claim is discussed by Ando and Zhang (2004) (though in less detail and formality) for a setting closer to that studied here, where  $\theta_*$  is an unknown parameter to be estimated.

There are also several results on transfer learning of a slightly different variety, in which, rather than having a prior distribution for the target concept, the learner initially has several potential concept spaces to choose from, and the role of transfer is to help the learner select from among these concept spaces (Baxter, 2000; Ando and Zhang, 2004). In this case, the idea is that one of these concept spaces has the best average minimum achievable error rate per learning problem,

and the objective of transfer learning is to perform nearly as well as if we knew which of the spaces has this property. In particular, if we assume the target functions for each task all reside in one of the concept spaces, then the objective of transfer learning is to perform nearly as well as if we knew which of the spaces contains the targets. Thus, transfer learning results in a sample complexity related to the number of learning problems, a complexity term for this best concept space, and a complexity term related to the diversity of concept spaces we have to choose from. In particular, as with Baxter (1997), these results can typically be interpreted as giving constant factor improvements from transfer in a passive learning context, at best reducing the complexity constants, from those for the union over the given concept spaces, down to the complexity constants of the single best concept space.

In addition to the above works, there are several analyses of transfer learning and multitask learning of an entirely different nature than our present discussion, in that the objectives of the analysis are somewhat different. Specifically, there is a branch of the literature concerned with task *relatedness*, not in terms of the underlying process that generates the target concepts, but rather directly in terms of relations between the target concepts themselves. In this sense, several tasks with related target concepts should be much easier to learn than tasks with unrelated target concepts. This is studied in the context of kernel methods by Micchelli and Pontil (2004); Evgeniou and Pontil (2004); Evgeniou, Micchelli, and Pontil (2005), and in a more general theoretical framework by Ben-David and Schuller (2003). As mentioned, our approach to transfer learning is based on the idea of estimating the distribution of target concepts. As such, though interesting and important, these notions of direct relatedness of target concepts are not as relevant to our present discussion.

As with Baxter (1997), the present work is interested in showing that as the number of tasks grows large, we can effectively achieve a sample complexity close to that achievable with direct access to  $\theta_*$ . However, in contrast, we are interested in a general approach to transfer learning and the analysis thereof, leading to concrete results for a variety of learning protocols such as active learning and semi-supervised learning. In particular, as we explain below, combining the results of this work with a result of Yang, Hanneke, and Carbonell (2010) reveals the interesting phenomenon that, in the context of active learning, transfer learning can sometimes improve the asymptotic dependence on  $\varepsilon$ , rather than merely the constant factors as in the analysis of Baxter (1997).

Additionally, unlike Baxter (1997), we study the benefits of transfer learning in terms of the asymptotics as the number of learning problems grows large, *without* necessarily requiring the number of labeled examples per learning problem to also grow large. That is, our analysis reveals benefits from transfer learning even if the number of labeled examples per learning problem is *bounded*. This is desirable for the following practical reasons. In many settings where transfer learning may be useful, it is desirable that the number of labeled examples we need to collect from each particular learning problem never be significantly larger than the number of such examples required to solve that particular problem (i.e., to learn that target concept to the desired accuracy). For instance, this is the case when the learning problems are not all solved by the same individual (or company, etc.), but rather a coalition of cooperating individuals (e.g., hospitals sharing data on clinical trials); each individual may be willing to share the data they used to learn their problem, in the interest of making others' learning problems easier; however, they may not be willing to collect significantly *more* data to advance this cause than they themselves need for their own learning problem. Given a desired error rate  $\varepsilon$  for each learning problem, the number of labeled examples required to learn each particular target concept to this desired error rate is always bounded by an  $\varepsilon$ -dependent value. Therefore, an analysis that requires a growing number of examples per learning

problem seems undesirable in these scenarios, since for some of the problems we would need to label a number of examples far beyond what is needed to learn a good classifier for that particular problem. We should therefore be particularly interested in studying transfer as a *by-product* of the usual learning process; failing this, we are interested in the minimum possible number of *extra* labeled examples per task to gain the benefits of transfer learning. To our knowledge, no result of this type (bounded sample size per learning problem) has yet been established at the level of generality studied here.

### 3. Estimating the Prior

The advantage of transfer learning in this setting is that each learning problem provides some information about  $\theta_*$ , so that after solving several of the learning problems, we might hope to be able to *estimate*  $\theta_*$ . Then with this estimate in hand, we can use the corresponding estimated prior distribution in the learning algorithm for subsequent learning problems, to help inform the learning process similarly to how direct knowledge of  $\theta_*$  might be helpful. However, the difficulty in approaching this is how to define such an estimator. Since we do not have direct access to the  $h_t^*$  values, but rather only indirect observations via a finite number of example labels, the standard results for density estimation from i.i.d. samples cannot be applied.

The idea we pursue below is to consider the distributions on  $\mathcal{Z}_{tk}(\theta_*)$ . These variables are directly observable, by requesting the labels of those examples. Thus, for any finite  $k \in \mathbb{N}$ , this distribution is estimable from observable data. That is, using the i.i.d. values  $\mathcal{Z}_{1k}(\theta_*), \dots, \mathcal{Z}_{tk}(\theta_*)$ , we can apply standard techniques for density estimation to arrive at an estimator of  $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_*)}$ . Then the question is whether the distribution  $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_*)}$  uniquely characterizes the prior distribution  $\pi_{\theta_*}$ : that is, whether  $\pi_{\theta_*}$  is *identifiable* from  $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_*)}$ .

As an example, consider the space of *half-open interval* classifiers on  $[0, 1]$ :  $\mathcal{C} = \{\mathbb{1}_{[a,b)}^\pm : 0 \leq a \leq b \leq 1\}$ , where  $\mathbb{1}_{[a,b)}^\pm(x) = +1$  if  $a \leq x < b$  and  $-1$  otherwise. In this case,  $\pi_{\theta_*}$  is *not* necessarily identifiable from  $\mathbb{P}_{\mathcal{Z}_{t1}(\theta_*)}$ ; for instance, the distributions  $\pi_{\theta_1}$  and  $\pi_{\theta_2}$  characterized by  $\pi_{\theta_1}(\{\mathbb{1}_{[0,1]}^\pm\}) = \pi_{\theta_1}(\{\mathbb{1}_\emptyset^\pm\}) = 1/2$  and  $\pi_{\theta_2}(\{\mathbb{1}_{[0,1/2)}^\pm\}) = \pi_{\theta_2}(\{\mathbb{1}_{[1/2,1)}^\pm\}) = 1/2$  are not distinguished by these one-dimensional distributions. However, it turns out that for this half-open intervals problem,  $\pi_{\theta_*}$  is uniquely identifiable from  $\mathbb{P}_{\mathcal{Z}_{t2}(\theta_*)}$ ; for instance, in the  $\theta_1$  vs  $\theta_2$  scenario, the conditional probability  $\mathbb{P}_{(Y_{t1}(\theta_i), Y_{t2}(\theta_i)) | (X_{t1}, X_{t2})}((+1, +1) | (1/4, 3/4))$  will distinguish  $\pi_{\theta_1}$  from  $\pi_{\theta_2}$ , and this can be calculated from  $\mathbb{P}_{\mathcal{Z}_{t2}(\theta_i)}$ . The crucial element of the analysis below is determining the appropriate value of  $k$  to uniquely identify  $\pi_{\theta_*}$  from  $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_*)}$  *in general*. As we will see,  $k = d$  is *always* sufficient, a key insight for the results that follow.

To be specific, in order to transfer knowledge from one task to the next, we use a few labeled data points from each task to gain information about  $\theta_*$ . For this, for each task  $t$ , we simply take the first  $d$  data points in the  $\mathcal{Z}_t(\theta_*)$  sequence. That is, we request the labels

$$Y_{t1}(\theta_*), Y_{t2}(\theta_*), \dots, Y_{td}(\theta_*)$$

and use the points  $\mathcal{Z}_{td}(\theta_*)$  to update an estimate of  $\theta_*$ .

The following result shows that this technique does provide a consistent estimator of  $\pi_{\theta_*}$ . Again, note that this result is not a straightforward application of the standard approach to consistent estimation, since the observations here are not the  $h_{t\theta_*}^*$  variables themselves, but rather a number of the  $Y_{ti}(\theta_*)$  values. The key insight in this result is that  $\pi_{\theta_*}$  is *uniquely identified* by the joint distribution

$\mathbb{P}_{\mathcal{Z}_{td}(\theta_*)}$  over the first  $d$  labeled examples; later, we prove this is *not* necessarily true for  $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_*)}$  for values  $k < d$ .

**Theorem 1** *There exists an estimator  $\hat{\theta}_{T\theta_*} = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta_*), \dots, \mathcal{Z}_{Td}(\theta_*))$ , and functions  $R : \mathbb{N}_0 \times (0, 1] \rightarrow [0, \infty)$  and  $\delta : \mathbb{N}_0 \times (0, 1] \rightarrow [0, 1]$ , such that for any  $\alpha > 0$ ,  $\lim_{T \rightarrow \infty} R(T, \alpha) = \lim_{T \rightarrow \infty} \delta(T, \alpha) = 0$  and for any  $T \in \mathbb{N}_0$  and  $\theta_* \in \Theta$ ,*

$$\mathbb{P} \left( \|\pi_{\hat{\theta}_{T\theta_*}} - \pi_{\theta_*}\| > R(T, \alpha) \right) \leq \delta(T, \alpha) \leq \alpha.$$

One important detail to note, for our purposes, is that  $R(T, \alpha)$  is independent from  $\theta_*$ , so that the value of  $R(T, \alpha)$  can be calculated and used within a learning algorithm. The proof of Theorem 1 will be established via the following sequence of lemmas. Lemma 2 relates distances in the space of priors to distances in the space of distributions on the full data sets. In turn, Lemma 3 relates these distances to distances in the space of distributions on a finite number of examples from the data sets. Lemma 4 then relates the distances between distributions on any finite number of examples to distances between distributions on  $d$  examples. Finally, Lemma 5 presents a standard result on the existence of a converging estimator, in this case for the distribution on  $d$  examples, for totally bounded families of distributions. Tracing these relations back, they relate convergence of the estimator for the distribution of  $d$  examples to convergence of the corresponding estimator for the prior itself.

**Lemma 2** *For any  $\theta, \theta' \in \Theta$  and  $t \in \mathbb{N}$ ,*

$$\|\pi_\theta - \pi_{\theta'}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\|.$$

**Proof** Fix  $\theta, \theta' \in \Theta$ ,  $t \in \mathbb{N}$ . Let  $\mathbb{X} = \{X_{t1}, X_{t2}, \dots\}$ ,  $\mathbb{Y}(\theta) = \{Y_{t1}(\theta), Y_{t2}(\theta), \dots\}$ , and for  $k \in \mathbb{N}$  let  $\mathbb{X}_k = \{X_{t1}, \dots, X_{tk}\}$  and  $\mathbb{Y}_k(\theta) = \{Y_{t1}(\theta), \dots, Y_{tk}(\theta)\}$ . For  $h \in \mathbb{C}$ , let  $c_{\mathbb{X}}(h) = \{(X_{t1}, h(X_{t1})), (X_{t2}, h(X_{t2})), \dots\}$ .

For  $h, g \in \mathbb{C}$ , define  $\rho_{\mathbb{X}}(h, g) = \lim_{m \rightarrow \infty} \frac{1}{m} \sum_{i=1}^m \mathbb{1}[h(X_{ti}) \neq g(X_{ti})]$  (if the limit exists), and  $\rho_{\mathbb{X}_k}(h, g) = \frac{1}{k} \sum_{i=1}^k \mathbb{1}[h(X_{ti}) \neq g(X_{ti})]$ . Note that since  $\mathbb{C}$  has finite VC dimension, so does the collection of sets  $\{\{x : h(x) \neq g(x)\} : h, g \in \mathbb{C}\}$ , so that the uniform strong law of large numbers implies that with probability one,  $\forall h, g \in \mathbb{C}$ ,  $\rho_{\mathbb{X}}(h, g)$  exists and has  $\rho_{\mathbb{X}}(h, g) = \rho(h, g)$  (Vapnik, 1982).

Consider any  $\theta, \theta' \in \Theta$ , and any  $A \in \mathcal{B}$ . Then since  $\mathcal{B}$  is the Borel  $\sigma$ -algebra induced by  $\rho$ , any  $h \notin A$  has  $\forall g \in A$ ,  $\rho(h, g) > 0$ . Thus, if  $\rho_{\mathbb{X}}(h, g) = \rho(h, g)$  for all  $h, g \in \mathbb{C}$ , then  $\forall h \notin A$ ,

$$\forall g \in A, \rho_{\mathbb{X}}(h, g) = \rho(h, g) > 0 \implies \forall g \in A, c_{\mathbb{X}}(h) \neq c_{\mathbb{X}}(g) \implies c_{\mathbb{X}}(h) \notin c_{\mathbb{X}}(A).$$

This implies  $c_{\mathbb{X}}^{-1}(c_{\mathbb{X}}(A)) = A$ . Under these conditions,

$$\mathbb{P}_{\mathcal{Z}_t(\theta)|\mathbb{X}}(c_{\mathbb{X}}(A)) = \pi_\theta(c_{\mathbb{X}}^{-1}(c_{\mathbb{X}}(A))) = \pi_\theta(A),$$

and similarly for  $\theta'$ .

