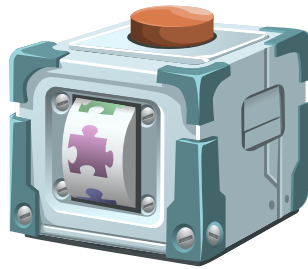


INRIA GRENOBLE-ALPES,

# Online Learning Lecture Notes



*Pierre Gaillard*

2024

## Abstract

These notes explore fundamental ideas in online learning, where data are processed in real-time, and algorithms are updated dynamically. Topics include online linear and convex optimization, as well as multi-armed bandits. The main algorithms in the field will be introduced, and we will delve into regret minimization concepts for theoretical analysis. Online learning algorithms play a central role in recent advancements in reinforcement learning.

## Useful information

- **Pierre Gaillard** (INRIA Grenoble-Alpes)
- Email: pierre.gaillard@inria.fr
- Relevant references: [Cesa-Bianchi and Lugosi \[2006\]](#), [Shalev-Shwartz et al. \[2012\]](#), [Hazan et al. \[2016\]](#), [Lattimore and Szepesvári \[2020\]](#), [Orabona \[2019\]](#)
- Content of the class: mostly theoretical (algorithms and proofs), sequential learning with adversarial data, stochastic bandits, adversarial bandits

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Setting of online learning . . . . .	5
1.2	How to measure the performance: the regret . . . . .	6
<b>2</b>	<b>Online Linear Optimization</b>	<b>8</b>
2.1	Simplex decision set . . . . .	8
2.1.1	The Exponentially Weighed Average forecaster . . . . .	8
2.1.2	Application to prediction with expert advice . . . . .	12
2.1.3	Non-convexity: Linearizing the loss through randomization . . . . .	14
2.2	Convex and compact decision set . . . . .	15
2.2.1	Online Gradient Descent . . . . .	16
2.2.2	Follow The Regularized Leader . . . . .	17
2.2.3	Online Mirror Descent . . . . .	19
<b>3</b>	<b>Online Convex Optimization</b>	<b>23</b>
3.1	The Gradient Trick (from linear to convex losses) . . . . .	23
3.2	Faster rate for losses with curvature (ONS, Non-linear Ridge, EWA) . . . . .	24
3.2.1	Online Newton Step (ONS) . . . . .	25
3.2.2	Non-linear Ridge Regression . . . . .	26
3.2.3	Variants of Exponential Weights for continuous action spaces . . . . .	27
<b>4</b>	<b>Adversarial Bandits</b>	<b>33</b>
4.1	Adversarial multi-armed bandits . . . . .	33
4.1.1	Pseudo-regret bound . . . . .	33
4.1.2	High probability bound on the regret . . . . .	36
4.2	Contextual adversarial multi-armed bandits . . . . .	40
4.2.1	Adversarial multi-armed bandits with experts . . . . .	40
4.2.2	Adversarial multi-armed bandits with side information . . . . .	42
4.3	Online convex optimization with bandit feedback . . . . .	43
4.3.1	Estimating gradients from value observations . . . . .	43
4.3.2	OGD with bandit feedback . . . . .	45
<b>5</b>	<b>Stochastic Multi-armed Bandits</b>	<b>48</b>
5.1	Setting: stochastic bandit with finitely many actions . . . . .	48
5.2	Explore-Then-Commit (ETC) . . . . .	51
5.3	Upper-Confidence-Bound (UCB) . . . . .	53
5.4	Other algorithms . . . . .	56
5.4.1	$\epsilon$ -greedy . . . . .	56
5.4.2	Thompson Sampling . . . . .	56
5.5	Lower bounds for multi-armed bandit . . . . .	56
5.5.1	Distribution-free lower bound . . . . .	56

5.5.2	Distribution-dependent lower bound . . . . .	57
<b>6</b>	<b>Stochastic Contextual bandits</b>	<b>58</b>
6.1	Continuous stochastic bandits . . . . .	58
6.1.1	Contextual bandits through discretization . . . . .	59
6.1.2	Stochastic Linear bandits . . . . .	61
6.2	Other possible extensions of bandits . . . . .	66

# 1 Introduction

In many applications, the data set is not available from the beginning to learn a model but it is observed sequentially as a flow of data. Furthermore, the environment may be so complex that it is unfeasible to choose a comprehensive model and use classical statistical theory and optimization. A classic example is the spam detection which can be seen as a game between spammer and spam filters. Each trying to fool the other one. Another example, is the prediction of processes that depend on human behaviors such as the electricity consumption. These problems are often not adversarial games but cannot be modeled easily and are surely not i.i.d. There is a necessity to take a robust approach by using a method that learns as ones goes along, learning from experiences as more aspects of the data and the problem are observed. This is the goal of online learning. The curious reader can know more about online learning in the books [Cesa-Bianchi and Lugosi \[2006\]](#), [Hazan et al. \[2016\]](#), [Shalev-Shwartz et al. \[2012\]](#).

## 1.1 Setting of online learning

In online learning, a player sequentially makes decisions based on past observations. After committing the decision, the player suffers a loss (or receives a reward depending on the problem). Every possible decision incurs a (possibly different) loss. The losses are unknown to the player beforehand and may be arbitrarily chosen by some adversary. More formally, an online learning problem can be formalized as in Figure 1.1.

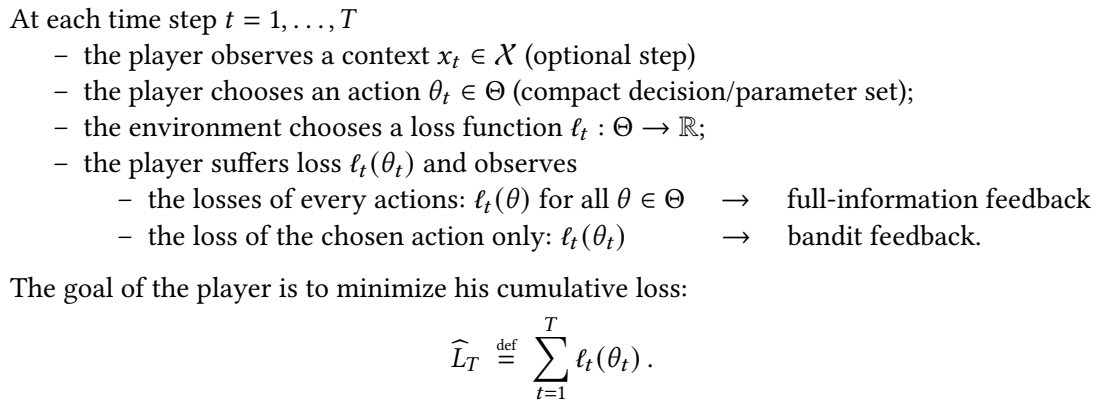


Figure 1.1: Setting of an online learning problem/online convex optimization

**Example 1.1** (Multi-armed bandit). *In  $K$ -armed bandit, the decision set are  $K$  actions (or arms)  $\Theta = \{1, \dots, K\}$  and the player only observes the performance of the chosen action (bandit feedback). In this problem, there is an exploration-exploitation trade-off: the player wants to select the best arm as often as possible but he also needs to explore all arms to estimate their performance.*

*This problem takes his name from slot machines (also known as one-armed bandits because they were originally operated by one lever on the side of the machine) in which some player explores several slot*

machines and tries to maximize his cumulative gain (or more likely minimize his loss!).

Originally, multi-armed bandit setting was introduced by Thompson in 1933 and motivated by clinical trials. For the  $t$ -th patient in some clinical study, one needs to choose the treatment to assign to this patient and observe the response. The goal is to maximize the number of patients healed during the study.

Nowadays, multi-armed bandit is motivated by many applications coming from internet (recommender systems, online advertisements,...). We will see more on multi-armed bandit in next lectures.

**Example 1.2** (Prediction with expert advice). In prediction with expert advice, there is some sequence of observations  $y_1, \dots, y_T \in [0, 1]$  to be predicted step by step with the help of expert forecasts. The setting can be formalized as follows: at each time step  $t \geq 1$

- the environment reveals experts forecasts  $x_t(k)$  for  $k = 1, \dots, d$
- the player chooses a weight vector  $p_t \in \Delta_d \stackrel{\text{def}}{=} \{p \in [0, 1]^d : \sum_{k=1}^d p_k = 1\}$   
(here  $\theta_t$  is denoted  $p_t$  and  $\Theta = \Delta_d$ )
- the player forecasts  $\hat{y}_t = \sum_{k=1}^d p_t(k)x_t(k)$
- the environment reveals  $y_t \in [0, 1]$  and the player suffers loss  $\ell_t(p_t) = \ell(\hat{y}_t, y_t)$  where  $\ell : [0, 1]^2 \rightarrow [0, 1]$  is a loss function.