Any measurable set  $C$  for the range of  $\mathcal{Z}_t(\theta)$  can be expressed as  $C = \{c_{\bar{x}}(h) : (h, \bar{x}) \in C'\}$  for some appropriate  $C' \in \mathcal{B} \otimes \mathcal{B}_{\mathcal{X}}^\infty$ . Letting  $C'_{\bar{x}} = \{h : (h, \bar{x}) \in C'\}$ , we have

$$\mathbb{P}_{\mathcal{Z}_t(\theta)}(C) = \int \pi_\theta(c_{\bar{x}}^{-1}(c_{\bar{x}}(C'_{\bar{x}}))) \mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}) = \int \pi_\theta(C'_{\bar{x}}) \mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}). = \mathbb{P}_{(h_{t\theta}^*, \mathbb{X})}(C').$$

Likewise, this reasoning holds for  $\theta'$ . Then

$$\begin{aligned} \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| &= \|\mathbb{P}_{(h_{t\theta}^*, \mathbb{X})} - \mathbb{P}_{(h_{t\theta'}^*, \mathbb{X})}\| \\ &= \sup_{C' \in \mathcal{B} \otimes \mathcal{B}_{\mathcal{X}}^\infty} \left| \int (\pi_\theta(C'_x) - \pi_{\theta'}(C'_x)) \mathbb{P}_{\mathbb{X}}(dx) \right| \\ &\leq \int \sup_{A \in \mathcal{B}} |\pi_\theta(A) - \pi_{\theta'}(A)| \mathbb{P}_{\mathbb{X}}(dx) = \|\pi_\theta - \pi_{\theta'}\|. \end{aligned}$$

Since we also have

$$\begin{aligned} \|\pi_\theta - \pi_{\theta'}\| &= \|\mathbb{P}_{(h_{t\theta}^*, \mathbb{X})}(\cdot \times \mathcal{X}^\infty) - \mathbb{P}_{(h_{t\theta'}^*, \mathbb{X})}(\cdot \times \mathcal{X}^\infty)\| \\ &\leq \|\mathbb{P}_{(h_{t\theta}^*, \mathbb{X})} - \mathbb{P}_{(h_{t\theta'}^*, \mathbb{X})}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\|, \end{aligned}$$

this means  $\|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| = \|\pi_\theta - \pi_{\theta'}\|$ . ■

**Lemma 3** *There exists a sequence  $r_k = o(1)$  such that  $\forall t, k \in \mathbb{N}, \forall \theta, \theta' \in \Theta$ ,*

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq \|\pi_\theta - \pi_{\theta'}\| \leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + r_k.$$

**Proof** The left inequality follows from Lemma 2 and the basic definition of  $\|\cdot\|$ , since  $\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(\cdot) = \mathbb{P}_{\mathcal{Z}_t(\theta)}(\cdot \times (\mathcal{X} \times \{-1, +1\})^\infty)$ , so that

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| = \|\pi_\theta - \pi_{\theta'}\|.$$

The remainder of this proof focuses on the right inequality. Fix  $\theta, \theta' \in \Theta$ , let  $\gamma > 0$ , and let  $B \subseteq (\mathcal{X} \times \{-1, +1\})^\infty$  be a measurable set such that

$$\|\pi_\theta - \pi_{\theta'}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| < \mathbb{P}_{\mathcal{Z}_t(\theta)}(B) - \mathbb{P}_{\mathcal{Z}_t(\theta')}(B) + \gamma.$$

Let  $\mathcal{A}$  be the collection of all measurable subsets of  $(\mathcal{X} \times \{-1, +1\})^\infty$  representable in the form  $A' \times (\mathcal{X} \times \{-1, +1\})^\infty$ , for some measurable  $A' \subseteq (\mathcal{X} \times \{-1, +1\})^k$  and some  $k \in \mathbb{N}$ . In particular, since  $\mathcal{A}$  is an algebra that generates the product  $\sigma$ -algebra, Carathéodory's extension theorem (Schervish, 1995) implies that there exist disjoint sets  $\{A_i\}_{i \in \mathbb{N}}$  in  $\mathcal{A}$  such that  $B \subseteq \bigcup_{i \in \mathbb{N}} A_i$  and

$$\mathbb{P}_{\mathcal{Z}_t(\theta)}(B) - \mathbb{P}_{\mathcal{Z}_t(\theta')}(B) < \sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i) + \gamma.$$

Additionally, as these sums are bounded, there must exist  $n \in \mathbb{N}$  such that

$$\sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) < \gamma + \sum_{i=1}^n \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i),$$

so that

$$\begin{aligned} \sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i) &< \gamma + \sum_{i=1}^n \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i=1}^n \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i) \\ &= \gamma + \mathbb{P}_{\mathcal{Z}_t(\theta)} \left( \bigcup_{i=1}^n A_i \right) - \mathbb{P}_{\mathcal{Z}_t(\theta')} \left( \bigcup_{i=1}^n A_i \right). \end{aligned}$$

As  $\bigcup_{i=1}^n A_i \in \mathcal{A}$ , there exists  $k' \in \mathbb{N}$  and measurable  $A' \subseteq (\mathcal{X} \times \{-1, +1\})^{k'}$  such that  $\bigcup_{i=1}^n A_i = A' \times (\mathcal{X} \times \{-1, +1\})^\infty$ , and therefore

$$\begin{aligned} \mathbb{P}_{\mathcal{Z}_t(\theta)} \left( \bigcup_{i=1}^n A_i \right) - \mathbb{P}_{\mathcal{Z}_t(\theta')} \left( \bigcup_{i=1}^n A_i \right) &= \mathbb{P}_{\mathcal{Z}_{tk'}(\theta)}(A') - \mathbb{P}_{\mathcal{Z}_{tk'}(\theta')}(A') \\ &\leq \|\mathbb{P}_{\mathcal{Z}_{tk'}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk'}(\theta')}\| \leq \lim_{k \rightarrow \infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|. \end{aligned}$$

In summary, we have  $\|\pi_\theta - \pi_{\theta'}\| \leq \lim_{k \rightarrow \infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + 3\gamma$ . Since this is true for an arbitrary  $\gamma > 0$ , taking the limit as  $\gamma \rightarrow 0$  implies

$$\|\pi_\theta - \pi_{\theta'}\| \leq \lim_{k \rightarrow \infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|.$$

In particular, this implies there exists a sequence  $r_k(\theta, \theta') = o(1)$  such that

$$\forall k \in \mathbb{N}, \|\pi_\theta - \pi_{\theta'}\| \leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + r_k(\theta, \theta').$$

This would suffice to establish the upper bound if we were allowing  $r_k$  to depend on the particular  $\theta$  and  $\theta'$ . However, to guarantee the same rates of convergence for all pairs of parameters requires an additional argument. Specifically, let  $\gamma > 0$  and let  $\Theta_\gamma$  denote a minimal subset of  $\Theta$  such that,  $\forall \theta \in \Theta, \exists \theta_\gamma \in \Theta_\gamma$  s.t.  $\|\pi_\theta - \pi_{\theta_\gamma}\| < \gamma$ : that is, a minimal  $\gamma$ -cover. Since  $|\Theta_\gamma| < \infty$  by assumption, defining  $r_k(\gamma) = \max_{\theta, \theta' \in \Theta_\gamma} r_k(\theta, \theta')$ , we have  $r_k(\gamma) = o(1)$ . Furthermore, for any  $\theta, \theta' \in \Theta$ , letting  $\theta_\gamma = \operatorname{argmin}_{\theta'' \in \Theta_\gamma} \|\pi_\theta - \pi_{\theta''}\|$  and  $\theta'_\gamma = \operatorname{argmin}_{\theta'' \in \Theta_\gamma} \|\pi_{\theta'} - \pi_{\theta''}\|$ , we have (by triangle inequalities)

$$\begin{aligned} \|\pi_\theta - \pi_{\theta'}\| &\leq \|\pi_\theta - \pi_{\theta_\gamma}\| + \|\pi_{\theta_\gamma} - \pi_{\theta'_\gamma}\| + \|\pi_{\theta'_\gamma} - \pi_{\theta'}\| \\ &< 2\gamma + r_k(\gamma) + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\|. \end{aligned}$$

By triangle inequalities and the left inequality from the lemma statement (established above), we also have

$$\begin{aligned} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\| &\leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}\| + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta')\theta_\gamma}\| \\ &\leq \|\pi_{\theta_\gamma} - \pi_\theta\| + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + \|\pi_{\theta'} - \pi_{\theta'_\gamma}\| \\ &< 2\gamma + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|. \end{aligned}$$

Defining  $r_k = \inf_{\gamma > 0} (4\gamma + r_k(\gamma))$ , we have the right inequality of the lemma statement, and since  $r_k(\gamma) = o(1)$  for each  $\gamma > 0$ , we have  $r_k = o(1)$ .  $\blacksquare$

**Lemma 4**  $\forall t, k \in \mathbb{N}, \forall \theta, \theta' \in \Theta$ ,

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq 4 \cdot 2^{2k+d} k^d \sqrt{\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|}.$$

**Proof** Fix any  $t \in \mathbb{N}$ , and let  $\mathbb{X} = \{X_{t1}, X_{t2}, \dots\}$  and  $\mathbb{Y}(\theta) = \{Y_{t1}(\theta), Y_{t2}(\theta), \dots\}$ , and for  $k \in \mathbb{N}$  let  $\mathbb{X}_k = \{X_{t1}, \dots, X_{tk}\}$  and  $\mathbb{Y}_k(\theta) = \{Y_{t1}(\theta), \dots, Y_{tk}(\theta)\}$ .

If  $k \leq d$ , then  $\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(\cdot) = \mathbb{P}_{\mathcal{Z}_{td}(\theta)}(\cdot \times (\mathcal{X} \times \{-1, +1\})^{d-k})$ , so that

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq \|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|,$$

and therefore the result trivially holds.

Now suppose  $k > d$ . For a sequence  $\bar{z}$  and  $I \subseteq \mathbb{N}$ , we will use the notation  $\bar{z}_I = \{\bar{z}_i : i \in I\}$ . Note that, for any  $k > d$  and  $\bar{x}^k \in \mathcal{X}^k$ , there is a sequence  $\bar{y}(\bar{x}^k) \in \{-1, +1\}^k$  such that no  $h \in \mathbb{C}$  has  $h(\bar{x}^k) = \bar{y}(\bar{x}^k)$  (i.e.,  $\forall h \in \mathbb{C}, \exists i \leq k$  s.t.  $h(\bar{x}_i^k) \neq \bar{y}_i(\bar{x}^k)$ ). Now suppose  $k > d$  and take as an inductive hypothesis that there is a measurable set  $A^* \subseteq \mathcal{X}^\infty$  of probability one with the property that  $\forall \bar{x} \in A^*$ , for every finite  $I \subset \mathbb{N}$  with  $|I| > d$ , for every  $\bar{y} \in \{-1, +1\}^\infty$  with  $\|\bar{y}_I - \bar{y}(\bar{x}_I)\|_1/2 \leq k-1$ ,

$$\begin{aligned} & |\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I)| \\ & \leq 2^{k-1} \cdot \max_{\tilde{y}^d \in \{-1, +1\}^d, D \in I^d} \left| \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\bar{x}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\bar{x}_D) \right|. \end{aligned}$$

This clearly holds for  $\|\bar{y}_I - \bar{y}(\bar{x}_I)\|_1/2 = 0$ , since  $\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) = 0$  in this case, so this will serve as our base case in the inductive proof. Next we inductively extend this to the value  $k > 0$ . Specifically, let  $A_{k-1}^*$  be the  $A^*$  guaranteed to exist by the inductive hypothesis, and fix any  $\bar{x} \in A^*$ ,  $\bar{y} \in \{-1, +1\}^\infty$ , and finite  $I \subset \mathbb{N}$  with  $|I| > d$  and  $\|\bar{y}_I - \bar{y}(\bar{x}_I)\|_1/2 = k$ . Let  $i \in I$  be such that  $\bar{y}_i \neq \bar{y}_i(\bar{x}_I)$ , and let  $\bar{y}' \in \{-1, +1\}$  have  $\bar{y}'_j = \bar{y}_j$  for every  $j \neq i$ , and  $\bar{y}'_i = -\bar{y}_i$ . Then

$$\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) = \mathbb{P}_{\mathbb{Y}_{I \setminus \{i\}}(\theta)|\mathbb{X}_{I \setminus \{i\}}}(\bar{y}_{I \setminus \{i\}}|\bar{x}_{I \setminus \{i\}}) - \mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I),$$

and similarly for  $\theta'$ . By the inductive hypothesis, this means

$$\begin{aligned} & |\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I)| \\ & \leq \left| \mathbb{P}_{\mathbb{Y}_{I \setminus \{i\}}(\theta)|\mathbb{X}_{I \setminus \{i\}}}(\bar{y}_{I \setminus \{i\}}|\bar{x}_{I \setminus \{i\}}) - \mathbb{P}_{\mathbb{Y}_{I \setminus \{i\}}(\theta')|\mathbb{X}_{I \setminus \{i\}}}(\bar{y}_{I \setminus \{i\}}|\bar{x}_{I \setminus \{i\}}) \right| \\ & \quad + \left| \mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I) \right| \\ & \leq 2^k \cdot \max_{\tilde{y}^d \in \{-1, +1\}^d, D \in I^d} \left| \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\bar{x}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\bar{x}_D) \right|. \end{aligned}$$

Therefore, by the principle of induction, this inequality holds for all  $k > d$ , for every  $\bar{x} \in A^*$ ,  $\bar{y} \in \{-1, +1\}^\infty$ , and finite  $I \subset \mathbb{N}$ , where  $A^*$  has  $\mathcal{D}^\infty$ -probability one.

In particular, we have that for  $\theta, \theta' \in \Theta$ ,

$$\begin{aligned} & \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \\ & \leq 2^k \mathbb{E} \left[ \max_{\bar{y}^k \in \{-1, +1\}^k} \left| \mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k}(\bar{y}^k|\mathbb{X}_k) - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}(\bar{y}^k|\mathbb{X}_k) \right| \right] \\ & \leq 2^{2k} \mathbb{E} \left[ \max_{\tilde{y}^d \in \{-1, +1\}^d, D \in \{1, \dots, k\}^d} \left| \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_D) \right| \right] \\ & \leq 2^{2k} \sum_{\tilde{y}^d \in \{-1, +1\}^d} \sum_{D \in \{1, \dots, k\}^d} \mathbb{E} \left[ \left| \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_D) \right| \right]. \end{aligned}$$

Exchangeability implies this is at most

$$\begin{aligned} & 2^{2k} \sum_{\tilde{y}^d \in \{-1,+1\}^d} \sum_{D \in \{1,\dots,k\}^d} \mathbb{E} \left[ \left| \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) \right| \right] \\ & \leq 2^{2k+d} k^d \max_{\tilde{y}^d \in \{-1,+1\}^d} \mathbb{E} \left[ \left| \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) \right| \right]. \end{aligned}$$

To complete the proof, we need only bound this value by an appropriate function of  $\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|$ . Toward this end, suppose

$$\mathbb{E} \left[ \left| \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) \right| \right] \geq \varepsilon,$$

for some  $\tilde{y}^d$ . Then either

$$\mathbb{P} \left( \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) \geq \varepsilon/4 \right) \geq \varepsilon/4,$$

or

$$\mathbb{P} \left( \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) \geq \varepsilon/4 \right) \geq \varepsilon/4.$$

For which ever is the case, let  $A_\varepsilon$  denote the corresponding measurable subset of  $\mathcal{X}^d$ , of probability at least  $\varepsilon/4$ . Then

$$\begin{aligned} \|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\| & \geq \left| \mathbb{P}_{\mathcal{Z}_{td}(\theta)}(A_\varepsilon \times \{\tilde{y}^d\}) - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}(A_\varepsilon \times \{\tilde{y}^d\}) \right| \\ & \geq (\varepsilon/4)\mathbb{P}_{\mathbb{X}_d}(A_\varepsilon) \geq \varepsilon^2/16. \end{aligned}$$

Therefore,

$$\mathbb{E} \left[ \left| \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) \right| \right] \leq 4\sqrt{\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|},$$

which means

$$\begin{aligned} & 2^{2k+d} k^d \max_{\tilde{y}^d \in \{-1,+1\}^d} \mathbb{E} \left[ \left| \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) \right| \right] \\ & \leq 4 \cdot 2^{2k+d} k^d \sqrt{\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|}. \end{aligned}$$

■

The following lemma is a standard result on the existence of converging density estimators for totally bounded families of distributions. For instance, the *skeleton* estimates described by Yatracos (1985); Devroye and Lugosi (2001) satisfy this; in fact, in many contexts (though certainly not all), even a simple maximum likelihood estimator would suffice. The reader is referred to (Yatracos, 1985; Devroye and Lugosi, 2001) for a proof of this lemma.

**Lemma 5** (Yatracos, 1985; Devroye and Lugosi, 2001) *Let  $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$  be a totally bounded family of probability measures on a measurable space  $(\Omega, \mathcal{F})$ , and let  $\{W_t(\theta)\}_{t \in \mathbb{N}, \theta \in \Theta}$  be  $\Omega$ -valued random variables such that  $\{W_t(\theta)\}_{t \in \mathbb{N}}$  are i.i.d.  $p_\theta$  for each  $\theta \in \Theta$ . Then there exists an estimator*

$\hat{\theta}_{T\theta_*} = \hat{\theta}_T(W_1(\theta_*), \dots, W_T(\theta_*))$  and functions  $R_{\mathcal{P}} : \mathbb{N}_0 \times (0, 1] \rightarrow [0, \infty)$  and  $\delta_{\mathcal{P}} : \mathbb{N}_0 \times (0, 1] \rightarrow [0, 1]$  such that  $\forall \alpha > 0$ ,  $\lim_{T \rightarrow \infty} R_{\mathcal{P}}(T, \alpha) = \lim_{T \rightarrow \infty} \delta_{\mathcal{P}}(T, \alpha) = 0$ , and  $\forall \theta_* \in \Theta$  and  $T \in \mathbb{N}_0$ ,

$$\mathbb{P} \left( \|p_{\hat{\theta}_{T\theta_*}} - p_{\theta_*}\| > R_{\mathcal{P}}(T, \alpha) \right) \leq \delta_{\mathcal{P}}(T, \alpha) \leq \alpha.$$

We are now ready for the proof of Theorem 1

**Proof** [Theorem 1] For  $\varepsilon > 0$ , let  $\Theta_\varepsilon \subseteq \Theta$  be any finite subset such that  $\forall \theta \in \Theta$ ,  $\exists \theta_\varepsilon \in \Theta_\varepsilon$  with  $\|\pi_{\theta_\varepsilon} - \pi_\theta\| < \varepsilon$ ; this exists by the assumption that  $\{\pi_\theta : \theta \in \Theta\}$  is totally bounded. Then Lemma 3 implies that  $\forall \theta \in \Theta$ ,  $\exists \theta_\varepsilon \in \Theta_\varepsilon$  with  $\|\mathbb{P}_{Z_{td}(\theta_\varepsilon)} - \mathbb{P}_{Z_{td}(\theta)}\| \leq \|\pi_{\theta_\varepsilon} - \pi_\theta\| < \varepsilon$ , so that  $\{\mathbb{P}_{Z_{td}(\theta_\varepsilon)} : \theta_\varepsilon \in \Theta_\varepsilon\}$  is a finite  $\varepsilon$ -cover of  $\{\mathbb{P}_{Z_{td}(\theta)} : \theta \in \Theta\}$ . Therefore,  $\{\mathbb{P}_{Z_{td}(\theta)} : \theta \in \Theta\}$  is totally bounded. Lemma 5 then implies that there exists an estimator  $\hat{\theta}_{T\theta_*} = \hat{\theta}_T(Z_{1d}(\theta_*), \dots, Z_{Td}(\theta_*))$  and functions  $R_d : \mathbb{N}_0 \times (0, 1] \rightarrow [0, \infty)$  and  $\delta_d : \mathbb{N}_0 \times (0, 1] \rightarrow [0, 1]$  such that  $\forall \alpha > 0$ ,  $\lim_{T \rightarrow \infty} R_d(T, \alpha) = \lim_{T \rightarrow \infty} \delta_d(T, \alpha) = 0$ , and  $\forall \theta_* \in \Theta$  and  $T \in \mathbb{N}_0$ ,

$$\mathbb{P} \left( \|\mathbb{P}_{Z_{(T+1)d}(\hat{\theta}_{T\theta_*})} - \mathbb{P}_{Z_{(T+1)d}(\theta_*)}\| > R_d(T, \alpha) \right) \leq \delta_d(T, \alpha) \leq \alpha. \quad (1)$$

Defining

$$R(T, \alpha) = \min_{k \in \mathbb{N}} \left( r_k + 4 \cdot 2^{2k+d} k^d \sqrt{R_d(T, \alpha)} \right),$$

and  $\delta(T, \alpha) = \delta_d(T, \alpha)$ , and combining (1) with Lemmas 4 and 3, we have

$$\mathbb{P} \left( \|\pi_{\hat{\theta}_{T\theta_*}} - \pi_{\theta_*}\| > R(T, \alpha) \right) \leq \delta(T, \alpha) \leq \alpha.$$

Finally, note that  $\lim_{k \rightarrow \infty} r_k = 0$  and  $\lim_{T \rightarrow \infty} R_d(T, \alpha) = 0$  imply that  $\lim_{T \rightarrow \infty} R(T, \alpha) = 0$ . ■

### 3.1. Identifiability from d Points

Inspection of the above proof reveals that the assumption that the family of priors is totally bounded is required only to establish the estimability and bounded rate guarantees. In particular, the implied identifiability condition is, in fact, *always* satisfied, as stated formally in the following corollary.

**Corollary 6** For any priors  $\pi_1, \pi_2$  on  $\mathbb{C}$ , if  $h_i^* \sim \pi_i$ ,  $X_1, \dots, X_d$  are i.i.d.  $\mathcal{D}$  independent from  $h_i^*$ , and  $Z_d(i) = \{(X_1, h_i^*(X_1)), \dots, (X_d, h_i^*(X_d))\}$  for  $i \in \{1, 2\}$ , then  $\mathbb{P}_{Z_d(1)} = \mathbb{P}_{Z_d(2)} \Rightarrow \pi_1 = \pi_2$ .

**Proof** The described scenario is a special case of our general setting, with  $\Theta = \{1, 2\}$ , in which case  $\mathbb{P}_{Z_d(i)} = \mathbb{P}_{Z_{1d}(i)}$ . Thus, if  $\mathbb{P}_{Z_d(1)} = \mathbb{P}_{Z_d(2)}$ , then Lemma 4 and Lemma 3 combine to imply that  $\|\pi_1 - \pi_2\| \leq \inf_{k \in \mathbb{N}} r_k = 0$ . ■

It is natural to wonder whether this identifiability remains true for some smaller number of points  $k < d$ , so that we might hope to create an estimator for  $\pi_{\theta_*}$  based on an estimator for  $\mathbb{P}_{Z_{tk}(\theta_*)}$ . However, one can show that  $d$  is actually the *minimum* possible value for which this remains true for all  $\mathcal{D}$  and all families of priors. Formally, we have the following result, holding for every VC class  $\mathbb{C}$ .

**Theorem 7** *There exists a data distribution  $\mathcal{D}$  and priors  $\pi_1, \pi_2$  on  $\mathbb{C}$  such that, for any positive integer  $k < d$ , if  $h_i^* \sim \pi_i$ ,  $X_1, \dots, X_k$  are i.i.d.  $\mathcal{D}$  independent from  $h_i^*$ , and  $Z_k(i) = \{(X_1, h_i^*(X_1)), \dots, (X_k, h_i^*(X_k))\}$  for  $i \in \{1, 2\}$ , then  $\mathbb{P}_{Z_k(1)} = \mathbb{P}_{Z_k(2)}$  but  $\pi_1 \neq \pi_2$ .*

**Proof** Note that it suffices to show this is the case for  $k = d - 1$ , since any smaller  $k$  is a marginal of this case. Consider a shatterable set of points  $S_d = \{x_1, x_2, \dots, x_d\} \subseteq \mathcal{X}$ , and let  $\mathcal{D}$  be uniform on  $S_d$ . Let  $\mathbb{C}[S_d]$  be any  $2^d$  classifiers in  $\mathbb{C}$  that shatter  $S_d$ . Let  $\pi_1$  be the uniform distribution on  $\mathbb{C}[S_d]$ . Now let  $S_{d-1} = \{x_1, \dots, x_{d-1}\}$  and  $\mathbb{C}[S_{d-1}] \subseteq \mathbb{C}[S_d]$  shatter  $S_{d-1}$  with the property that  $\forall h \in \mathbb{C}[S_{d-1}], h(x_d) = \prod_{j=1}^{d-1} h(x_j)$ . Let  $\pi_2$  be uniform on  $\mathbb{C}[S_{d-1}]$ . Now for any  $k < d$  and distinct indices  $t_1, \dots, t_k \in \{1, \dots, d\}$ ,  $\{h_i^*(x_{t_1}), \dots, h_i^*(x_{t_k})\}$  is distributed uniformly in  $\{-1, +1\}^k$  for both  $i \in \{1, 2\}$ . This implies  $\mathbb{P}_{Z_{d-1}(1)|X_1, \dots, X_{d-1}} = \mathbb{P}_{Z_{d-1}(2)|X_1, \dots, X_{d-1}}$ , which implies  $\mathbb{P}_{Z_{d-1}(1)} = \mathbb{P}_{Z_{d-1}(2)}$ . However,  $\pi_1$  is clearly different from  $\pi_2$ , since even the sizes of the supports are different.  $\blacksquare$

## 4. Transfer Learning

In this section, we look at an application of the techniques from the previous section to transfer learning. Like the previous section, the results in this section are general, in that they are applicable to a variety of learning protocols, including passive supervised learning, passive semi-supervised learning, active learning, and learning with certain general types of data-dependent interaction (Hanneke, 2009). For simplicity, we restrict our discussion to the active learning formulation; the analogous results for these other learning protocols follow by similar reasoning.

The result of the previous section implies that an estimator for  $\theta_*$  based on  $d$ -dimensional joint distributions is consistent with a bounded rate of convergence  $R$ . Therefore, for certain prior-dependent learning algorithms, their behavior should be similar under  $\pi_{\hat{\theta}_{T\theta_*}}$  to their behavior under  $\pi_{\theta_*}$ .

To make this concrete, we formalize this in the active learning protocol as follows. A *prior-dependent* active learning algorithm  $\mathcal{A}$  takes as inputs  $\varepsilon > 0$ ,  $\mathcal{D}$ , and a distribution  $\pi$  on  $\mathbb{C}$ . It initially has access to  $X_1, X_2, \dots$  i.i.d.  $\mathcal{D}$ ; it then selects an index  $i_1$  to request the label for, receives  $Y_{i_1} = h^*(X_{i_1})$ , then selects another index  $i_2$ , etc., until it eventually terminates and returns a classifier. Denote by  $\mathcal{Z} = \{(X_1, h^*(X_1)), (X_2, h^*(X_2)), \dots\}$ . To be *correct*, the algorithm  $\mathcal{A}$  must guarantee that for  $h^* \sim \pi$ ,  $\forall \varepsilon > 0$ ,  $\mathbb{E}[\rho(\mathcal{A}(\varepsilon, \mathcal{D}, \pi), h^*)] \leq \varepsilon$ . We define the random variable  $N(\mathcal{A}, f, \varepsilon, \mathcal{D}, \pi)$  as the number of label requests  $\mathcal{A}$  makes before terminating, when given  $\varepsilon$ ,  $\mathcal{D}$ , and  $\pi$  as inputs, and when  $h^* = f$  is the value of the target function; we make the particular data sequence  $\mathcal{Z}$  the algorithm is run with implicit in this notation. We will be interested in the *expected sample complexity*  $SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi) = \mathbb{E}[N(\mathcal{A}, h^*, \varepsilon, \mathcal{D}, \pi)]$ .

We propose the following algorithm  $\mathcal{A}_\tau$  for transfer learning, defined in terms of a given correct prior-dependent active learning algorithm  $\mathcal{A}_a$ . We discuss interesting specifications for  $\mathcal{A}_a$  in the next section, but for now the only assumption we require is that for any  $\varepsilon > 0$  and  $\mathcal{D}$ , there is a value  $s_\varepsilon < \infty$  such that for every  $\pi$  and  $f \in \mathbb{C}$ ,  $N(\mathcal{A}_a, f, \varepsilon, \mathcal{D}, \pi) \leq s_\varepsilon$ ; this is a very mild requirement, and any active learning algorithm can be converted into one that satisfies this without significantly increasing its sample complexities for the priors it is already good for (Balcan, Hanneke, and Vaughan, 2010). We denote by  $m_\varepsilon = \frac{16d}{\varepsilon} \ln\left(\frac{24}{\varepsilon}\right)$ , and  $B(\theta, \gamma) = \{\theta' \in \Theta : \|\pi_\theta - \pi_{\theta'}\| \leq \gamma\}$ .

---

**Algorithm 1**  $\mathcal{A}_\tau(T, \varepsilon)$ : an algorithm for transfer learning, specified in terms of a generic subroutine  $\mathcal{A}_a$ .

---

```

for  $t = 1, 2, \dots, T$  do
    Request labels  $Y_{t1}(\theta_\star), \dots, Y_{td}(\theta_\star)$ 
    if  $R(t-1, \varepsilon/2) > \varepsilon/8$  then
        Request labels  $Y_{t(d+1)}(\theta_\star), \dots, Y_{tm_\varepsilon}(\theta_\star)$ 
        Take  $\hat{h}_t$  as any  $h \in \mathbb{C}$  s.t.  $\forall i \leq m_\varepsilon$ ,  $h(X_{ti}) = Y_{ti}(\theta_\star)$ 
    else
        Let  $\check{\theta}_{t\theta_\star} \in B\left(\hat{\theta}_{(t-1)\theta_\star}, R(t-1, \varepsilon/2)\right)$  be such that
         $SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}}) \leq \min_{\theta \in B(\hat{\theta}_{(t-1)\theta_\star}, R(t-1, \varepsilon/2))} SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_\theta) + 1/t$ 
        Run  $\mathcal{A}_a(\varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}})$  with data sequence  $\mathcal{Z}_t(\theta_\star)$  and let  $\hat{h}_t$  be the classifier it returns
    end if
end for

```

---

**Theorem 8** *The algorithm  $\mathcal{A}_\tau$  is correct. Furthermore, if  $S_T(\varepsilon)$  is the total number of label requests made by  $\mathcal{A}_\tau(T, \varepsilon)$ , then  $\limsup_{T \rightarrow \infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \leq SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\theta_\star}) + d$ .*

The remarkable implication of Theorem 8 is that, via transfer learning, it is possible to achieve almost the *same* long-run average sample complexity as would be achievable if the target's prior distribution were *known* to the learner. We will see in the next section that this is sometimes significantly better than the single-task sample complexity.