Considering  $\Theta := \Delta_d$  and  $\theta_t := p_t$ , this setting can be recovered by the online learning setting of Figure 1.1. The inputs correspond to the expert advice  $x_t(k)$  that are often revealed before the learner makes his decision  $p_t$ .

Player's performance is then measured via a loss function  $\ell_t(p_t) = \ell(\hat{y}_t, y_t)$  which measures the distance between the prediction  $\hat{y}_t$  and the output  $y_t$ . Typical loss functions are the squared loss  $\ell(\hat{y}_t, y_t) = (\hat{y}_t - y_t)^2$ , the absolute loss  $\ell(\hat{y}_t, y_t) = |\hat{y}_t - y_t|$  or the absolute percentage of error  $\ell(\hat{y}_t, y_t) = |\hat{y}_t - y_t|/|y_t|$ . All these loss functions are convex, which will play an important role in the analysis.

## 1.2 How to measure the performance: the regret

Of course, if the environment chooses large losses  $\ell_t(x)$  for all decisions  $\theta \in \Theta$ , it is impossible for the player to ensure small cumulative loss. Therefore, one needs a relative criterion: the regret of the player is the difference between the cumulative loss he incurred and that of the best fixed decision in hindsight.

### Definition 1.1

### Regret

The regret of the player with respect to a fixed parameter  $\theta^* \in \Theta$  after  $T$  time steps is

$$R_T(\theta^*) \stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(\theta_t) - \sum_{t=1}^T \ell_t(\theta^*).$$

The regret (or uniform regret) is defined as  $R_T \stackrel{\text{def}}{=} \sup_{\theta^* \in \Theta} R_T(\theta^*)$ .

We have some bias-variance decomposition:

$$\sum_{t=1}^T \ell_t(\theta_t) = \underbrace{\inf_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta)}_{\text{Approximation error = how good the possible actions are.}} + \underbrace{R_T}_{\text{Sequential estimation error of the best action}}$$

We will focus on the regret in these lectures. The goal of the player is to ensure a sublinear regret  $R_T = o(T)$  as  $T \rightarrow \infty$  and this for any possible sequence of losses  $\ell_1, \dots, \ell_T$ . In this case, the average performance of the player will approach on the long term the one of the best decision.

**Remarks** Let us make some remarks:

- Except in the stochastic bandit part, we will not make any random assumption on the process generating the losses  $\ell_t$ . The latter are deterministic and may be chosen by some adversary. Typically, the problem can be seen as a game between the player who aims at optimizing with respect to  $\theta_1, \dots, \theta_T$  against an environment who aims at maximizing with respect to  $\ell_t, \dots, \ell_T$  and  $\theta^*$ . Player's goal is to approach the quantity:

$$\inf_{\theta_1} \sup_{\ell_1} \inf_{\theta_2} \sup_{\ell_2} \dots \inf_{\theta_T} \sup_{\ell_T} \sup_{\theta^* \in \Theta} R_T(\theta^*).$$

- Note that the loss functions  $\ell_t$  depend on the round  $t$ . This may be caused by many phenomena. We provide here some possible reasons. This may be because
  - of some observation to be predicted if  $\ell_t(x) = \ell(x, y_t)$ . For instance, if the goal is to predict the evolution of the temperature  $y_1, \dots, y_T$ , the latter changes over time and a prediction  $x$  is evaluated with  $\ell_t(x) = (x - y_t)^2$ .
  - the environment is stochastic and the variation over time  $t$  models some noise effect.
  - of a changing environment. For instance, if the player is playing a game against some adversary that evolves and adapts to its strategy. A typical example is the case of spam detections. If the player tries to detect spams, while some spammers (the environment) try at the same time to fool the player with new spam strategies.

**Exercise 1.1.** Instead considering the regret with respect to a fixed  $\theta^* \in \Theta$ , one would be tempted to minimize the quantity

$$R_T^* \stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(\theta_t) - \sum_{t=1}^T \inf_{\theta \in \Theta} \ell_t(\theta)$$

where the infimum is inside the sum. Show that the environment can ensure  $R_T^*$  to be linear in  $T$  by choosing properly the loss functions  $\ell_t$ .

## 2 Online Linear Optimization

In this part, we assume that  $\Theta \subset \mathbb{R}^d$  and that the loss functions  $\ell_t : \Theta \rightarrow \mathbb{R}$  are linear

$$\forall \theta \in \Theta, \quad \ell_t(\theta) = \langle \theta, g_t \rangle \quad (2.1)$$

where  $g_t \in \mathbb{R}^d$  is a loss vector chosen by the environment at round  $t$  and which is revealed to the player at the end of the round. This setting might seem restrictive but we will see later that it can easily be generalized to more complex frameworks.

### 2.1 Simplex decision set

Here, we start by presenting an algorithm when the decision set  $\Theta$  is the  $d$ -dimensional simplex

$$\Delta_d \stackrel{\text{def}}{=} \left\{ p \in [0, 1]^d : \sum_{k=1}^d p_k = 1 \right\}.$$

Since the decisions  $\theta_t$  are probability distributions over  $[d] \stackrel{\text{def}}{=} \{1, \dots, d\}$ , in this part we will denote them by  $p_t$  instead of  $\theta_t$ . The simplex is a versatile decision set that includes distributions, enables weighted averages for method combination, and can represent any convex hull, making it powerful.

#### 2.1.1 The Exponentially Weighted Average forecaster

At round  $t$  the player needs to choose a weight vector  $p_t \in \Delta_d$ . The question is how to choose it? The idea is to give more weight to actions that performed well in the past. But we should not give all the weight to the current best action, otherwise it would not work (see exercises). The exponentially weighted average forecaster (EWA) also called Hedge performs this trade-off by choosing a weight that decreases exponentially fast with the past errors.

Parameter:  $\eta > 0, p_1 \in \Delta_d$

For  $t = 1, \dots, T$

- select  $p_t$ ; incur loss  $\ell_t(p_t) = \langle p_t, g_t \rangle$  and observe  $g_t \in \mathbb{R}^d$ ;
- update for all  $k \in \{1, \dots, d\}$

$$p_{t+1}(k) = \frac{p_1(k) e^{-\eta \sum_{s=1}^t g_s(k)}}{\sum_{j=1}^d p_1(j) e^{-\eta \sum_{s=1}^t g_s(j)}}.$$

**Algorithm 2.1:** The Exponentially Weighted Average forecaster (EWA)

**Exercise 2.1.** Consider the strategy, called “Follow The Leader” (FTL) that puts all the mass on the best action so far:

$$p_t \in \arg \min_{p \in \Theta} \sum_{s=1}^{t-1} \ell_s(p). \quad (\text{FTL})$$



1. Show that  $p_t(k) > 0$  implies that  $k \in \arg \min_j \sum_{s=1}^{t-1} g_s(j)$
2. Show that the regret of FTL might be linear: i.e., there exists a sequence  $g_1, \dots, g_T \in [-1, 1]^d$  such that  $R_T \geq (1 - 1/d)T$ .

The following theorem proves that EWA, which is a smoothed version of FTL (it performs a soft-max), achieves sublinear regret.

### Theorem 2.1

Let  $T \geq 1$ . For all sequences of loss vectors  $(g_t) \in \mathbb{R}^d$  with  $\|g_t\|_\infty \leq G_\infty$ , if  $\eta \leq G_\infty^{-1}$ , and  $p_1 = (1/d, \dots, 1/d)$ , EWA achieves the regret upper bound

$$R_T \stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(p_t) - \min_{p \in \Delta_d} \sum_{t=1}^T \ell_t(p) \leq \eta \sum_{t=1}^T \sum_{k=1}^d p_t(k) g_t(k)^2 + \frac{\log d}{\eta}, \quad (2.2)$$

where we recall  $\ell_t : p \in \Delta_d \mapsto \langle p, g_t \rangle$ . Therefore, for the choice  $\eta = \frac{1}{G_\infty} \sqrt{\frac{\log d}{T}}$ , EWA satisfies the regret bound  $R_T \leq 2G_\infty \sqrt{T \log d}$ .

This regret bound is optimal up to multiplicative constants (see [Cesa-Bianchi and Lugosi \[2006\]](#)).

*Proof.* First, we remark that if the losses are linear  $\ell_t(p) = \langle p, g_t \rangle$ , then

$$\min_{p \in \Delta_d} \sum_{t=1}^T \ell_t(p) = \min_{1 \leq k \leq d} \sum_{t=1}^T g_t(k).$$

Indeed, for any  $p \in \Delta_d$ ,

$$\sum_{t=1}^T \ell_t(p) = \sum_{t=1}^T \langle p, g_t \rangle = \sum_{k=1}^d p(k) \sum_{t=1}^T g_t(k) \geq \sum_{k=1}^d p(k) \min_{1 \leq k \leq d} \sum_{t=1}^T g_t(k) = \min_{1 \leq k \leq d} \sum_{t=1}^T g_t(k).$$

Therefore

$$R_T = \sum_{t=1}^T \langle p_t, g_t \rangle - \min_{1 \leq k \leq d} \sum_{t=1}^T g_t(k).$$

We denote  $W_t(j) = e^{-\eta \sum_{s=1}^t g_s(j)}$  and  $W_t = \sum_{j=1}^d W_t(j)$ . The proof will consist in upper-bounding and lower-bounding  $W_T$ . We have

$$\begin{aligned} W_t &= \sum_{j=1}^d W_{t-1}(j) e^{-\eta g_t(j)} && \leftarrow W_t^{(j)} = W_{t-1}(j) e^{-\eta g_t(j)} \\ &= W_{t-1} \sum_{j=1}^d \frac{W_{t-1}(j)}{W_{t-1}} e^{-\eta g_t(j)} \\ &= W_{t-1} \sum_{j=1}^d p_t(j) e^{-\eta g_t(j)} && \leftarrow p_t(j) = \frac{e^{-\eta \sum_{s=1}^{t-1} g_s(j)}}{\sum_{k=1}^d e^{-\eta \sum_{s=1}^{t-1} g_s(k)}} = \frac{W_{t-1}(j)}{W_{t-1}} \\ &\leq W_{t-1} \sum_{j=1}^d p_t(j) (1 - \eta g_t(j) + \eta^2 g_t(j)^2) && \leftarrow e^x \leq 1 + x + x^2 \text{ for } x \leq 1 \end{aligned}$$

$$= W_{t-1}(1 - \eta \langle p_t, g_t \rangle + \eta^2 \langle p_t, g_t^2 \rangle),$$

where we used in the inequality  $-\eta g_t(j) \leq -g_t(j)/G_\infty \leq 1$  and where we denote  $g_t = (g_t(1), \dots, g_t(d))$ ,  $g_t^2 = (g_t(1)^2, \dots, g_t(d)^2)$  and  $p_t = (p_t(1), \dots, p_t(d))$ . Now, using  $1 + x \leq e^x$ , we get:

$$W_t \leq W_{t-1} \exp(-\eta \langle p_t, g_t \rangle + \eta^2 \langle p_t, g_t^2 \rangle).$$

By induction on  $t = 1, \dots, T$ , this yields using  $W_0 = d$

$$W_T \leq d \exp\left(-\eta \sum_{t=1}^T \langle p_t, g_t \rangle + \eta^2 \sum_{t=1}^T \langle p_t, g_t^2 \rangle\right). \quad (2.3)$$

On the other hand, upper-bounding the maximum with the sum,

$$\exp\left(-\eta \min_{j \in [d]} \sum_{t=1}^T g_t(j)\right) \leq \sum_{j=1}^d \exp\left(-\eta \sum_{t=1}^T g_t(j)\right) \leq W_T.$$

Combining the above inequality with Inequality (2.3) and taking the log, we get

$$-\eta \min_{j \in [d]} \sum_{t=1}^T g_t(j) \leq -\eta \sum_{t=1}^T \langle p_t, g_t \rangle + \eta^2 \sum_{t=1}^T \langle p_t, g_t^2 \rangle + \log d. \quad (2.4)$$

Dividing by  $\eta$  and reorganizing the terms proves the first inequality:

$$R_T = \sum_{t=1}^T \langle p_t, g_t \rangle - \min_{1 \leq j \leq d} \sum_{t=1}^T g_t(j) \leq \eta \sum_{t=1}^T \langle p_t, g_t^2 \rangle + \frac{\log d}{\eta}$$

Optimizing  $\eta$  and upper-bounding  $\langle p_t, g_t^2 \rangle \leq G_\infty^2$  concludes the second inequality.  $\square$

**Exercise 2.2.** Using Hoeffding's lemma (Lem. 5.1), show that

$$\sum_{j=1}^d p_t(j) e^{-\eta g_t(j)} \leq e^{\frac{\eta^2 G_\infty^2}{8} - \langle p, g_t \rangle}.$$

Deduce that for a well-chosen  $\eta$ , the regret upper bound of EWA can be improved to

$$R_T \leq G_\infty \sqrt{\frac{T}{2} \log d}.$$

**Anytime algorithm (the doubling trick)** The previous algorithm EWA depends on a parameter  $\eta > 0$  that needs to be optimized according to  $d$ ,  $G_\infty$ , and  $T$ . For instance, for EWA using the value

$$\eta = \frac{1}{G_\infty} \sqrt{\frac{\log d}{T}}.$$

the bound of Theorem 2.1 is only valid for horizon  $T$ . However, the learner might not know  $G_\infty$  or the time horizon  $T$  in advance and one might want an algorithm with guarantees valid simultaneously for all  $T \geq 1$ . We can avoid the assumption that  $T$  is known in advance, at the cost of a constant factor, by using the so-called *doubling trick*. The general idea is the following. Whenever we reach a time step  $t$  which is a power of 2, we restart the algorithm (forgetting all the information gained in the past) setting  $\eta$  to  $G_\infty^{-1} \sqrt{\log d/t}$ . Let us denote EWA-doubling this algorithm.

### Theorem 2.2

### Anytime bound on the regret

For all  $T \geq 1$ , the pseudo-regret of EWA-doubling is then upper-bounded as:

$$R_T \leq 7G_\infty \sqrt{T \log d}.$$

The same trick can be used to turn most online algorithms into anytime algorithms (even in more general settings: bandits, general loss,...). We can use the *doubling trick* whenever we have an algorithm with a regret of order  $O(T^\alpha)$  for some  $\alpha > 0$  with a known horizon  $T$  to turn it into an algorithm with a regret  $O(T^\alpha)$  for all  $T \geq 1$ .

Another solution is to use time-varying parameters  $\eta_t$  replacing  $T$  with the current value of  $t$ . The analysis is however less straightforward.

**Exercise 2.3.** Prove a regret bound for the time-varying choice  $\eta_t = \sqrt{\log d / (1 + \sum_{s=1}^t \|g_s\|_\infty^2)}$  in EWA.

*Proof of Theorem 2.2.* For simplicity we assume  $T = 2^{M+1} - 1$ . The regret of EWA-doubling is then upper-bounded as:

$$\begin{aligned} R_T &= \sum_{t=1}^T \ell_t(p_t) - \min_{p \in \Delta_d} \sum_{t=1}^T \ell_t(p) \\ &\leq \sum_{t=1}^T \ell_t(p_t) - \sum_{m=0}^M \min_{p \in \Delta_d} \sum_{t=2^m}^{2^{m+1}-1} \ell_t(p) \\ &= \underbrace{\sum_{m=0}^M \sum_{t=2^m}^{2^{m+1}-1} \ell_t(p_t) - \min_{p \in \Delta_d} \sum_{t=2^m}^{2^{m+1}-1} \ell_t(p)}_{R_m}. \end{aligned}$$

Now, we remark that each term  $R_m$  corresponds to the expected regret of an instance of EWA over the  $2^m$  rounds  $t = 2^m, \dots, 2^{m+1} - 1$  and run with the optimal parameter  $\eta = \sqrt{\log d / 2^m}$ . Therefore, using Theorem 2.1, we get  $R_m \leq 2\sqrt{2^m \log d}$ , which yields:

$$R_T \leq \sum_{m=0}^M 2G_\infty \sqrt{2^m \log d} \leq 2(1 + \sqrt{2})G_\infty \sqrt{2^{M+1} \log d} \leq 7G_\infty \sqrt{T \log d}.$$