The algorithm  $\mathcal{A}_\tau$  is stated in a simple way here, but Theorem 8 can be improved with some obvious modifications to  $\mathcal{A}_\tau$ . The extra “ $+d$ ” in Theorem 8 is not actually necessary, since we could stop updating the estimator  $\check{\theta}_{t\theta_\star}$  (and the corresponding  $R$  value) after some  $o(T)$  number of rounds (e.g.,  $\sqrt{T}$ ), in which case we would not need to request  $Y_{t1}(\theta_\star), \dots, Y_{td}(\theta_\star)$  for  $t$  larger than this, and the extra  $d \cdot o(T)$  number of labeled examples vanishes in the average as  $T \rightarrow \infty$ . Additionally, the  $\varepsilon/4$  term can easily be improved to any value arbitrarily close to  $\varepsilon$  (even  $(1 - o(1))\varepsilon$ ) by running  $\mathcal{A}_a$  with argument  $\varepsilon - 2R(t-1, \varepsilon/2) - \delta(t-1, \varepsilon/2)$  instead of  $\varepsilon/4$ , and using this value in the  $SC$  calculations in the definition of  $\check{\theta}_{t\theta_\star}$  as well. In fact, for many algorithms  $\mathcal{A}_a$  (e.g., with  $SC(\mathcal{A}_a, \varepsilon, \mathcal{D}, \pi_{\theta_\star})$  continuous in  $\varepsilon$ ), combining the above two tricks yields  $\limsup_{T \rightarrow \infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \leq SC(\mathcal{A}_a, \varepsilon, \mathcal{D}, \pi_{\theta_\star})$ .

Returning to our motivational remarks from Subsection 2.1, we can ask how many *extra* labeled examples are required from each learning problem to gain the benefits of transfer learning. This question essentially concerns the initial step of requesting the labels  $Y_{t1}(\theta_\star), \dots, Y_{td}(\theta_\star)$ . Clearly this indicates that from each learning problem, we need at most  $d$  extra labeled examples to gain the benefits of transfer. Whether these  $d$  label requests are indeed *extra* depends on the particular learning algorithm  $\mathcal{A}_a$ ; that is, in some cases (e.g., certain passive learning algorithms),  $\mathcal{A}_a$  may itself use these initial  $d$  labels for learning, so that in these cases the benefits of transfer learning are essentially gained as a *by-product* of the learning processes, and essentially no additional labeling effort need be expended to gain these benefits. On the other hand, for some active learning algorithms, we may expect that at least some of these initial  $d$  labels would not be requested by the algorithm, so that some extra labeling effort is expended to gain the benefits of transfer in these cases.

**Proof** [Theorem 8] Recall that, to establish correctness, we must show that  $\forall t \leq T$ ,  $\mathbb{E} [\rho(\hat{h}_t, h_{t\theta_*}^*)] \leq \varepsilon$ , regardless of the value of  $\theta_* \in \Theta$ . Fix any  $\theta_* \in \Theta$  and  $t \leq T$ . If  $R(t-1, \varepsilon/2) > \varepsilon/8$ , then classic results from passive learning indicate that  $\mathbb{E} [\rho(\hat{h}_t, h_{t\theta_*}^*)] \leq \varepsilon$  (Vapnik, 1982). Otherwise, by Theorem 1, with probability at least  $1 - \varepsilon/2$ , we have  $\|\pi_{\theta_*} - \pi_{\hat{\theta}_{(t-1)\theta_*}}\| \leq R(t-1, \varepsilon/2)$ . On this event, if  $R(t-1, \varepsilon/2) \leq \varepsilon/8$ , then by a triangle inequality  $\|\pi_{\hat{\theta}_{t\theta_*}} - \pi_{\theta_*}\| \leq 2R(t-1, \varepsilon/2) \leq \varepsilon/4$ . Thus,

$$\mathbb{E} [\rho(\hat{h}_t, h_{t\theta_*}^*)] \leq \mathbb{E} [\mathbb{E} [\rho(\hat{h}_t, h_{t\theta_*}^*) | \check{\theta}_{t\theta_*}] \mathbb{1} [\|\pi_{\check{\theta}_{t\theta_*}} - \pi_{\theta_*}\| \leq \varepsilon/4]] + \varepsilon/2. \quad (2)$$

For  $\theta \in \Theta$ , let  $\hat{h}_{t\theta}$  denote the classifier that would be returned by  $\mathcal{A}_a(\varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta}})$  when run with data sequence  $\{(X_{t1}, h_{t\theta}^*(X_{t1})), (X_{t2}, h_{t\theta}^*(X_{t2})), \dots\}$ . Note that for any  $\theta \in \Theta$ , any measurable function  $F : \mathbb{C} \rightarrow [0, 1]$  has

$$\mathbb{E} [F(h_{t\theta}^*)] \leq \mathbb{E} [F(h_{t\theta})] + \|\pi_\theta - \pi_{\theta_*}\|. \quad (3)$$

In particular, supposing  $\|\pi_{\check{\theta}_{t\theta_*}} - \pi_{\theta_*}\| \leq \varepsilon/4$ , we have

$$\begin{aligned} \mathbb{E} [\rho(\hat{h}_t, h_{t\theta_*}^*) | \check{\theta}_{t\theta_*}] &= \mathbb{E} [\rho(\hat{h}_{t\theta_*}, h_{t\theta_*}^*) | \check{\theta}_{t\theta_*}] \\ &\leq \mathbb{E} [\rho(\hat{h}_{t\check{\theta}_{t\theta_*}}, h_{t\check{\theta}_{t\theta_*}}^*) | \check{\theta}_{t\theta_*}] + \|\pi_{\check{\theta}_{t\theta_*}} - \pi_{\theta_*}\| \leq \varepsilon/4 + \varepsilon/4 = \varepsilon/2. \end{aligned}$$

Combined with (2), this implies  $\mathbb{E} [\rho(\hat{h}_t, h_{t\theta_*}^*)] \leq \varepsilon$ .

We establish the sample complexity claim as follows. First note that convergence of  $R(t-1, \varepsilon/2)$  implies that  $\lim_{T \rightarrow \infty} \sum_{t=1}^T \mathbb{1}[R(t, \varepsilon/2) > \varepsilon/8]/T = 0$ , and that the number of labels used for a value of  $t$  with  $R(t-1, \varepsilon/2) > \varepsilon/8$  is bounded by a finite function  $m_\varepsilon$  of  $\varepsilon$ . Therefore,

$$\begin{aligned} &\limsup_{T \rightarrow \infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \\ &\leq d + \limsup_{T \rightarrow \infty} \sum_{t=1}^T \mathbb{E} [N(\mathcal{A}_a, h_{t\theta_*}^*, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_*}})] \mathbb{1}[R(t-1, \varepsilon/2) \leq \varepsilon/8]/T \\ &\leq d + \limsup_{T \rightarrow \infty} \sum_{t=1}^T \mathbb{E} [N(\mathcal{A}_a, h_{t\theta_*}^*, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_*}})] / T. \end{aligned} \quad (4)$$

By the definition of  $R, \delta$  from Theorem 1, we have

$$\begin{aligned} &\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E} [N(\mathcal{A}_a, h_{t\theta_*}^*, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_*}}) \mathbb{1} [\|\pi_{\hat{\theta}_{(t-1)\theta_*}} - \pi_{\theta_*}\| > R(t-1, \varepsilon/2)]] \\ &\leq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T s_{\varepsilon/4} \mathbb{P} (\|\pi_{\hat{\theta}_{(t-1)\theta_*}} - \pi_{\theta_*}\| > R(t-1, \varepsilon/2)) \\ &\leq s_{\varepsilon/4} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \delta(t-1, \varepsilon/2) = 0. \end{aligned}$$

Combined with (4), this implies

$$\begin{aligned} \limsup_{T \rightarrow \infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} &\leq d + \\ \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ N(\mathcal{A}_a, h_{t\theta_*}^*, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_*}}) \mathbb{1} \left[ \|\pi_{\hat{\theta}_{(t-1)\theta_*}} - \pi_{\theta_*}\| \leq R(t-1, \varepsilon/2) \right] \right]. \end{aligned}$$

For any  $t \leq T$ , on the event  $\|\pi_{\hat{\theta}_{(t-1)\theta_*}} - \pi_{\theta_*}\| \leq R(t-1, \varepsilon/2)$ , we have (by the property (3) and a triangle inequality)

$$\begin{aligned} \mathbb{E} \left[ N(\mathcal{A}_a, h_{t\theta_*}^*, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_*}}) \middle| \check{\theta}_{t\theta_*} \right] &\leq \mathbb{E} \left[ N(\mathcal{A}_a, h_{t\check{\theta}_{t\theta_*}}^*, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_*}}) \middle| \check{\theta}_{t\theta_*} \right] + 2R(t-1, \varepsilon/2) \\ &= SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_*}}) + 2R(t-1, \varepsilon/2) \\ &\leq SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\theta_*}) + 1/t + 2R(t-1, \varepsilon/2), \end{aligned}$$

where the last inequality follows by definition of  $\check{\theta}_{t\theta_*}$ . Therefore,

$$\begin{aligned} \limsup_{T \rightarrow \infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} &\leq d + \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\theta_*}) + 1/t + 2R(t-1, \varepsilon/2) \\ &= d + SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\theta_*}). \end{aligned}$$

■

#### 4.1. Application to Self-Verifying Active Learning

Recent work of Yang, Hanneke, and Carbonell (2010) shows that there is a correct prior-dependent active learning algorithm  $\mathcal{A}$  such that, for any prior  $\pi$  over  $\mathbb{C}$ ,  $SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi) = o(1/\varepsilon)$ . This is interesting, in that it contrasts with established results for correct prior-independent active learning algorithms, where there are known problems  $(\mathbb{C}, \mathcal{D})$  for which any prior-independent active learning algorithm  $\mathcal{A}'$  that is correct (in the sense studied above) has some prior  $\pi$  for which  $SC(\mathcal{A}', \varepsilon, \mathcal{D}, \pi) = \Omega(1/\varepsilon)$ ; for instance, the class of interval classifiers on  $[0, 1]$  under a uniform distribution  $\mathcal{D}$  satisfies this (Balcan, Hanneke, and Vaughan, 2010).

Combined with the results above for transfer learning, we get an immediate corollary that, running  $\mathcal{A}_\tau$  with the active learning algorithm  $\mathcal{A}$  having this  $o(1/\varepsilon)$  sample complexity guarantee, we have

$$\limsup_{T \rightarrow \infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} = o(1/\varepsilon).$$

Thus, in the case of active learning, there are scenarios where transfer learning (of the type studied here) can provide significant improvements in the average expected sample complexity, including improvements to the asymptotic dependence on  $\varepsilon$ .

## 5. Conclusions

We have shown that when learning a sequence of i.i.d. target concepts from a known VC class, with an unknown distribution from a known totally bounded family, transfer learning can lead to amortized expected sample complexity close to that achievable by an algorithm with direct knowledge of the targets' distribution. Furthermore, the number of extra labeled examples per task, beyond what is needed for learning that task, is bounded by the VC dimension of the class. The key insight leading to this result is that the prior distribution is uniquely identifiable based on the joint distribution over the first VC dimension number of points. This is not necessarily the case for the distribution over any number of points less than the VC dimension. As a particularly interesting application, we note that in the context of active learning, transfer learning of this type can even lead to improvements in the asymptotic dependence on the desired error rate guarantee  $\varepsilon$  in the average expected sample complexity, and in particular can guarantee this average is  $o(1/\varepsilon)$ .

## Acknowledgments

We extend our sincere thanks to Avrim Blum for several thought-provoking discussions on this topic.

## References

- R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. Technical Report RC23462, IBM T.J. Watson Research Center, 2004.
- M.-F. Balcan, S. Hanneke, and J. Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 80(2–3):111–139, September 2010.
- J. Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28:7–39, 1997.
- J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12: 149–198, 2000.
- S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *Conference on Learning Theory*, 2003.
- J. G. Carbonell. Learning by analogy: Formulating and generalizing plans from past experience. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning, An Artificial Intelligence Approach*. Tioga Press, Palo Alto, CA, 1983.
- J. G. Carbonell. Derivational analogy: A theory of reconstructive problem solving and expertise acquisition. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning, An Artificial Intelligence Approach, Volume II*. Morgan Kaufmann, 1986.
- R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.
- L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New York, NY, USA, 2001.

- T. Evgeniou and M. Pontil. Regularized multi-task learning. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004.
- T. Evgeniou, C. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.
- S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009.
- J. Kolodner (Ed). *Case-Based Learning*. Kluwer Academic Publishers, The Netherlands, 1993.
- C. Micchelli and M. Pontil. Kernels for multi-task learning. In *Advances in Neural Information Processing 18*, 2004.
- M. J. Schervish. *Theory of Statistics*. Springer, New York, NY, USA, 1995.
- D. L. Silver. *Selective Transfer of Neural Network Task Knowledge*. PhD thesis, Computer Science, University of Western Ontario, 2000.
- S. Thrun. Is learning the n-th thing any easier than learning the first? In *In Advances in Neural Information Processing Systems 8*, 1996.
- V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.
- M. M. Veloso and J. G. Carbonell. Derivational analogy in prodigy: Automating case acquisition, storage and utilization. *Machine Learning*, 10:249–278, 1993.
- L. Yang, S. Hanneke, and J. Carbonell. The sample complexity of self-verifying bayesian active learning. Technical Report CMU-ML-10-105, Carnegie Mellon University, 2010.
- Y. G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov’s entropy. *The Annals of Statistics*, 13:768–774, 1985.

## **Part II**

# **Open problems**



# Does an Efficient Calibrated Forecasting Strategy Exist?

**Jacob Abernethy**

JAKE@CS.BERKELEY.EDU UC Berkeley, Div. of Computer Science

**Shie Mannor**

SHIE@EE.TECHNION.AC.IL Technion, Department of Electrical Engineering

**Editor:** Sham Kakade, Ulrike von Luxburg

**Abstract** We recall two previously-proposed notions of *asymptotic calibration* for a forecaster making a sequence of probability predictions. We note that the existence of efficient algorithms for calibrated forecasting holds only in the case of binary outcomes. We pose the question: do there exist such efficient algorithms for the general (non-binary) case?