□

**Improvement for small losses** The first inequality in Theorem 2.1 is sometimes called improvement for small losses when losses are non-negative. Let's define  $\widehat{L}_T \stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(p_t)$  the loss of the algorithm and  $L_T^* \stackrel{\text{def}}{=} \min_{p \in \Delta_d} \sum_{t=1}^T \ell_t(p)$  the optimal loss. Then, optimizing in  $\eta = (\log(d) / \sum_{t=1}^T \langle p_t, g_t^2 \rangle)^{1/2}$ , the regret is upper-bounded by

$$R_T \stackrel{\text{def}}{=} \widehat{L}_T - L_T^* \leq \frac{\log d}{\eta} + \eta \sum_{t=1}^T \langle p_t, g_t^2 \rangle = 2\sqrt{(\log d) \sum_{t=1}^T \langle p_t, g_t^2 \rangle} \leq 2\sqrt{(\log d) G_\infty \widehat{L}_T}.$$

Therefore, using that  $x^2 \leq a + cx$  implies  $x \leq \sqrt{a} + c$  when  $a, c \geq 0$ , we get with  $x = \sqrt{\widehat{L}_T}$  that

$$\sqrt{\widehat{L}_T} \leq \sqrt{L_T^*} + 2\sqrt{G_\infty \log d},$$

which yields

$$R_T \leq 4\sqrt{G_\infty \log(d)L_T^*} + 4G_\infty \log d,$$

which is small whenever the optimal loss  $L_T^*$  is small.

### 2.1.2 Application to prediction with expert advice

The preceding section considers linear loss functions. Yet, it can yield non-trivial regret bounds for general convex losses. We consider here an application to the setting of prediction with expert advice detailed in Example 1.2. The goal is to minimize the regret with respect to the best expert

$$R_T^{\text{expert}} \stackrel{\text{def}}{=} \sum_{t=1}^T \ell(\widehat{y}_t, y_t) - \min_{1 \leq k \leq d} \sum_{t=1}^T \ell(x_t(k), y_t),$$

where  $\widehat{y}_t = \langle p_t, x_t \rangle$  are the predictions of the algorithm and  $y_t$  the observations to be predicted sequentially.

**Convex loss function** We state bellow a corollary to Theorem 2.1 when the loss functions  $\ell(\cdot, \cdot)$  are convex in their first argument.

#### Corollary 2.3 Regret of EWA for prediction with expert advice and convex loss

Let  $T \geq 1, B > 0$ . Assume that the loss function  $\ell : (x, y) \in \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$  is convex and takes values in  $[-B, B]$ . Then, EWA applied with the vector vectors  $g_t = (\ell(x_t(1), y_t), \dots, \ell(x_t(d), y_t)) \in [-B, B]^d$  has a regret upper-bounded by

$$R_T^{\text{expert}} \leq 2B\sqrt{T \log d}$$

where  $\widehat{y}_t = \langle p_t, x_t \rangle$  and were  $\eta > 0$  is well-tuned.

Therefore, the average error of the algorithm will converge to the average error of the best expert. This is the case for the square loss, the absolute loss or the absolute percentage of error.

*Proof.* It suffices to remark that by convexity of  $\ell(\cdot, \cdot)$  in its first argument

$$\begin{aligned} R_T^{\text{expert}} &= \sum_{t=1}^T \ell(\langle p_t, x_t \rangle, y_t) - \min_{1 \leq k \leq d} \sum_{t=1}^T \ell(x_t(k), y_t) \\ &\leq \sum_{t=1}^T \langle p_t, g_t \rangle - \min_{1 \leq k \leq d} \sum_{t=1}^T g_t(k) \stackrel{\text{def}}{=} R_T. \end{aligned}$$

The result is then obtained by Theorem 2.1. □

**Exp-concave loss function** Here, we show that a faster rate can be obtained (with EWA) if the loss function are exp-concave.

**Definition 2.1** **$\eta$ -exp-concavity**

For  $\eta \in \mathbb{R}$ , a function  $f$  is said to be  $\eta$ -exp-concave if  $x \mapsto e^{-\eta f(x)}$  is concave.

Exp-concavity is stronger than convexity but weaker than strong convexity. Indeed, exp-concave functions are convex because  $-\log$  is convex and decreasing. Furthermore, any  $\eta$ -exp-concave function is also  $\eta'$ -exp-concave for  $0 \leq \eta' \leq \eta$ .

In prediction with expert advice, if the loss are generated from a fixed loss function  $\ell_t(p) = \ell(\langle p, x_t \rangle, y_t)$ , then  $\ell_t$  are  $\eta$ -exp-concave if  $\hat{y} \mapsto \ell(\hat{y}, y_t)$  are  $\eta$ -exp-concave for all  $y_t$ . We can compute  $\eta$  for some common loss functions:

- the squared loss:  $\ell : (\hat{y}, y) \in [0, 1]^2 \mapsto (\hat{y} - y)^2$ , then  $\ell_t$  are  $1/2$ -exp-concave. Indeed, let  $y \in [0, 1]$  and denote  $G : \hat{y} \mapsto \exp(-\eta(\hat{y} - y)^2)$ . Then,  $G''(\hat{y}) = G(\hat{y})(4\eta^2(\hat{y} - y)^2 - 2\eta)$ . Thus  $G$  is concave if and only if  $(\hat{y} - y)^2 \leq 1/(2\eta)$  which is satisfied for  $\eta = 1/2$ . This is also the case in higher dimensions with  $\ell(\hat{y}, y) = \|\hat{y} - y\|^2$ . If the observations and prediction  $\hat{y}, y \in [0, B]$ , then the  $\ell_t$  are  $1/(2B^2)$ -exp-concave
- the relative entropy (or Kullback–Leibler divergence):  $\ell : (\hat{y}, y) \in [0, 1]^2 \mapsto y \log(y/\hat{y}) - (1 - y) \log((1 - y)/(1 - \hat{y}))$ . Then the functions  $\ell_t$  are 1-exp-concave. This loss can for instance used for density estimation of the sequence  $y_1, \dots, y_T$ .
- the linear loss  $\ell(\hat{y}, y) = \hat{y} \cdot y$ , the absolute loss  $\ell(\hat{y}, y) = |\hat{y} - y|$  or the absolute percentage of error are however not  $\eta$ -exp-concave for any  $\eta > 0$ .

**Corollary 2.4      Regret of EWA for prediction with expert advice and exp-concave loss**

In the setting of prediction with expert advice, if the loss functions  $\ell(\cdot, y_t)$  are  $\eta$ -exp-concave for all  $y_t$ , then EWA run with vectors  $g_t = (\ell(x_t(1), y_t), \dots, \ell(x_t(d), y_t)) \in \mathbb{R}^d$  with parameter  $\eta > 0$  and  $p_1 = (1/d, \dots, 1/d)$  satisfies

$$R_T^{\text{expert}} \leq \frac{\log d}{\eta},$$

for all  $T \geq 1$ .

The worst-case regret does not increase with  $T$  but grows logarithmically in the dimension  $d$ .

*Proof.* The proof is similar to the original proof of EWA. We define  $W_t(i) = e^{-\eta \sum_{s=1}^t g_s(i)}$  and  $W_t = \sum_{i=1}^d W_t(i)$ . We have

$$\begin{aligned} W_t &= \sum_{j=1}^N W_{t-1}(j) e^{-\eta g_t(j)} && \leftarrow W_t(j) = W_{t-1}(j) e^{-\eta g_t(j)} \\ &= W_{t-1} \sum_{j=1}^N \frac{W_{t-1}(j)}{W_{t-1}} e^{-\eta g_t(j)} \\ &= W_{t-1} \sum_{j=1}^N p_t(j) e^{-\eta g_t(j)} && \leftarrow p_t(j) = \frac{e^{-\eta \sum_{s=1}^{t-1} g_s(j)}}{\sum_{k=1}^N e^{-\eta \sum_{s=1}^{t-1} g_s(k)}} = \frac{W_{t-1}(j)}{W_{t-1}} \\ &\leq W_{t-1} \exp(-\eta \ell(\langle p_t, x_t \rangle, y_t)) && \leftarrow \eta\text{-exp-concavity} \end{aligned}$$

Now, by induction on  $t = 1, \dots, T$ , this yields using  $W_0 = d$

$$W_T \leq d \exp\left(-\eta \sum_{t=1}^T \ell(\hat{y}_t, y_t)\right). \quad (2.5)$$

On the other hand, upper-bounding the maximum with the sum,

$$\exp\left(-\eta \min_{j \in [d]} \sum_{t=1}^T g_t(j)\right) \leq \sum_{j=1}^d \exp\left(-\eta \sum_{t=1}^T g_t(j)\right) \leq W_T.$$

Combining the above inequality with Inequality (2.5) and taking the log concludes the proof.  $\square$

### 2.1.3 Non-convexity: Linearizing the loss through randomization

**Setting:**  $\Theta$  finite, non-convex loss functions  $\ell_t : \Theta \rightarrow [-B, B]$ .

In this section, we consider a *finite set of decision*  $\Theta = \{1, \dots, d\}$  and we assume that the player is restricted to play an action in  $\Theta$ . In other words, the player cannot play convex combinations of the actions as it was done for prediction with expert advice. For instance, we may want to build a recommender system to recommend movies to customers. The loss function are *arbitrary bounded loss functions*  $\ell_t : \Theta \rightarrow [-B, B]$ .

**Need of a random strategy** The following proposition shows that the choice  $\theta_t$  cannot be deterministic in this setting. Otherwise, the adversary may fool the player by taking  $\ell_t$  depending on  $\theta_t$ .

#### Proposition 2.5

*Any deterministic algorithm may incur a linear regret. In other words, we can find some sequence of losses  $\ell_t$  such that  $R_T \gtrsim T$ .*

*Proof.* Since  $\theta_t$  is deterministic, the loss function  $\ell_t$  can depend on  $\theta_t$ . We then choose  $\ell_t(\theta_t) = 1$  and  $\ell_t(\theta) = 0$  for  $\theta \neq \theta_t$ . Then one of the chosen actions was picked less than  $T/d$  times so that  $\max_{1 \leq k \leq d} \ell_t(k) \leq T/d$ . Therefore,  $R_T \geq (1 - 1/d)T$ .  $\square$

From the above proposition, we see that the strategy of the learner needs to be random. Therefore, instead of choosing an action in  $\{1, \dots, d\}$ , the player chooses a probability distribution  $p_t \in \Delta_d := \{p \in [0, 1]^d : \sum_k p_k = 1\}$  and draws  $\theta_t \sim p_t$ . And we recover the setting with actions played in the simplex  $\Delta_d$ .

**A random regret** The regret  $R_T$  will be here a random quantity that depends on the randomness of the algorithm (and eventually of the data). We will thus focus on upper-bounding the regret:

- with high-probability:  $R_T \leq \varepsilon$  with probability at least  $1 - \delta$ ;
- in expectation:  $\mathbb{E}[R_T] \leq \varepsilon$ .

*From high-probability bound to expected bound.* Note that since the losses are bounded in  $[0, 1]$  a bound in high probability entails a bound in expectation. If  $R_T \leq \varepsilon$  with probability at least  $1 - \delta$ , then

$$\mathbb{E}[R_T] \leq \mathbb{E}[R_T | R_T \leq \varepsilon] \mathbb{P}(R_T \leq \varepsilon) + \mathbb{E}[R_T | R_T \geq \varepsilon] \mathbb{P}(R_T \geq \varepsilon) \leq \varepsilon + T\delta. \quad (2.6)$$

Another useful (and often better) tool to transform a high-probability bound into a bound in expectation is the inequality  $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq \varepsilon) d\varepsilon$  for nonnegative random variable  $X$ .

From expected bound to high-probability bound. On the other hand, since the losses are bounded, using Hoeffding's inequality a bound in expectation entails a bound in high probability at the cost of an additive term of order  $\sqrt{T \log(1/\delta)}$  in the regret.

### Proposition 2.6

Let  $\Theta$  be finite of cardinal  $d$  and  $\ell_1, \dots, \ell_T : \Theta \rightarrow [-B, B]$  be an arbitrary sequence of losses. Then, applying EWA with a well-chosen  $\eta$  and  $p_1 = (1/d, \dots, 1/d)$  and sampling  $\theta_t \sim p_t$  at every round, satisfies the expected regret

$$\mathbb{E}[R_T] = \mathbb{E} \left[ \sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta) \right] \leq 2B\sqrt{T \log d}$$

for  $\eta$  well tuned.

**Exercise 2.4.** Using Hoeffding's inequality, provide a bound on the regret  $R_T$  with probability  $1 - \delta$ .

*Proof.* Using  $g_t = (\ell_t(1), \dots, \ell_t(d)) \in [-B, B]^d$ , from Theorem 1 of last class, we have

$$\sum_{t=1}^T \langle p_t, g_t \rangle - \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta) \leq 2B\sqrt{T \log d}.$$

It suffices then to take the expectation and remark that

$$\mathbb{E}[\ell_t(\theta_t)] = \mathbb{E}[\mathbb{E}[\ell_t(\theta_t) | p_t]] = \mathbb{E}[\langle p_t, g_t \rangle].$$

□

It is worth pointing out that we did not make any assumption on the loss function  $\ell_t$  beside boundedness. In particular, it can be non-convex.

**Example 2.1** (Online classification). Assume that you may want to predict a sequence of labels  $y_1, \dots, y_T \in \{0, 1\}$  (such as spams) based on expert advice  $x_t(k) \in \{0, 1\}$  (such as different spam detectors). Then, using the losses  $\ell_t(k) = \mathbb{1}_{x_t(k) \neq y_t}$ , EWA ensures

$$\mathbb{E} \left[ \sum_{t=1}^T \mathbb{1}_{\theta_t \neq y_t} - \min_{1 \leq k \leq d} \sum_{t=1}^T \mathbb{1}_{x_t(k) \neq y_t} \right] \leq 2\sqrt{T \log d}.$$

Hence, the expected number of mistakes of the algorithms will not be much larger than the one of the best expert. This is valid though the loss function is nonconvex.

## 2.2 Convex and compact decision set

**Setting:** linear loss function, convex and compact decision set  $\Theta$ .

In this section, we generalize the preceding sections to any compact convex decision set  $\Theta \subset \mathbb{R}^d$ . We still assume that the loss functions take a linear form

$$\forall \theta \in \Theta, \quad \ell_t(\theta) = \langle \theta, g_t \rangle,$$

for some  $g_t \in \mathbb{R}^d$ . We provide a few well-known algorithms in this setting.

### 2.2.1 Online Gradient Descent

We introduce, Online Gradient Descent, and is due to [Zinkevich \[2003\]](#) in the online learning setting. It is an online variant of the well-known Gradient Descent algorithm in optimization.

Parameter:  $\eta > 0, \theta_1 \in \Theta$   
 For  $t = 1, \dots, T$   
   – select  $\theta_t$ ; incur loss  $\ell_t(\theta_t) = \langle \theta_t, g_t \rangle$  and observe  $g_t \in \mathbb{R}^d$ ;  
   – update  
       
$$\theta_{t+1} = \text{Proj}_\Theta (\theta_t - \eta g_t),$$
  
   where  $\text{Proj}_\Theta$  is the Euclidean projection onto  $\Theta$ .

**Algorithm 2.2:** Online Gradient Descent (OGD) for linear losses  $\ell_t(\theta) = \langle \theta, g_t \rangle$ .

#### Theorem 2.7

Then for any sequence  $g_1, \dots, g_T \in \mathbb{R}^d$ , if the losses are linear  $\ell_t(\theta) = \langle \theta, g_t \rangle$ , the regret of OGD satisfies

$$R_T = \sum_{t=1}^T \ell_t(\theta_t) - \ell_t(\theta^*) \leq \frac{\|\theta^* - \theta_1\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|g_t\|^2,$$

where  $\theta^* \in \arg \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta)$ . In particular, if  $\max_{\theta \in \Theta} \|\theta - \theta_1\| \leq D_2$  and  $\|g_t\| \leq G_2$  for all  $t$ , for  $\eta = \frac{D_2}{G_2 \sqrt{T}}$ , we have  $R_T \leq D_2 G_2 \sqrt{T}$ .