## Review of Calibrated Forecasting

Glenn Brier, writing in the journal *Monthly Weather Review*, observed a challenge in assessing sequential probability forecasts (Brier, 1950):

*Verification of weather forecasts has been a controversial subject for more than a half century. There are a number of reasons why this problem has been so perplexing to meteorologists and others but one of the most important difficulties seems to be in reaching an agreement on the specification of a scale of goodness for weather forecasts.*

In response, he proposed an objective scoring function which, if implemented by the forecaster, would lead to “calibrated” predictions. Yet a question which presumably did not occur to Brier is whether the latter *is even computationally feasible*. This question will be topic of the present note.

Precisely, we would like to know whether there exists an efficient algorithm for forecasting sequences of outcomes that is *asymptotically calibrated*. As the forecaster observes the sequence  $y_1, y_2, \dots$ , from some given set  $K$ , she outputs a sequence of (potentially randomized) probability predictions  $\mathbf{p}_1, \mathbf{p}_2, \dots \in \Delta(K)$ , where  $\mathbf{p}_t$  may depend only on the past  $y_1, y_2, \dots, y_{t-1}$ . Roughly speaking, a forecaster is asymptotically calibrated if her probability predictions match the outcome frequencies. Let  $\tau_T(\mathbf{p}) := \{t : \mathbf{p}_t \approx \mathbf{p}\}$ , and let  $\delta_y \in \Delta(K)$  be a point mass distribution on  $y$ , then our forecaster is calibrated if, for all  $\mathbf{p} \in \Delta(K)$ ,

$$\frac{\sum_{t \in \tau_T(\mathbf{p})} \delta_{y_t}}{|\tau_T(\mathbf{p})|} \rightarrow \mathbf{p} \quad \text{as } T \rightarrow \infty.$$

In the above definition we have been intentionally vague about the meaning of  $\mathbf{p}_t \approx \mathbf{p}$ , as well as our notion of convergence in the limit. To be more precise we shall distinguish between two notions of calibration: strong and weak.

**Definition 1** A function  $g : \Delta(K) \rightarrow \mathbb{R}$  is called *lipschitz* if there exists some  $c > 0$  for which  $|g(\mathbf{p}) - g(\mathbf{q})| \leq c\|\mathbf{p} - \mathbf{q}\|_1$  for every  $\mathbf{p}, \mathbf{q} \in \Delta(K)$ . We shall call a forecaster weakly

calibrated if for every lipschitz “test function”  $g$  and any sequence  $y_1, y_2, \dots \in K$  we have

$$CS_T(g) := \text{calibration score w.r.t. } g$$

$$\overbrace{\left\| \frac{1}{T} \sum_{t=1}^T g(\mathbf{p}_t)(\mathbf{p}_t - \delta_{y_t}) \right\|_1} \rightarrow 0 \quad \text{almost surely as } T \rightarrow \infty. \quad (1)$$

Let  $I_{\mathbf{p},\epsilon} : \Delta(K) \rightarrow \mathbb{R}$  be the indicator defined as  $I_{\mathbf{p},\epsilon}(\mathbf{q}) := 1$  when  $\|\mathbf{p} - \mathbf{q}\|_1 \leq \epsilon$  and  $I_{\mathbf{p},\epsilon}(\mathbf{q}) := 0$  otherwise. A forecaster is called strongly calibrated if equation (1) holds for the indicator test functions  $g(\mathbf{q}) = I_{\mathbf{p},\epsilon}(\mathbf{q})$  for every  $\mathbf{p} \in \Delta(K)$  and  $\epsilon > 0$ . Given a fixed  $\epsilon$ , a forecaster is  $\epsilon$ -strongly calibrated if, for every  $\mathbf{p} \in \Delta(K)$ , the calibration score  $CS_T(I_{\mathbf{p},\epsilon}) \leq \epsilon$  for large enough  $T$ .

It should be clear from the definitions that strong calibration implies weak, but it may not be as obvious that the reverse does not hold. It has been shown that for the binary case ( $K = \{0, 1\}$ ) there exists a deterministic weakly calibrated forecaster (Kakade and Foster, 2008; Mannor et al., 2007), yet any strongly calibrated forecaster must be randomized. The latter is demonstrated by a well-known counter example of Dawid (1982). (For a complete survey on strongly calibrated forecasting, see Cesa-Bianchi and Lugosi (2006).)

The (strong or weak) calibration property leaves much to be desired as a measure of “performance” of a forecaster. For one, a forecaster can be calibrated yet completely inaccurate even on “easy” inputs: If the outcomes simply alternates as  $(0, 1, 0, 1, 0, 1, \dots)$  then the trivial forecaster predicting 0.5 achieves asymptotic calibration. Yet calibrated forecasting remains a useful tool in many circumstances, particularly given that it is robust to arbitrary and potentially adversarial inputs. It can be shown, for instance, that no-regret learning algorithms can be constructed from strongly-calibrated forecasters; the same trick can be applied towards a strategy for Blackwell’s Approachability problem.

### Open Problem: Can We Calibrate Efficiently?

The existence of a calibrated forecaster has been established for some time (strong (Foster and Vohra, 1998) and weak (Kakade and Foster, 2008)) in most cases via construction. Unfortunately, these constructions have typically been characterized by very greedy methods that give rise to inefficient algorithms. The most common approach is to take the probability space  $\Delta(K)$  and cover it with an  $\epsilon$ -grid and, for each grid point  $p$ , the algorithm maintains some statistic. To achieve  $\epsilon$ -strongly-calibrated forecaster, the algorithm will potentially have to process the entire grid *for each prediction*  $p_t$ .

This story may have a happy ending. Recent results suggest that calibration may be achieved via efficient methods. Mannor et al. (2007) developed a weakly calibrated forecasting algorithm that requires constant time and space for each prediction. In the present year’s COLT proceedings, Abernethy et al. (2011) developed an  $\epsilon$ -strongly calibrated forecasting algorithm which requires  $O(\log \frac{1}{\epsilon})$  time yet  $O(\frac{1}{\epsilon})$  space per prediction.

What’s the catch? Each of the above algorithm are applicable only to *binary* forecasting and do not extend to more general  $K$ . Mannor and Stoltz (2010) have recently introduced work addressing the case where  $|K| > 2$ , utilizing Blackwell approachability, to obtain  $\epsilon$ -strong calibration with time and space complexity that behaves like  $O(1/\epsilon^{|K|})$ . As described

in the table below, very little progress has been made towards efficient algorithms that achieve calibration in general.

	Strong	Weak
Binary ( $ K  = 2$ )	$O(\log \frac{1}{\epsilon})$ time, $O(\frac{1}{\epsilon})$ space	$O(1)$ time/space
Finite-alphabet ( $ K  = n > 2$ )	$O(1/\epsilon^{ K })$ time/space	?

We use the term “efficient” to denote an algorithm whose per-step complexity and memory requirements are poly in  $|K|$  and poly-logarithmic in  $(1/\epsilon)$ . Our concrete questions are:

1. Is there an efficient time and memory algorithm for  $\epsilon$ -strong calibration for  $|K| = 2$ ?
2. Is there an efficient time and memory algorithm for weak calibration for any  $|K|$ ?
3. Is there an efficient time and memory algorithm for strong calibration for any  $|K|$ ?

Our best guess is that such an algorithm likely exists, for both the weak and strong case, in particular because we have not placed any restrictions on the rate at which the algorithm must achieve the calibration objective. On the other hand, the previously-discovered tricks which lead to efficient calibrated forecasters may be very special to the binary case and we would not be surprised if no such efficient algorithms exist when  $|K| > 2$ .

## References

- J. Abernethy, P.L. Bartlett, and E. Hazan. Blackwell approachability and low-regret learning are equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.
- G.W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950. ISSN 1520-0493.
- Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. 2006. ISBN 0521841089, 9780521841085.
- A. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77:605–613, 1982.
- D. P Foster and R. V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379, 1998.
- S.M. Kakade and D.P. Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008.
- S. Mannor and G. Stoltz. A geometric proof of calibration. *Mathematics of Operations Research*, 35(4):721–727, 2010. URL [http://webee.technion.ac.il/people/shie/public/papers/J\\_MShimkin03Bayes.pdf](http://webee.technion.ac.il/people/shie/public/papers/J_MShimkin03Bayes.pdf).
- S. Mannor, J.S. Shamma, and G. Arslan. Online calibrated forecasts: Memory efficiency versus universality for learning in games. *Machine Learning*, 67(1):77–115, 2007. ISSN 0885-6125.

ABERNETHY MANNOR

## Bounds on Individual Risk for Log-loss Predictors

**Peter D. Grünwald** PDG@CWI.NL and **Wojciech Kotłowski** KOTLOWSK@CWI.NL  
*Centrum Wiskunde & Informatica*  
*Amsterdam, the Netherlands*

**Editor:** Sham Kakade, Ulrike von Luxburg

### Abstract

In sequential prediction with log-loss as well as density estimation with risk measured by KL divergence, one is often interested in the *expected instantaneous loss*, or, equivalently, the *individual risk* at a given fixed sample size  $n$ . For Bayesian prediction and estimation methods, it is often easy to obtain bounds on the *cumulative risk*. Such results are based on bounding the individual sequence regret, a technique that is very well known in the COLT community. Motivated by the easiness of proofs for the cumulative risk, our open problem is to use the results on cumulative risk to prove corresponding individual-risk bounds.

**Background** We consider sequential prediction (online learning) with log-loss (Cesa-Bianchi and Lugosi, 2006). In each iteration  $n = 1, 2, \dots$ , after observing a sequence of past outcomes  $x^n = x_1, x_2, \dots, x_n \in \mathcal{X}^n$ , a prediction strategy assigns a probability distribution on  $\mathcal{X}$ , denoted  $\hat{P}(\cdot | x^n)$ . Then, a next outcome  $x_{n+1}$  is revealed and the strategy incurs the *log loss*  $-\log \hat{P}(x_{n+1} | x^n)$ . The goal of the prediction strategy is to be not much worse than the best in a reference set of distributions (also called “experts”), which we call the *model*  $\mathcal{M}$ .

In online learning, the performance of a prediction strategy is usually measured by the *regret*, which is the difference between the accumulated loss of the prediction strategy and the best distribution in the model. The goal is then to minimize the regret in the worst case over all possible data sequences. This problem is relatively well-explored as it has been investigated in such fields as statistics, information theory, finance and machine learning. For example, it is known that: (1) when the model is finite (contains a finite number of distributions, say  $N$ ), it is possible to obtain a constant bound  $\log N$  on the regret, (2) when the model is infinite, but parametric (e.g. exponential families), a bound of the form  $\frac{k}{2} \log n + O(1)$  is usually possible, where  $k$  is the number of parameters (Grünwald, 2007).

In statistics, more focus is traditionally put on the *instantaneous* rather than cumulative losses of the prediction strategy: one wants the loss when predicting  $x_n$  to be small for fixed  $n$ , and to go to 0 at a fast rate as  $n$  increases. Since it is not possible to meaningfully bound instantaneous loss for adversarial data, one assumes that the data are sampled from a distribution  $P^*$ . Then, it is reasonable to define the *individual risk* or *instantaneous redundancy* in the  $n$ -th iteration as the difference between the expected loss of the prediction strategy and the expected loss of the best (w.r.t.  $P^*$ ) distribution in the model:

$$RISK_n(\hat{P}, P^*) = \mathbb{E}_{P^*}[-\log \hat{P}(X_{n+1} | X^n)] - \inf_{P \in \mathcal{M}} \{\mathbb{E}_{P^*}[-\log P(X_{n+1} | X^n)]\}$$

(note that we use capitals to denote both distributions and their densities/mass functions). Although our questions can be phrased more generally, for simplicity we will assume that data are i.i.d. and that  $P^* \in \mathcal{M}$ . Then the infimum in the above is attained by  $P^*$  and the expression simplifies to:

$$RISK_n(\hat{P}, P^*) = \mathbb{E}_{X^{n+1} \sim P^*}[-\log \hat{P}(X_{n+1}|X^n)] - \mathbb{E}_{X \sim P^*}[-\log P^*(X)]$$

The aforementioned results about regret immediately imply results about *cumulative* risk. For example, for  $k$ -parameter exponential families, Bayesian, ML (maximum likelihood prediction), NML (normalized maximum likelihood (“Shtarkov”)) and several other well-known strategies achieve cumulative risk  $\sum_{i=1}^n RISK_i(\hat{P}, P^*) = (k/2) \log n + O(1)$ . In general, especially for Bayesian strategies, it is easy to obtain bounds on the cumulative risk. For example, if  $\mathcal{M}$  is countable and  $\hat{P}$  is the Bayesian predictive distribution based on prior  $W$  such that  $W(P^*) > 0$ , then one has that

$$\sum_{i=1}^n RISK_i(\hat{P}, P^*) \leq -\log W(P^*).$$

The proof is completely straightforward using an individual-sequence regret argument (Grünwald, 2007, Chapter 6). The question we ask is what can be said about the individual risk of a prediction strategy  $\hat{P}$ , given the performance in terms of such cumulative risk. In particular, let  $\hat{P}$  be defined as a Bayesian predictive distribution. We ask:

## Our Questions

1. When the model is a  $k$ -parameter exponential family, is a bound of the form  $\frac{k}{2n} + O(1/n^2)$  possible?
2. When the model is countably infinite, is it possible to obtain a bound on the individual risk of the form  $\frac{-\log W(P^*)}{n^\gamma}$  for some  $\gamma > 0$  (preferably  $\gamma \geq 1$ )?

## Known results.

1. **Cesaro** As noted by e.g. Barron (1998); Yang (2000) (see also Catoni (1997)) it is possible to establish a relationship between the cumulative risk of a prediction strategy and the individual risk of a modified strategy using the notion of *Cesaro averaging*. Let  $\hat{P}$  be any prediction strategy and define:  $\hat{P}_{\text{Cesaro}}(x_{n+1}|x^n) = \frac{1}{n} \sum_{i=1}^n \hat{P}(x_{i+1}|x^i)$ . It turns out that  $RISK_n(\hat{P}_{\text{Cesaro}}, P^*) \leq \frac{1}{n} \sum_{i=1}^n RISK_i(\hat{P}, P^*)$ . Unfortunately, this statement does not say anything about the individual risk of the original strategy  $\hat{P}$ , only about its Cesaro average  $\hat{P}_{\text{Cesaro}}$ . In practice, the Cesaro average will often perform worse than the original  $\hat{P}$ : *Cesaro averaging is good to prove things, not to improve things*. Moreover, in question 1 above the Cesaro-strategy, when applied to Bayesian strategies, gives a rate bounded by  $n^{-1}(k/2) \log n$ , which is suboptimal by a factor of  $\log n$ : it is known that, e.g. with the ML estimator, an individual risk of  $O(1/n)$  is achievable.