**Exercise 2.5.** Prove an upper-bound on the regret of OGD

- a) when  $\eta$  is calibrated with a doubling trick.
- b) when  $\eta$  is calibrated using a time-varying parameter  $\eta_t$

**Exercise 2.6.** Prove an upper-bound on the regret of OGD with respect to any sequence of points  $\theta_1^*, \dots, \theta_t^* \in \Theta$  such that  $\sum_{t=2}^T \|\theta_t^* - \theta_{t-1}^*\| \leq X$

$$\sum_{t=1}^T \ell_t(\theta_t) - \sum_{t=1}^T \ell_t(\theta_t^*) \leq \dots$$

**Remark.** Assume that  $\Theta = \Delta_d$  is the simplex and the loss functions are of the form  $\ell_t(\theta) = \langle \theta, g_t \rangle$  where  $\|g_t\|_\infty \leq G_\infty$ . Then both EWA and OGD are possible algorithms (see Theorems 2.1 and 2.7). We saw in Theorem 2.1 that EWA has a regret bound  $R_T \leq 2G_\infty \sqrt{T \log d}$ . In this case, for all  $p, p' \in \Delta_d$

$$\|p - p'\| = \sum_{k=1}^d (p(i) - p'(i))^2 \leq \sum_{i=1}^d |p(i) - p'(i)| \leq \sum_{i=1}^d p(i) + p'(i) = 2,$$

and  $\|g_t\| \leq \sqrt{d} \|g_t\|_\infty \leq \sqrt{d} G_\infty$ . Therefore, the regret of OGD is upper-bounded by  $R_T \leq G_\infty \sqrt{2dT}$ . To summarize

$$\text{EWA: } R_T \leq 2G_\infty \sqrt{T \log d} \quad \text{and} \quad \text{OGD: } R_T \leq \sqrt{2dT}.$$

The dependence on  $d$  of OGD is suboptimal in this case. This is solved by OMD, a generalization of both algorithms.



*Proof of Theorem 2.7.* Let  $\theta^* \in \arg \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta)$  and denote  $z_{t+1} = \theta_t - \eta g_t$  so that by definition of  $\theta_{t+1}$  in the algorithm, we have  $\theta_{t+1} = \text{Proj}_{\Theta}(z_{t+1})$ . By convexity, the regret can be upper-bounded as

$$\begin{aligned} R_T &\stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(\theta_t) - \ell_t(\theta^*) = \sum_{t=1}^T \langle g_t, \theta_t - \theta^* \rangle \\ &= \frac{1}{\eta} \sum_{t=1}^T \langle z_{t+1} - \theta_t, \theta_t - \theta^* \rangle \quad \leftarrow g_t = \frac{z_{t+1} - \theta_t}{\eta}. \end{aligned}$$

Then, we use the equality  $\|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2\langle x, y \rangle$  for all  $x, y \in \Theta$  so that

$$\langle x, y \rangle = \frac{\|x\|^2 + \|y\|^2 - \|x - y\|^2}{2}.$$

Applying it with  $x = z_{t+1} - \theta_t$  and  $y = \theta_t - \theta^*$  en substituting into the above regret bound, this yields

$$R_T \leq \frac{1}{2\eta} \sum_{t=1}^T (\|z_{t+1} - \theta_t\|^2 + \|\theta^* - \theta_t\|^2 - \|z_{t+1} - \theta^*\|^2)$$

Then, using  $\|z_{t+1} - \theta_t\| = \eta \|g_t\|$  and  $\|\theta^* - \theta_t\| \leq \|\theta^* - z_t\|$  because  $\Theta$  is convex and  $\theta_t = \text{Proj}_{\Theta}(z_t)$ , we get

$$R_T \leq \frac{1}{2\eta} \sum_{t=1}^T (\eta^2 \|g_t\|^2 + \|\theta^* - z_t\|^2 - \|z_{t+1} - \theta^*\|^2).$$

The last terms telescope, therefore summing over  $t$  concludes the proof

$$R_T \leq \frac{\eta}{2} \sum_{t=1}^T \|g_t\|^2 + \frac{\|\theta^* - \theta_1\|^2}{2\eta}.$$

□

### 2.2.2 Follow The Regularized Leader

In the preceding section, EWA and OGD were introduced. Here, we will present more versatile algorithms that rely on a regularization function  $\psi : \Theta \rightarrow \mathbb{R}$ , appearing to be generalizations of the earlier algorithms. We consider the FTRL algorithm defined in Algorithm 2.3, which depends on a strongly convex, smooth, and twice differentiable regularization function  $\psi : \Theta \rightarrow \mathbb{R}$ .

Input:  $\eta > 0$ , regularization function  $\psi$   
 Let  $\theta_1 = \arg \min_{\theta \in \Theta} \{\psi(\theta)\}$   
 For  $t = 1$  to  $T$  do  
   – Play  $\theta_t$  incur loss  $\ell_t(\theta_t) = \langle \theta_t, g_t \rangle$  and observe  $g_t \in \mathbb{R}^d$   
   – Update  

$$\theta_{t+1} = \arg \min_{\theta \in \Theta} \left\{ \eta \sum_{s=1}^t \langle \theta, g_s \rangle + \psi(\theta) \right\}$$
  
 end for

**Algorithm 2.3:** Follow the Regularized Leader (FTRL)

Before stating the theorem, let us first recall a few basic definitions useful for the analysis.

**Definition 2.2****Strong convexity**

We say that a differentiable function  $f : \Theta \rightarrow \mathbb{R}$  is  $\mu$ -strongly convex with respect to a norm  $\|\cdot\|$  if for all  $\theta, \theta' \in \Theta$

$$f(\theta) \geq f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle + \frac{\mu}{2} \|\theta - \theta'\|^2.$$

**Theorem 2.8****FTRL regret upper-bound**

Let  $T \geq 1$ . Let  $\psi : \Theta \rightarrow \mathbb{R}$  be a regularization which is  $\mu$ -strongly convex with respect to some norm  $\|\cdot\|$  and let  $D \geq \sqrt{\max_{\theta} \psi(\theta) - \min_{\theta} \psi(\theta)}$ . Then, for a well-chosen  $\eta > 0$ , the regret of FTRL is upper-bounded as

$$R_T \leq D \sqrt{\frac{2}{\mu} \sum_{t=1}^T \|g_t\|_*^2},$$

where  $\|\cdot\|_* \stackrel{\text{def}}{=} \sup_{\|\theta\| \leq 1} \langle \cdot, \theta \rangle$  is the dual norm of  $\|\cdot\|$ .

*Proof.* Let  $T \geq 1$ . Define for all  $t \geq 1$  the function  $\Phi_t$  such that for all  $\theta \in \Theta$

$$\Phi_t(\theta) = \eta \sum_{s=1}^t \langle \theta, g_s \rangle + \psi(\theta).$$

Note that by definition  $\theta_{t+1} = \arg \min_{\theta} \Phi_t(\theta)$ . Therefore, for any  $\theta$ ,  $\langle \nabla \Phi_t(\theta_{t+1}), \theta_{t+1} - \theta \rangle \leq 0$ , which entails by  $\mu$ -strong convexity of  $\psi$  and thus of  $\Phi_t$ ,

$$\Phi_t(\theta_{t+1}) - \Phi_t(\theta) \leq -\frac{\mu}{2} \|\theta - \theta_{t+1}\|^2, \quad \forall t \in [T], \forall \theta \in \Theta. \quad (2.7)$$

Let  $\theta \in \Theta$ . Then,

$$\begin{aligned} R_T(\theta) &\stackrel{\text{def}}{=} \sum_{t=1}^T \langle \theta_t, g_t \rangle - \sum_{t=1}^T \langle \theta, g_t \rangle \\ &= \sum_{t=1}^T \langle \theta_t, g_t \rangle - \frac{\Phi_T(\theta)}{\eta} + \frac{\psi(\theta)}{\eta} \\ &\stackrel{(2.7)}{\leq} \sum_{t=1}^T \langle \theta_t, g_t \rangle - \frac{\Phi_T(\theta_{T+1})}{\eta} - \frac{\mu}{2\eta} \|\theta_{T+1} - \theta\|^2 + \frac{\psi(\theta)}{\eta} \\ &= \sum_{t=1}^T \langle \theta_t, g_t \rangle - \langle \theta_{T+1}, g_T \rangle - \frac{\Phi_{T-1}(\theta_{T+1})}{\eta} - \frac{\mu}{2\eta} \|\theta_{T+1} - \theta\|^2 + \frac{\psi(\theta)}{\eta} \\ &\stackrel{(2.7)}{\leq} \sum_{t=1}^T \langle \theta_t, g_t \rangle - \langle \theta_{T+1}, g_T \rangle - \frac{\Phi_{T-1}(\theta_T)}{\eta} - \frac{\mu}{2\eta} \|\theta_T - \theta_{T+1}\|^2 - \frac{\mu}{2\eta} \|\theta_{T+1} - \theta\|^2 + \frac{\psi(\theta)}{\eta} \\ &\leq \sum_{t=1}^T \left( \langle \theta_t - \theta_{t+1}, g_t \rangle - \frac{\mu}{2\eta} \|\theta_t - \theta_{t+1}\|^2 \right) - \frac{\mu}{2\eta} \|\theta_{T+1} - \theta\|^2 + \frac{\psi(\theta) - \psi(\theta_1)}{\eta} \quad \leftarrow \text{by induction on } t \\ &\leq \sum_{t=1}^T \left( \|\theta_t - \theta_{t+1}\| \|g_t\|_* - \frac{\mu}{2\eta} \|\theta_t - \theta_{t+1}\|^2 \right) + \frac{\psi(\theta) - \psi(\theta_1)}{\eta} \quad \leftarrow \text{by Cauchy-Schwarz} \end{aligned}$$

$$\leq \sum_{t=1}^T \left( \|\theta_t - \theta_{t+1}\| \|g_t\|_* - \frac{\mu}{2\eta} \|\theta_t - \theta_{t+1}\|^2 \right) + \frac{D^2}{\eta} \quad \leftarrow \text{by definition of } D$$

(2.8)

Now using Young's inequality

$$\forall a, b, \gamma > 0, \quad ab \leq \frac{\gamma a^2}{2} + \frac{b^2}{2\gamma},$$

this yields

$$\begin{aligned} R_T(\theta) &\leq \sum_{t=1}^T \left( \frac{\mu}{2\eta} \|\theta_t - \theta_{t+1}\|^2 + \frac{\eta}{2\mu} \|g_t\|_*^2 - \frac{\mu}{2\eta} \|\theta_t - \theta_{t+1}\|^2 \right) + \frac{D^2}{\eta} \\ &= \frac{\eta}{2\mu} \sum_{t=1}^T \|g_t\|_*^2 + \frac{D^2}{\eta}. \end{aligned}$$

Optimizing  $\eta$  concludes the proof.  $\square$

### 2.2.3 Online Mirror Descent

Online Mirror Descent (OMD) is another generalization of OGD to better exploit the geometry of the decision space  $\Theta$ . OMD is the online counterpart of the *Mirror Descent* algorithm from convex optimization. The generality of OMD comes from the updates being performed into a dual space which is defined by a convex differentiable regularization function  $\psi : \Theta \rightarrow \mathbb{R}$ , called mirror map.

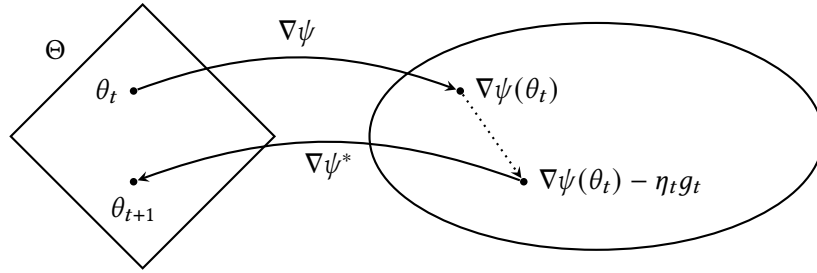


Figure 2.1: Illustration of Online Mirror Descent (Agile version)

Before stating the algorithm, we need to define the Bregman divergence.

#### Definition 2.3

For any continuously differentiable convex function  $\psi$ , the Bregman divergence with respect to  $\psi$  is defined as

$$D_\psi(x, y) \leq \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle \quad \forall x, y \in \Theta.$$

It is the difference between the value of the regularization function at  $x$  and the value of its first order Taylor approximation. It is nonnegative but not symmetric.

Online Mirror Descent is then defined as follows.

The lazy version performs all gradient steps in the dual space before computing the projection over  $\Theta$  to form predictions  $\theta_t$ , while the agile version alternates projections and gradient steps in the dual space.

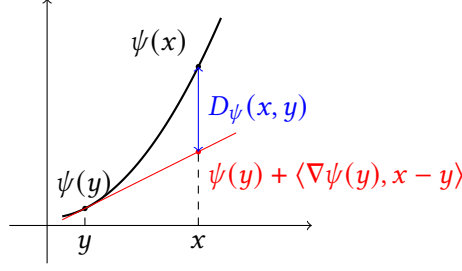


Figure 2.2: Illustration of the Bregman divergence

Parameters:  $\eta > 0$ , mirror map  $\psi$

Initialize:  $\theta_1 \in \Theta$ ,  $z_1 = x_1$

For  $t = 1, \dots, T$

- select  $\theta_t$ ; incur loss  $\ell_t(\theta_t) = \langle \theta_t, g_t \rangle$  and observe  $g_t \in \mathbb{R}^d$
- update  $z_t$  such that

$$\nabla\psi(z_{t+1}) = \nabla\psi(\theta_t) - \eta g_t \quad \leftarrow \quad \text{agile version}$$

$$\nabla\psi(z_{t+1}) = \nabla\psi(z_t) - \eta g_t \quad \leftarrow \quad \text{lazy version}$$

- project according to the Bregman divergence

$$\theta_{t+1} \in \arg \min_{\theta \in \Theta} D_\psi(\theta, z_{t+1}).$$

**Algorithm 2.4:** Online Mirror Descent (OMD) – Lazy and agile versions

**Equivalent formulation of OMD agile** OMD (agile version) is equivalent to the following update

$$\theta_{t+1} = \arg \min_{\theta \in \Theta} \left\{ \langle g_t, \theta - \theta_t \rangle + \frac{1}{\eta} D_\psi(\theta, \theta_t) \right\} \quad (2.9)$$

Indeed,

$$\begin{aligned} & \arg \min_{\theta \in \Theta} \left\{ \langle g_t, \theta - \theta_t \rangle + \frac{1}{\eta} D_\psi(\theta, \theta_t) \right\} \\ &= \arg \min_{\theta \in \Theta} \left\{ \eta \langle g_t, \theta - \theta_t \rangle + \psi(\theta) - \cancel{\psi(\theta_t)} - \langle \nabla\psi(\theta_t), \theta - \cancel{\theta_t} \rangle \right\} \\ &= \arg \min_{\theta \in \Theta} \left\{ \langle \eta g_t - \nabla\psi(\theta_t), \theta \rangle + \psi(\theta) \right\} \\ &= \arg \min_{\theta \in \Theta} \left\{ \langle -\nabla\psi(z_{t+1}), \theta \rangle + \psi(\theta) \right\} \quad \text{since } \nabla\psi(z_{t+1}) = \nabla\psi(\theta_t) - \eta g_t \\ &= \arg \min_{\theta \in \Theta} \left\{ -\cancel{\psi(z_{t+1})} - \langle \nabla\psi(z_{t+1}), \theta - \cancel{z_{t+1}} \rangle + \psi(\theta) \right\} \\ &= \arg \min_{\theta \in \Theta} \left\{ D_\psi(\theta, z_{t+1}) \right\} \end{aligned}$$

**Equivalent formulation of OMD (lazy version) – linearized FTRL** OMD (lazy version) is the linearized version of Follow The Regularized Leader (FTRL, Alg. 2.3)

$$\theta_{t+1} = \arg \min_{\theta \in \Theta} \left\{ \sum_{s=1}^t \langle g_s, \theta - \theta_s \rangle + \frac{1}{\eta} D_\psi(\theta, \theta_1) \right\} \quad (2.10)$$

Indeed, for the lazy version  $\nabla \psi(z_{t+1}) = \nabla \psi(z_t) - \eta g_t = \nabla \psi(z_1) - \eta \sum_{s=1}^t g_s$ , and  $z_1 = x_1$ . Thus

$$\begin{aligned} & \arg \min_{\theta \in \Theta} \left\{ \sum_{s=1}^t \langle g_s, \theta - \theta_s \rangle + \frac{1}{\eta} D_\psi(\theta, \theta_1) \right\} \\ &= \arg \min_{\theta \in \Theta} \left\{ \eta \sum_{s=1}^t \langle g_s, \theta - \theta_s \rangle + \psi(\theta) - \cancel{\psi(\theta_1)} - \langle \nabla \psi(\theta_1), \theta - \cancel{\theta_1} \rangle \right\} \\ &= \arg \min_{\theta \in \Theta} \left\{ \left\langle \sum_{s=1}^t \eta g_s - \nabla \psi(\theta_1), \theta \right\rangle + \psi(\theta) \right\} \\ &= \arg \min_{\theta \in \Theta} \left\{ \langle -\nabla \psi(z_{t+1}), \theta \rangle + \psi(\theta) \right\} \\ &= \arg \min_{\theta \in \Theta} \left\{ D_\psi(\theta, z_{t+1}) \right\}. \end{aligned}$$

The regret bound of OMD lazy follows thus from Theorem 2.8.