**2. Follow the Leader** So far, the only case for which individual risk results are relatively well-studied seems to be the maximum likelihood (also known as “follow the leader”) strategy for exponential families. Grünwald and de Rooij (2005) proved that for one-parameter exponential family, when the data are generated i.i.d. by a distribution  $P^*$ , possibly outside  $\mathcal{M}$ , then the individual risk of the maximum likelihood decreases as:

$$RISK_n(\hat{P}, P^*) = \frac{1}{2n} \text{var}(P^*) \cdot I(\bar{P}) + O(1/n^2), \quad (1)$$

where  $\text{var}(P^*)$  is a variance of  $P^*$ ,  $I$  is a Fisher information, while  $\bar{P}$  is the element in  $\mathcal{M}$  closest to  $P^*$  in terms of KL-divergence  $D(P^* \| P)$ . In particular, if  $P^* \in \mathcal{M}$ , then  $P^* = \bar{P}$  and  $\text{var}(P^*) = I^{-1}(P^*)$ , so that the bound takes the form  $\frac{1}{2n} + O(1/n^2)$ , which is the optimal rate for a one-dimensional exponential family. Forster and Warmuth (2002) considered maximum likelihood for  $k$ -dimensional exponential families and managed to prove the bound of the form:

$$RISK_n(\hat{P}, P^*) \leq \frac{1}{2(n-1)} \text{tr}\{\text{cov}(P^*)\} \cdot \sup_{P \in \mathcal{M}} \|I(P)\|, \quad (2)$$

where  $\text{cov}(P^*)$  is the covariance matrix for  $P^*$ . The bounds (1) and (2) are very similar, but essentially incomparable. The latter is a true bound, which holds for all  $n$  and any exponential family, but the constant in front of  $O(\frac{1}{n})$  is not optimal. The former is an asymptotic expansion of the individual risk with the optimal constant in front of  $O(\frac{1}{n})$ . Both results concern only a particular prediction strategy (ML), which is known to be suboptimal when  $P^* \notin \mathcal{M}$ , and cannot be easily extended to say anything about any asymptotically optimal strategy, such as Bayes.

**3. 2-part MDL** If  $\hat{P}(X_{n+1} | X^n)$  is taken to be the 2-part MDL estimator achieving  $\min_{P \in \mathcal{M}} -\log W(P) - \log P(X^n)$ , then one can use a result due to Barron, Cover, Li and Zhang to get a bound on the squared Hellinger distance between  $\hat{P}$  and  $P^*$ . If all distributions in  $\mathcal{M}$  have uniformly bounded density ratios, i.e.  $\sup_{x \in \mathcal{X}, P, Q \in \mathcal{M}} P(x)/Q(x) < \infty$ , then this translates into a bound on the instantaneous risk. With the original bound (see (Grünwald, 2007, Chapter 15) for a simple statement and proof), one gets a bound  $O(-\log W(P^*)(\log n)/n)$  on the individual risk for the two-part MDL prediction strategy. This can be refined (Zhang, 2006) to get  $O(-\log W(P^*)/n)$ . Strangely, if  $\hat{P}$  is set to be a Bayesian predictive distribution (which usually works better in practice), then nothing is known about the individual risk.

**4. Decreasing risk!?** Let  $a_1, a_2, \dots$  be any sequence of numbers such that  $\sum_{i=1}^n a_i \leq C \log n$ . It can be easily shown (Grünwald, 2007) that such a sequence does not necessarily converge to 0. Bounding  $\sum_{i=1}^n a_i \leq C$  does imply that  $a_n$  converges to 0, but it can converge at arbitrarily slow rate. However, if we additionally assume that the sequence  $a_n$  is non-increasing, we immediately get optimal-rate bounds  $a_n \leq \frac{C}{n}$  in the first question. Thus, one strategy to address our questions for a given model  $\mathcal{M}$  would be to first show that individual risks of  $\hat{P}$  are monotonically *decreasing*. It is known that e.g. if  $\mathcal{M}$  is the Gaussian location family, then the risk of the ML predictions is strictly decreasing; on the other hand in some cases the risk of the Bayesian strategy can slightly increase at some  $n$ . Consider e.g. the Bernoulli model with a uniform prior, and assume the data is a sequence

of independent fair coin flips, i.e. they are i.i.d. Bernoulli  $1/2$ . In that case the risk at sample size 1 is 0, because the Bayesian predictive distribution based on the uniform prior and no data is  $P(X_1 = 1) = 1/2$ . At sample size 2, the Bayesian predictive distribution is  $P(X_2 = 1 | X_1 = x)$  which is either  $2/3$  (if  $x = 1$ ) or  $1/3$  (if  $x = 0$ ). In both cases, the risk increases Barron (1998); Grünwald (2007). So increasing risk is possible. Still, no examples are known of substantially increasing risk at large  $n$ . Thus, maybe one might prove that some tight enough upper bound on the risk is still decreasing...

## References

- A.R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In A.P. Dawid J.M. Bernardo, J.O. Berger and A.F.M. Smith, editors, *Bayesian Statistics*, volume 6, pages 27–52. Oxford University Press, Oxford, 1998.
- O. Catoni. A mixture approach to universal model selection. preprint LMENS 97-30, 1997. Available from <http://www.dma.ens.fr/edition/preprints/Index.97.html>.
- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.
- Jürgen Forster and Manfred K. Warmuth. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences*, 64(1):76–102, 2002.
- P. D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.
- P. D. Grünwald and S. de Rooij. Asymptotic log-loss of prequential maximum likelihood codes. In *Conference on Learning Theory (COLT 2005)*, pages 652–667, 2005.
- Y. Yang. Mixing strategies for density estimation. 28(1):75–87, 2000.
- Tong Zhang. From  $\epsilon$ -entropy to KL entropy: analysis of minimum information complexity density estimation. 34(5):2180–2210, 2006.

# A simple multi-armed bandit algorithm with optimal variation-bounded regret

**Elad Hazan**

*Technion*

*Haifa, Israel*

EHAZAN@IE.TECHNION.AC.IL

**Satyen Kale**

*Yahoo! Research*

*Santa Clara, CA 95054*

SKALE@YAHOO-INC.COM

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

We pose the question of whether it is possible to design a simple, linear-time algorithm for the basic multi-armed bandit problem in the adversarial setting which has a regret bound of  $O(\sqrt{Q \log T})$ , where  $Q$  is the total quadratic variation of all the arms.

We are interested in the fundamental multi-armed bandit (MAB) problem: iteratively at times  $t = 1, 2, \dots, T$  the decision maker has to choose (possibly randomly) one of  $n$  arms,  $i_t$ , and then receives the payoff of the arm, assumed to be in the range  $[0, 1]$ . The payoffs are constructed adversarially, as in (Auer et al., 2003), and we denote the payoff at time  $t$  for arm  $i$  by  $f_t(i) \in [0, 1]$ . The decision maker can only see her own payoff, and does not have access to the entire payoff vector  $f_t$  (otherwise it would have been the usual “experts” problem”). The goal is to minimize regret:

$$\text{Regret} = \max_{i \in [n]} \sum_{t=1}^T f_t(i) - \sum_{t=1}^T f_t(i_t),$$

where  $i_t$  is the arm chosen by the algorithm in round  $t$ . If the algorithm is randomized, then the aim is to minimize the expected regret over the internal randomization of the algorithm.

The EXP3 algorithm of Auer et al. (2003), is an efficient (linear in  $n$  time) algorithm that obtains regret of  $O(\sqrt{nT \log T})$ . This is optimal in  $T, n$  up to logarithmic factors.

But the quest for an optimal algorithm for this fundamental problem is not over. Cesa-Bianchi et al. (2007) conjectured that it should be possible to bound the regret of online learning algorithms by the *quadratic variation* in payoffs. For the MAB problem as defined above, we can define the quadratic variation by:

$$Q = \sum_{i=1}^T \|f_t - \mu\|^2,$$

where  $\mu = \frac{1}{T} \sum_{t=1}^T f_t$  is the mean payoff. This is a natural parameter for measuring the difficulty of an online instance, as argued in Cesa-Bianchi et al. (2007), since it is related to the statistical properties of the underlying payoff sequences. Essentially, we may think

of the quadratic variability as the variance in our data, and the difficulty of learning should be proportional to how much the data deviates from the mean rather than the length of the prediction sequence. As a special case,  $Q$  can be a constant independent of the data sequence, in which case we would like our regret to remain constant independent of the number of iterations (this is motivated by financial applications, as in (Hazan and Kale, 2009b)).

Recently, online learning algorithms that bound the regret as a function of  $Q$  rather than  $T$  have been developed. For the online linear optimization setting this was obtained in (Hazan and Kale, 2008), and for the MAB setting, the following theorem was proven in (Hazan and Kale, 2009a):

**Theorem 1** *There exists a polynomial-time MAB algorithm whose regret is bounded by  $O(n^2\sqrt{Q}\log T)$ .*

Since our payoffs are bounded by one, it holds that  $Q \leq nT$ , and hence as  $T$  grows large the above bound is superior to the EXP3 bound (for certain ranges of  $Q$ ). However, the algorithm used to obtain this bound is rather complicated: it is based on self-concordant barrier functions as regularizers, which were introduced to learning theory in (Abernethy et al., 2008), and applies to a more general setting of bandit online linear optimization than MAB. This technology makes the algorithm poly-time, but not nearly linear time and simple as EXP3. More importantly, the above bound is sub-optimal in terms of  $n$ .

**The open question is to design a simple, linear-time algorithm for MAB which has a regret bound of  $O(\sqrt{Q}\log T)$ , hence improving upon EXP3.**

We conjecture that such an algorithm exists, and it should not use any self-concordance technology. Rather, it should be basic, perhaps based on the multiplicative updates method, and bear resemblance to EXP3. We note that EXP3 itself has  $\Omega(\sqrt{T})$  regret, since it mixes with the uniform distribution every iteration to enable sufficient exploration. Hence, the desired algorithm should be a little different from EXP3, incorporating just enough exploration proportional to the variation in the data.

One possible feature of the new algorithm is to use an unbiased estimator for the payoff vector  $f_t$  constructed by estimating the empirical mean and the deviation from the mean separately, as done in (Hazan and Kale, 2009a). An unbiased estimator for the mean can be constructed using the reservoir sampling ideas in (Hazan and Kale, 2009a). The deviation from the mean can be computed using importance weighted sampling as in EXP3.

## References

- J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, 2008.
- P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003. ISSN 0097-5397.
- N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66(2-3):321–352, 2007. ISSN 0885-6125.

- E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. In *COLT*, 2008.
- E. Hazan and S. Kale. Better algorithms for benign bandits. In *SODA*, pages 38–47, 2009a.
- E. Hazan and S. Kale. On stochastic and worst-case models for investing. In *NIPS*, 2009b.

HAZAN KALE

# Minimax Algorithm for Learning Rotations

**Wojciech Kotłowski**

*Centrum Wiskunde & Informatica*

KOTLOWSK@CWI.NL

**Manfred K. Warmuth**

*Department of Computer Science, UC Santa Cruz*

MANFRED@CSE.UCSC.EDU

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

It is unknown what is the most suitable regularization for rotation matrices and how to maintain uncertainty over rotations in an online setting. We propose to address these questions by studying the minimax algorithm for rotations and begin by working out the 2-dimensional case.

The problem of online learning of rotations is defined as follows. In every iteration  $t = 1, 2, \dots, T$ , the learner is given a unit vector  $\mathbf{x}_t$  ( $\|\mathbf{x}_t\| = 1$ ). The learner is then required to predict (deterministically or randomly), with a rotation matrix  $\mathbf{R}_t \in \mathcal{SO}(n)$ . The choice of  $\mathbf{R}_t$  determines the predicted unit vector  $\hat{\mathbf{y}}_t = \mathbf{R}_t \mathbf{x}_t$ . Finally, the algorithm obtains the “true” rotated unit vector  $\mathbf{y}_t$  and incurs loss

$$L_t(\mathcal{R}_t) = \frac{1}{2} \mathbb{E} [\|\mathbf{R}_t \mathbf{x}_t - \mathbf{y}_t\|^2] = \frac{1}{2} \mathbb{E} [\underbrace{\|\mathbf{R}_t \mathbf{x}_t\|^2}_1 + \underbrace{\|\mathbf{y}_t\|^2}_1 - 2(\mathbf{R}_t \mathbf{x}_t) \cdot \mathbf{y}_t] = 1 - (\mathbb{E}[\mathbf{R}_t] \mathbf{x}_t) \cdot \mathbf{y}_t,$$

where  $\cdot$  is the dot product,  $\mathcal{R}_t$  is the distribution over  $\mathcal{SO}(n)$  from which  $\mathbf{R}_t$  is drawn, and  $\mathbb{E}[\cdot]$  is the expectation wrt  $\mathcal{R}_T$ . We seek on-line algorithms which have small bounded regret

$$\sum_{t=1}^T L_t(\mathcal{R}_t) - \min_{\mathbf{R} \in \mathcal{SO}(n)} \sum_{t=1}^T L_t(\mathbf{R})$$

for arbitrary sequences of examples  $(\mathbf{x}_t, \mathbf{y}_t)$  of length  $T$ . Recently, a regret bound of  $2\sqrt{nT}$  has been proven for a randomized<sup>1</sup> algorithm which does a gradient descent step in each iteration and then projects into a suitable chosen convex set (Hazan et al., 2010). Even though the regret of this algorithm was shown to be optimal within a constant factor (Hazan et al., 2010), many questions remain for this archetypical machine learning problem that has many applications in robotics, vision, matrix completion, subspace tracking, etc (See e.g. Arora (2009); Hazan et al. (2010)):

1. We don’t know the proper way to regularize rotations. Is there some kind of entropy defined over  $\mathcal{SO}(n)$ ? The algorithm of (Hazan et al., 2010) is based on regularizing wrt the squared Euclidean distance, which does not take the structure of the  $\mathcal{SO}(n)$  into account.

---

1. Any deterministic algorithm can be forced to have regret  $\Omega(T)$  (Hazan et al., 2010).

2. The parameter space  $\mathcal{SO}(n)$  is not convex and we don't know the "correct" way to maintain uncertainty over this space. The algorithm of (Hazan et al., 2010) projects using inequality constraints which means that it "forgets" information about the past examples.
3. A good on-line algorithm should intuitively exploit the elegant Lie group and Lie algebra connection (via the exponential map) between  $\mathcal{SO}(n)$  and skew symmetric matrices, respectively (Arora, 2009).

We propose to resolve some of these issues by finding the minimax algorithm for learning rotations and we hope that this algorithm will give insights for learning other matrix classes:

$$\mathcal{R}_t = \operatorname{argmin}_{\mathcal{R}_t} \max_{\mathbf{y}_t} \max_{\mathbf{x}_{t+1}} \min_{\mathcal{R}_{t+1}} \max_{\mathbf{y}_{t+1}} \dots \max_{\mathbf{x}_T} \min_{\mathcal{R}_T} \max_{\mathbf{y}_T} \left( \sum_{q=t}^T L_q(\mathcal{R}_q) - \min_{\mathbf{R} \in \mathcal{SO}(n)} \sum_{t=1}^T L_t(\mathbf{R}) \right),$$

where the  $\mathcal{R}_q$  are distributions over  $\mathcal{SO}(n)$  and the unit vectors  $\mathbf{x}_q, \mathbf{y}_q$  are chosen deterministically.

So far, we have obtained the following partial result sketched below: If the instances  $\mathbf{x}_t$  are restricted to be a fixed unit, say  $\mathbf{e}_1 = (1, 0, \dots, 0)^\top$ , then we can give the minimax algorithm. In the case of  $n = 2$ , the two problems coincide because any black box for solving the fixed instance problem can be used to solve the variable instance problem with the same regret. This holds because if  $\mathbf{R}_{\mathbf{x}_t}$  rotates  $\mathbf{x}_t$  onto  $\mathbf{e}_1$  (i.e.  $\mathbf{e}_1 = \mathbf{R}_{\mathbf{x}_t} \mathbf{x}_t$ ), then processing  $(\mathbf{x}_t, \mathbf{y}_t)$  is the same as processing  $(\mathbf{e}_1, \mathbf{y}'_t)$ , where  $\mathbf{y}'_t := \mathbf{R}_{\mathbf{x}_t} \mathbf{y}_t$ :

$$\|\mathbf{R}_t \mathbf{x}_t - \mathbf{y}_t\| = \|\mathbf{R}_t \mathbf{R}_{\mathbf{x}_t}^{-1} \mathbf{e}_1 - \mathbf{y}_t\| = \|\mathbf{R}_{\mathbf{x}_t}^{-1} \mathbf{R}_t \mathbf{e}_1 - \mathbf{y}_t\| = \|\mathbf{R}_{\mathbf{x}_t}^{-1} (\mathbf{R}_t \mathbf{e}_1 - \mathbf{y}'_t)\| = \|\mathbf{R}_t \mathbf{e}_1 - \mathbf{y}'_t\|.$$

Note that in the 2nd equality we used  $\mathbf{R}_t \mathbf{R}_{\mathbf{x}_t}^{-1} = \mathbf{R}_{\mathbf{x}_t}^{-1} \mathbf{R}_t$ , which only holds for  $n = 2$ .

When  $\mathbf{x}_t = \mathbf{e}_1$  for all  $t$ , then the loss  $1 - (\mathbb{E}[\mathbf{R}_t] \mathbf{e}_1) \cdot \mathbf{y}_t$  can be rewritten as  $1 - \mathbf{w}_t \cdot \mathbf{y}_t$ , where  $\mathbf{w}_t = \mathbb{E}[\mathbf{R}_t] \mathbf{e}_1$  is a new parameter vector which has norm at most 1. With this parameter vector, the regret simplifies to

$$\sum_{t=1}^T (1 - \mathbf{w}_t \cdot \mathbf{y}_t) - \inf_{\mathbf{w}: \|\mathbf{w}\| \leq 1} \left\{ \sum_{t=1}^T (1 - \mathbf{w} \cdot \mathbf{y}_t) \right\} = - \sum_{t=1}^T \mathbf{w}_t \cdot \mathbf{y}_t + \mathbf{w}^* \cdot \mathbf{s}_T = \sum_{t=1}^T -\mathbf{w}_t \cdot \mathbf{y}_t + \|\mathbf{s}_T\|,$$

where  $\mathbf{s}_T = \sum_{t=1}^T \mathbf{y}_t$ , and  $\mathbf{w}^* = \arg \max_{\mathbf{w}, \|\mathbf{w}\| \leq 1} \mathbf{w} \cdot \mathbf{s}_T = \frac{\mathbf{s}_T}{\|\mathbf{s}_T\|}$ . To find the optimal strategy of the forecaster, we proceed backwards. Fix  $\mathbf{s}_{T-1}$  and  $\mathbf{w}_1, \dots, \mathbf{w}_{T-1}$ . We want to solve the following minimax problem in the last iteration:

$$\min_{\mathbf{w}: \|\mathbf{w}\| \leq 1} \max_{\mathbf{y}: \|\mathbf{y}\|=1} \{-\mathbf{w} \cdot \mathbf{y} + \|\mathbf{s}_{T-1} + \mathbf{y}\|\}. \quad (1)$$

A more involved analysis reveals that the optimal solution  $\mathbf{w}_T$  must be  $\mathbf{s}_{T-1}$  times a shrinking factor:

$$\mathbf{w}_T = \frac{\mathbf{s}_{T-1}}{\sqrt{\|\mathbf{s}_{T-1}\|^2 + 1}},$$

while the optimal (worst-case) outcome  $\mathbf{y}_T$  is *orthogonal* to  $\mathbf{s}_{T-1}$ . Plugging  $\mathbf{w}_T$  and  $\mathbf{y}_T$  into (1) and using  $\mathbf{y}_T \cdot \mathbf{s}_{T-1} = \mathbf{y}_T \cdot \mathbf{w}_T = 0$  gives the optimal value of the regret increase in the

last iteration:  $\sqrt{\|\mathbf{s}_{T-1}\|^2 + 1}$ . In the second to the last step

$$\mathbf{w}_{T-1} = \operatorname{argmin}_{\mathbf{w}: \|\mathbf{w}\| \leq 1} \max_{\mathbf{y}: \|\mathbf{y}\|=1} \left\{ -\mathbf{w} \cdot \mathbf{y} + \sqrt{\|\mathbf{s}_{T-2} + \mathbf{y}\|^2 + 1} \right\} = \frac{\mathbf{s}_{T-2}}{\sqrt{\|\mathbf{s}_{T-2}\|^2 + 2}}$$

and  $\mathbf{y}_{T-1}$  is orthogonal to  $\mathbf{s}_{T-2}$ . Plugging  $\mathbf{w}_{T-1}$  and  $\mathbf{y}_{T-1}$  into the optimized expression leads to the worst-case regret increase in the last two iterations which is  $\sqrt{\|\mathbf{s}_{T-2}\|^2 + 2}$ . Continuing the backward induction, in the  $k$ -th step from the end, we optimize

$$\mathbf{w}_{T-k+1} = \operatorname{argmin}_{\mathbf{w}: \|\mathbf{w}\| \leq 1} \max_{\mathbf{y}: \|\mathbf{y}\|=1} \left\{ -\mathbf{w} \cdot \mathbf{y} + \sqrt{\|\mathbf{s}_{T-k} + \mathbf{y}\|^2 + k - 1} \right\} = \frac{\mathbf{s}_{T-k}}{\sqrt{\|\mathbf{s}_{T-k}\|^2 + k}},$$

and the worst-case regret increase in the last  $k$  iterations equals  $\sqrt{\|\mathbf{s}_{T-k}\|^2 + k}$ . The value of the minimax regret can be obtained for  $k = T$  and is equal to  $\sqrt{T}$ .

Summarizing, we were able to prove that when the input  $\mathbf{x}_t$  is restricted to be a fixed vector, then the minimax regret is  $\sqrt{T}$  and does not depend on the dimension. The optimal strategy for this case is to choose  $\mathbf{w}_t$  as the current “sufficient statistic”  $\mathbf{s}_{t-1} = \sum_{q=1}^{t-1} \mathbf{y}_q$  times a shrinking factor that is related to the randomization. The worst-case data sequence for minimax algorithm is any sequence where the outcomes are always orthogonal to the current sufficient statistic (and the vector chosen by the optimal strategy).

For  $n = 2$ , the minimax regret for the fixed instance problem coincides with the minimax of the original rotation problem<sup>2</sup> and the open problem is to determine the minimax regret for dimension  $n > 2$ .

## Acknowledgments

This research was supported by NSF grant IIS-0917397.

## References

- R. Arora. On learning rotations. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 55–63. MIT Press, 2009.
- E. Hazan, S. Kale, and M. K. Warmuth. Learning rotations with little regret. In *COLT '10*, June 2010. A corrigendum can be found at the conference website.

---

2. For  $n = 2$ , the algorithm of (Hazan et al., 2010) has a regret bound of  $2\sqrt{2T}$ , whereas we show that the minimax regret for learning rotations is  $\sqrt{T}$ .

KOTŁOWSKI WARMUTH

# Missing Information Impediments to Learnability

**Loizos Michael**

*Open University of Cyprus*

LOIZOS@OUC.AC.CY

**Editor:** Sham Kakade, Ulrike von Luxburg

## Abstract

To what extent is learnability impeded when information is missing in learning instances? We present relevant known results and concrete open problems, in the context of a natural extension of the PAC learning model that accounts for arbitrarily missing information.

**Keywords:** PAC learning, missing information, masking process, open problems.

## 1. Learning from Partial Observations

In the PAC learning model (Valiant, 1984), *examples* are drawn from some unknown fixed probability distribution  $\mathcal{D}$  over  $\{0, 1\}^n$ . A boolean-valued label for each example is determined by applying an unknown fixed *target* function  $f \in \mathcal{C}$  on the example; the class  $\mathcal{C}$  of all such targets is the *concept class*. Given access to a set of labeled examples during a training phase, a learner seeks to produce, efficiently and with high probability, a *hypothesis* function  $h \in \mathcal{H}$  that predicts, with high probability, the labels of examples drawn from  $\mathcal{D}$  and labeled according to  $f$ ; the class  $\mathcal{H}$  of all such hypotheses is the *hypothesis class*.

Explicit in the definition of the PAC learning model is the requirement that each example offers sufficient information to determine its label; the primary challenge of learning is, thus, to efficiently identify *how* to determine the label. In certain settings (e.g., in a typical medical database), however, not all information necessary to determine the label is available in an example (e.g., due to medical tests that were not performed). Furthermore, this happens during *both* the training *and* the testing phase, and the manner in which information is missing may critically depend on the information itself. In the spirit of supervised learning, we consider only settings where example labels are never missing during the training phase.

These partial (but noiseless) views of examples we shall call *observations*. We represent them as ternary vectors  $\text{obs} \in \{0, 1, *\}^n$ , with the value  $*$  indicating that the corresponding attribute was not observed. Examples are mapped to observations through a *masking process*, a stochastic process  $\text{mask} : \{0, 1\}^n \rightarrow \{0, 1, *\}^n$  that induces a probability distribution over observations, which may depend on the example being mapped. The noiseless nature of observations implies that whenever an observation  $\text{obs}$  is drawn from  $\text{mask}(\text{exm})$ , it holds that  $\text{obs}[i] \in \{\text{exm}[i], *\}$ , where  $\text{obs}[i]$  and  $\text{exm}[i]$  correspond, respectively, to the value of the  $i$ -th attribute according to  $\text{obs}$  and  $\text{exm}$ . Such an observation  $\text{obs}$  is said to *mask* the example  $\text{exm}$ , and each attribute with  $\text{obs}[i] = *$  is said to be *masked* in  $\text{obs}$ .

Each observation is assumed to be drawn from the oracle  $\text{sense}(\mathcal{D}; f; \text{mask})$  in unit time, by means of the following process: (i) an example  $\text{exm}$  is drawn from  $\mathcal{D}$ ; (ii) the label of  $\text{exm}$  is computed to be  $f(\text{exm})$ ; (iii) an observation  $\text{obs}$  that masks  $\text{exm}$  is drawn from  $\text{mask}(\text{exm})$ ; (iv) the label of  $\text{obs}$  is assigned to equal  $f(\text{exm})$ ; (v) both  $\text{obs}$  and  $f(\text{exm})$  are returned.

As in the PAC learning model, a learner in the model that we consider seeks to produce a hypothesis for predicting the labels. The hypothesis is a boolean-valued function over the *boolean attributes*, and encodes (learned) knowledge about the structure of the *underlying examples* — not knowledge about the structure of observations and the way the masking process hides information (cf. Schuurmans and Greiner, 1994). The PAC learning model can be viewed as the special case of this model when the masking process `mask` is an identity.

Since a hypothesis  $h$  is defined over boolean attributes but evaluated on observations, its prediction  $h(\text{obs})$  on observation `obs` may possibly remain undefined; this occurs exactly when  $h(\text{exm})$  is not constant across all examples `exm` masked by `obs`. In such a case,  $h$  abstains from making a prediction. Abstentions are not penalized, as they are not actively chosen by the hypothesis. We shall say that a hypothesis  $h$  has a *consistency conflict* with an observation `obs` if  $h$  does not abstain, and  $h(\text{obs})$  differs from the label of `obs`.

A hypothesis  $h$  is  $\varepsilon$ -*inconsistent* w.r.t. oracle `sense`( $\mathcal{D}; f; \text{mask}$ ) if  $h$  has a consistency conflict with an observation `obs` drawn from `sense`( $\mathcal{D}; f; \text{mask}$ ) with probability at most  $\varepsilon$ .

**Definition 1** A concept class  $\mathcal{C}$  is *consistently learnable* by a hypothesis class  $\mathcal{H}$  if there exists an algorithm  $\mathcal{L}$  such that for every natural number  $n$ , every probability distribution  $\mathcal{D}$  over  $\{0, 1\}^n$ , every target function  $f \in \mathcal{C}$  over  $n$  attributes, every masking process `mask` over  $n$  attributes, and every pair of real numbers  $\delta, \varepsilon \in (0, 1]$ , algorithm  $\mathcal{L}$  is such that:

given the parameters  $n, \mathcal{C}, \mathcal{H}, \delta, \varepsilon$  as input, and given access to the oracle `sense`( $\mathcal{D}; f; \text{mask}$ ), algorithm  $\mathcal{L}$  runs in time polynomial in  $n, 1/\delta, 1/\varepsilon$ , and the size of  $f$ , and returns, with probability at least  $1 - \delta$ , a hypothesis  $h \in \mathcal{H}$  that is  $\varepsilon$ -inconsistent w.r.t. `sense`( $\mathcal{D}; f; \text{mask}$ ).

The definition of consistent learnability insists that the typical PAC guarantees hold, but for *every* masking process. It is worth pointing out that the resulting learning requirements are not overly demanding, since exactly when learnability may suffer due to less information in observations, hypotheses may abstain more and avoid consistency conflicts. Abstentions cannot, however, be abused, as they cannot be actively invoked. It is the masking process that effectively determines when hypotheses abstain, and this is beyond the learner's control.

The model of consistent learnability presented herein is a special case of the *autodidactic learning model* (Michael, 2008, 2010), where there is no distinguished label for observations, and the aim of the learning process is to complete the values of the masked attributes. The results in the section that follows were obtained in the context of the latter model. Proofs of the results, details about that model, and comparison to other extensions of the PAC learning model that accommodate missing information, can be found in the cited works.

## 2. Known Results and Open Problems

Since consistent learnability implies PAC learnability, the latter is a necessary condition for the former. PAC learnability in conjunction with either the monotone or the read-once property holding for the concept class is a sufficient condition for consistent learnability.

**Theorem 2** A concept class  $\mathcal{C}$  that comprises either monotone or read-once formulas is consistently learnable by a hypothesis class  $\mathcal{H}$ , assuming that  $\mathcal{C}$  is PAC learnable by  $\mathcal{H}$ .

Thus, the concept classes of conjunctions and linear thresholds (Kearns and Vazirani, 1994) are consistently learnable. Unlike what holds in the PAC learning model, a learning reduction cannot be readily employed to establish the learnability of  $k$ -CNFs for constant values of  $k \geq 2$ . This holds because  $k$ -CNFs cannot be evaluated *modularly* on observations (unlike on examples). Indeed, the value of a certain conjunction of two subformulas on some observation may not be determinable only by the values of the subformulas (e.g., when they are undefined), but may require knowledge of the subformulas themselves. Hence:

**Problem 3** *Is the concept class  $\mathcal{C}$  of 2-CNFs consistently learnable by a hypothesis class  $\mathcal{H}$ ? Is the question true for any concept class of formulas that are not modularly evaluable?*

The case of learning 3-CNFs presents an additional challenge when compared to the case of learning 2-CNFs, since the former formulas are not believed to be evaluable *efficiently*. Indeed, their evaluation on the observation  $*^n$  implies deciding their satisfiability. Hence:

**Problem 4** *Is the concept class  $\mathcal{C}$  of 3-CNFs consistently learnable by a hypothesis class  $\mathcal{H}$ ? Is the question true for any concept class of formulas that are not efficiently evaluable?*

Despite being a necessary condition, PAC learnability is not, by itself, a sufficient condition for consistent learnability — at least not when the hypothesis class  $\mathcal{H}$  and the concept class  $\mathcal{C}$  are required to coincide, and the complexity condition  $\text{RP} \neq \text{NP}$  is assumed.

**Theorem 5** *The concept class  $\mathcal{C}$  that comprises either parities or monotone-term 1-decision lists is not consistently learnable by the hypothesis class  $\mathcal{H} = \mathcal{C}$ , unless  $\text{RP} = \text{NP}$ .*

The negative result holds despite  $\mathcal{C}$  being PAC learnable by  $\mathcal{H} = \mathcal{C}$  (Kearns and Vazirani, 1994), and even when at most three attributes are masked in each observation. Hence:

**Problem 6** *Is the concept class  $\mathcal{C}$  that comprises either parities or monotone-term 1-decision lists consistently learnable by a hypothesis class  $\mathcal{H}$  that differs from  $\mathcal{C}$ ?*

Refining the necessary and sufficient conditions for consistent learnability would help clarify which PAC learnability results remain true when information is missing arbitrarily, and, hence, which can be applied in realistic settings where the masking process is unknown.

## References

- Michael Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, Cambridge, Massachusetts, U.S.A., 1994.
- Loizos Michael. *Autodidactic Learning and Reasoning*. PhD thesis, School of Engineering and Applied Sciences, Harvard University, U.S.A., May 2008.
- Loizos Michael. Partial Observability and Learnability. *Artificial Intelligence*, 174(11): 639–669, July 2010.
- Dale Schuurmans and Russell Greiner. Learning Default Concepts. In *Proceedings of the Tenth Canadian Conference on Artificial Intelligence (AI'94)*, pages 99–106, May 1994.
- Leslie Valiant. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

MICHAEL

# Monotone multi-armed bandit allocations

Aleksandrs Slivkins

SLIVKINS@MICROSOFT.COM

*Microsoft Research Silicon Valley, Mountain View, CA 94043, USA*

We present a novel angle for multi-armed bandits (henceforth abbreviated MAB) which follows from the recent work on *MAB mechanisms* (Babaioff et al., 2009; Devanur and Kakade, 2009; Babaioff et al., 2010). The new problem is, essentially, about designing MAB algorithms under an additional constraint motivated by their application to MAB mechanisms.

This note is self-contained, although some familiarity with MAB is assumed; we refer the reader to Cesa-Bianchi and Lugosi (2006) for more background.

## 1. Problem formulation

We start with a slightly non-standard formalism for MAB.

**Definition 1 (MAB)** *In each round, an algorithm selects among  $k$  alternatives (arms), and collects a reward. The rewards are fixed in advance, but not revealed to the algorithm. Specifically, one fixes a table whose  $(i, t)$ -th entry is the reward of arm  $i$  in round  $t$ ; such table is called a realization (of the rewards). The realization is generated by a random process (generator). The algorithm only knows that the generator belongs to some set (of generators) called the MAB domain.<sup>1</sup>*

Two natural special cases are *stochastic rewards*: for each arm  $i$ , the reward of this arm in every given round is an independent sample from the same distribution  $\mathcal{D}_i$ , and *adversarial rewards*: the realization is chosen by an adversary.

**Definition 2 (MAB allocation rule)** *An MAB allocation rule is an MAB algorithm which initially inputs a vector of bids  $b_i \in [0, 1]$  for each arm  $i$ ; the rewards received from arm  $i$  (raw rewards) are then scaled by a factor of  $b_i$ . An MAB allocation rule is called monotone (resp., ex-post monotone) if for each agent arm  $i$ , increasing  $b_i$  and keeping all other bids fixed cannot decrease the raw reward from arm  $i$ , in expectation over the realization (resp., for every realization).*

Now we are ready to state the problem:

**Problem 1** *For a given MAB domain, design an MAB allocation rule so as to maximize the total reward subject to the constraint of (ex-post) monotonicity.*

---

1. The notation “MAB domain” is not standard; we adopt it here for the ease of presentation.

It is worth emphasizing that Problem 1 is well-defined for each of the numerous MAB domains studied in the literature: stochastic or adversarial, with or without priors on rewards, with or without auxiliary or contextual information, and so forth. For each MAB domain there is a corresponding version of Problem 1. The main qualitative issue is whether and by how much the additional constraint of (ex-post) monotonicity impacts performance.

## 2. Brief motivation

The motivating example is the following idealized model for selling ad slots to potential advertisers. Several advertisers, each with a single ad, are competing for a single ad slot that is displayed to multiple users over time. Each advertiser derives value only if her ad is clicked by a user. Thus, each arm corresponds to an advertiser, the bid corresponds to the value per click, and the raw reward in each round is 1 if clicked and 0 otherwise.

To motivate the (ex-post) monotonicity property we need to extend the above model to a simple auction, called *MAB auction*: first all advertisers submit their bids, then the ad slots are allocated using an MAB allocation rule, and finally the ad platform assigns how much each agent needs to pay. Crucially, the value per click ( $v_i$ ) is a *private information*: it is known to the advertiser, but not to the ad platform. Thus, an advertiser may *lie* about it ( $b_i \neq v_i$ ) if she thinks it may benefit her. This setting have been defined independently and concurrently in Devanur and Kakade (2009) and Babaioff et al. (2009), and further studied in Babaioff et al. (2010); see Appendix A for some background and motivation.

In an MAB auction, the monotonicity property of the MAB allocation rule is necessary and sufficient (with appropriate payments) to incentivize the advertisers to submit truthful bids. Monotonicity corresponds to incentives in expectation over realizations, whereas a stronger and more desirable ex-post monotonicity corresponds to incentives for every given realization. The details are deferred to Appendix A; these details are not necessary for understanding Problem 1.

It is worth noting that the *social welfare* of an MAB auction – a standard performance benchmark for auctions – coincides with the total reward of the corresponding allocation rule. Social welfare is defined as the total utility: the sum of utilities of all advertisers and the utility (profit) of the ad platform (so payments cancel out).

## 3. Current status

The problem has been resolved for stochastic rewards in the strongest possible sense: there exists an ex-post monotone MAB allocation rule whose regret is essentially optimal among all MAB allocation rules (Babaioff et al., 2010). Moreover, an MAB allocation rule based on (a version of) a well-known MAB algorithm UCB1 (Auer et al., 2002a) is proved to be monotone; this result extends to a more general family of MAB allocation rules. However, UCB1-based MAB allocation rule is *not* ex-post monotone as is; making it ex-post monotone seems to require significant modifications.

For all other MAB domains the problem is open. One particularly appealing target is adversarial rewards.<sup>2</sup> Babaioff et al. (2009) exhibit an ex-post monotone MAB allocation rule (based on the MAB algorithm in Awerbuch and Kleinberg (2008)) which separates

---

2. Note that for adversarial rewards, the notions of ex-post monotonicity and monotonicity coincide.

exploration and exploitation and has regret  $\tilde{O}(T^{2/3})$ . This is in contrast to optimal MAB algorithms such as EXP3 (Auer et al., 2002b) that achieve regret  $\tilde{O}(\sqrt{T})$ . We conjecture that EXP3-based MAB allocation rule is *not* ex-post monotone. It is not clear whether an MAB allocation rule with regret  $\tilde{O}(\sqrt{T})$  can be ex-post monotone.

## Acknowledgments

We thank Robert Kleinberg for his comments and suggestions.

## References

- Aaron Archer and Éva Tardos. Truthful mechanisms for one-parameter agents. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 482–491, 2001.
- Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a. Preliminary version in *15th ICML*, 1998.
- Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b. Preliminary version in *36th IEEE FOCS*, 1995.
- Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, February 2008. Preliminary version appeared in *36th ACM STOC*, 2004.
- Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *10th ACM Conf. on Electronic Commerce (EC)*, pages 79–88, 2009. Full version available at <http://arxiv.org/abs/0812.2291>.
- Moshe Babaioff, Robert Kleinberg, and Aleksandrs Slivkins. Truthful mechanisms with implicit payment computation. In *11th ACM Conf. on Electronic Commerce (EC)*, pages 43–52, 2010. Best Paper Award. Full version available at <http://arxiv.org/abs/1004.3630>.
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- Nikhil Devanur and Sham M. Kakade. The price of truthfulness for pay-per-click auctions. In *10th ACM Conf. on Electronic Commerce (EC)*, pages 99–106, 2009.
- Roger B. Myerson. Optimal Auction Design. *Mathematics of Operations Research*, 6:58–73, 1981.
- N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani (eds.). *Algorithmic Game Theory*. Cambridge University Press., 2007.
- Tim Roughgarden. An algorithmic game theory primer. IFIP International Conference on Theoretical Computer Science (TCS). An invited survey., 2008.
- Balasubramanian Sivan and Christopher A. Wilkens. Single-call mechanisms, 2010. Available at <http://arxiv.org/abs/1011.6134>.

## Appendix A. MAB auctions: motivation and background

**Sponsored search and MAB auctions.** Sponsored search on the Web is a billions dollar market in which ad slots are sold to potential advertisers. The ad slots are usually sold via auctions, with numerous auctions running every second. In a given auction, multiple advertisers compete for a well-defined selection of ad slots, such as ad slots that appear next to search results on a given keyword. The predominant model, called *pay-per-click (PPC)* auctions, is that an advertiser pays only if her ad is clicked by a user. The underlying assumption here is that an advertiser derives value only if her ad is clicked, and the PPC model shields advertisers from the risk that the allocation of ads to ad slots may be suboptimal for a given advertiser. A typical PPC auction proceeds as follows: advertisers submit their preferences, then the ad platform allocates ad slots between the advertisers, records the clicks, and then the payments are assigned based on the submitted preferences, the allocation, and the observed clicks.

In general, optimizing performance in a PPC auction requires knowing or estimating the rates at which the ads are clicked (*click-through rates*, or *CTRs*). Most prior work assumes that these rates are known or estimated externally (see Babaioff et al. (2009) for a more thorough discussion).

A recent line of work (Devanur and Kakade, 2009; Babaioff et al., 2009, 2010) considers the problem of designing truthful PPC auctions when the process of learning the CTRs is explicitly treated as a part of the game. The motivation is that the self-interested advertisers would take this process into account, as it influences their utility. These papers are mainly interested in the interplay of the two key issues: the *strategic* issue (the issue of incentives) and the issue of learning CTRs from the clicks that are observed over time. They define and study a clean model, called *MAB auctions*, which abstracts these two issues. This model (which we described in Section 2) can be seen as a natural *strategic* version of MAB.

**Incentives and monotonicity.** As we alluded in Section 2, the issue of incentives is crucial. The goal is to design mechanisms that are *incentive-compatible* in the following sense: every agent maximizes her expected utility by bidding truthfully, for any bids of the others. This is a standard notion in the literature on Mechanism Design.<sup>3</sup> A stronger notion, *ex-post incentive-compatibility*, is the same for each realization of the clicks (rather than in expectation). Thus, we design MAB auctions so as to maximize performance subject to the constraint of (ex-post) incentive-compatibility. Two standard performance measures in the literature are social welfare and total profit.

A well-known general result in Mechanism Design (Myerson, 1981; Archer and Tardos, 2001) implies that an MAB allocation rule can be “extended” to an (ex-post) incentive-compatible MAB mechanism if and only if the MAB allocation rule is (ex-post) monotone. However, naively computing the payments in the “if” direction of this result requires information about the realization that is not revealed during a given run of the mechanism. Devanur and Kakade (2009) and Babaioff et al. (2009) show that incomputability of pay-

---

3. Mechanism Design is a branch of Economics that is concerned with, essentially, the design of auctions. (A *mechanism* is a slightly more general technical term.) A Computer Science take on Mechanism Design (termed *Algorithmic Mechanism Design*) is a design of algorithms whose inputs are provided by self-interested agents, which brings about additional incentive constraints. For background see the book Nisan et al. (2007) and a survey Roughgarden (2008).

ments is a real obstacle. They show that ex-post truthfulness implies severe limitations on the structure of a deterministic MAB allocation rule: essentially, it must separate exploration and exploitation. For the natural example of stochastic clicks it leads to very suboptimal performance (compared to arbitrary MAB allocation rules), both in terms of social welfare (Babaioff et al., 2009) and in terms of profit (Devanur and Kakade, 2009).

Babaioff et al. (2010) overcome this obstacle for *randomized* MAB mechanisms: they provide a general procedure to transform any (ex-post) monotone MAB mechanism to a randomized (ex-post) incentive-compatible MAB mechanism with a very minor loss in performance.<sup>4</sup> This result motivates Problem 1.

---

4. The downside of this transformation is a relatively high variance in payments, see Babaioff et al. (2010) for further discussion. However, a very recent result of Sivan and Wilkens (2010) shows that this amount of variance is essentially optimal for any such general transformation.

SLIVKINS