### Theorem 2.9

### Regret of OMD

Let  $T \geq 1$ . Let  $\Theta$  be a compact and convex,  $\theta_1 \in \Theta$ , and  $\psi : \Theta \rightarrow \mathbb{R}$  be  $\mu$ -strongly convex mirror map with respect to some norm  $\|\cdot\|$ . Assume

$$D_\psi(\theta, \theta_1)^{1/2} \leq D \quad \text{and} \quad \|g_t\|_* \leq G_*.$$

Then, for any convex  $\ell_1, \dots, \ell_T : \Theta \rightarrow \mathbb{R}$ , the regret of OMD (agile or lazy version) with  $\eta = DG_*^{-1} \sqrt{2\mu/T}$  is upper bounded as

$$\text{Reg}_T \leq DG_* \sqrt{\frac{2T}{\mu}}.$$

**Remark.** OMD is the generalization of two well-studied algorithms in online learning:

- The agile version with  $\Theta \subset \mathbb{R}^d$  and  $\psi(\theta) = \frac{1}{2} \|\theta\|^2$  leads to Online Gradient Descent (Alg. 2.2)

$$\theta_t = \text{Proj}_\Theta (\theta_{t-1} - \eta \nabla \ell_{t-1}(\theta_{t-1})), \quad (2.11)$$

where  $\text{Proj}_\Theta$  is the Euclidean projection over  $\Theta$ . The regret bound is then  $\text{Reg}_T \lesssim D_2 G_2 \sqrt{T}$ . The lazy version is slightly different but provides the same guarantee.

- $\Theta = \Delta_d$  and the neg-entropy regularization  $\psi(\theta) = \sum_{i=1}^d \theta(i) \log \theta(i)$ , entails the exponentiated gradient algorithm (EG, Alg. 2.1)

$$\theta_t(i) = \frac{\theta_{t-1}(i) e^{-\eta g_{t-1}(i)}}{\sum_{j=1}^d e^{-\eta g_{t-1}(\theta_{t-1})(j)}}, \quad (2.12)$$

which has the advantage of being closed-form. Indeed, we can check that the above update solves the OMD optimization step (agile or lazy update), for any  $\theta \in \Delta_d$ ,

$$\begin{aligned}
\langle g_{t-1}, \theta_t \rangle + \frac{1}{\eta} KL(\theta_t \| \theta_{t-1}) &\stackrel{(2.12)}{=} \frac{\langle g_{t-1}, e^{-\eta g_{t-1}} \theta_{t-1} \rangle}{\eta \langle \theta_{t-1}, e^{-\eta g_{t-1}} \rangle} + \frac{\langle e^{-\eta g_{t-1}} \theta_{t-1}, -\eta g_{t-1} - \log \langle \theta_{t-1}, e^{-\eta g_{t-1}} \rangle \rangle}{\eta \langle \theta_{t-1}, e^{-\eta g_{t-1}} \rangle} \\
&= -\frac{1}{\eta} \log \left( \langle \theta_{t-1}, e^{-\eta g_{t-1}} \rangle \right) \\
&\leq -\frac{1}{\eta} \log \left( \langle \theta, e^{-\eta g_{t-1} + \log \frac{\theta_{t-1}}{\theta}} \rangle \right) \\
&\stackrel{\text{Jensen}}{\leq} -\frac{1}{\eta} \log \left( e^{\langle \theta, -\eta g_{t-1} + \log \frac{\theta_{t-1}}{\theta} \rangle} \right) = \langle g_{t-1}, \theta \rangle + \frac{1}{\eta} KL(\theta \| \theta_{t-1}).
\end{aligned} \tag{2.13}$$

The mirror map  $\psi$  is 1-strongly convex with respect to the  $\ell_1$ -norm. Indeed, by Pinsker's inequality

$$D_\psi(x, y) \stackrel{\text{def}}{=} \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle = KL(x \| y) \geq \frac{\|x - y\|_1^2}{2}.$$

Thus, Theorem 2.9 provides the regret upper bound  $\text{Reg}_T \lesssim D_1 G_\infty \sqrt{T \log d}$ .

*Proof of Theorem 2.9.* Let  $t \geq 1$  and  $\theta \in \Theta$ . Denote  $\Phi_t(\theta) = \langle g_t, \theta \rangle + \frac{1}{\eta} D_\psi(\theta, \theta_t)$ . From the optimality condition on  $\theta_{t+1}$ ,  $\langle \nabla \Phi_t(\theta_{t+1}), \theta_{t+1} - \theta \rangle \leq 0$ , which entails  $\langle g_t, \theta_{t+1} - \theta \rangle \leq \frac{1}{\eta} \langle \nabla \psi(\theta_{t+1}) - \nabla \psi(\theta_t), \theta - \theta_{t+1} \rangle$  and yields

$$\begin{aligned}
\ell_t(\theta_t) - \ell_t(\theta) &= \langle g_t, \theta_t - \theta \rangle = \langle g_t, \theta_t - \theta_{t+1} \rangle + \langle \nabla \ell_t(\theta_t), \theta_{t+1} - \theta \rangle \\
&\leq \langle g_t, \theta_t - \theta_{t+1} \rangle + \frac{1}{\eta} \langle \nabla \psi(\theta_{t+1}) - \nabla \psi(\theta_t), \theta - \theta_{t+1} \rangle \\
&= \langle g_t, \theta_t - \theta_{t+1} \rangle + \frac{1}{\eta} \left( D_\psi(\theta, \theta_t) - D_\psi(\theta, \theta_{t+1}) - D_\psi(\theta_{t+1}, \theta_t) \right) \\
&\leq \langle g_t, \theta_t - \theta_{t+1} \rangle + \frac{1}{\eta} \left( D_\psi(\theta, \theta_t) - D_\psi(\theta, \theta_{t+1}) - \frac{\mu}{2} \|\theta_t - \theta_{t+1}\|^2 \right),
\end{aligned}$$

where the last inequality because  $\psi$  is  $\mu$ -strongly convexity. Moreover,

$$\langle g_t, \theta_t - \theta_{t+1} \rangle \stackrel{\text{Cauchy-Schwarz}}{\leq} \|\theta_t - \theta_{t+1}\| \|g_t\|_* \stackrel{\text{Young's}}{\leq} \frac{\mu}{2\eta} \|\theta_t - \theta_{t+1}\|^2 + \frac{\eta}{2\mu} \|g_t\|_*^2.$$

Combining the last two upper-bounds and summing over  $t = 1, \dots, T$  entails the stated regret upper bound.  $\square$

## 3 Online Convex Optimization

In this section, we aim at generalizing the previous algorithms beyond linear losses  $\ell_t : \Theta \rightarrow \mathbb{R}$ .

### 3.1 The Gradient Trick (from linear to convex losses)

**Setting:** compact convex decision set  $\Theta$ , convex and sub-differentiable loss functions.

Here, we show how to generalize Chapter 2 to sub-differentiable loss functions  $\ell_t : \Theta \subset \mathbb{R}^d \rightarrow \mathbb{R}$ .

#### Definition 3.1

#### Sub-gradient

A sub-gradient of a convex function  $\ell_t$  at point  $\theta \in \Theta$  is a point in  $\mathbb{R}^d$ , denoted  $\nabla \ell_t(\theta)$ , that satisfies the convexity inequality:

$$\forall \theta' \in \Theta, \quad \ell_t(\theta) - \ell_t(\theta') \leq \langle \nabla \ell_t(\theta), \theta - \theta' \rangle.$$

To do so, we introduce the gradient trick that consist in linearizing the losses to apply the results of the previous section. The idea is to apply previous algorithms on the linear losses  $\langle g_t, \theta \rangle$  with  $g_t = \nabla \ell_t(\theta_t)$ . In this case, we have from the definition of the sub-gradients,

$$R_T = \sum_{t=1}^T \ell_t(\theta_t) - \ell_t(\theta^*) \leq \sum_{t=1}^T \langle g_t, \theta_t \rangle - \langle g_t, \theta^* \rangle,$$

where  $\theta^* = \arg \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta)$ . Hence, upper-bounding the regret with linear losses  $\theta \mapsto \langle g_t, \theta \rangle$  also upper-bound the regret with the true losses  $\ell_t$ . The following theorem follows.

#### Theorem 3.1

#### Gradient trick

Let  $\mathcal{A}$  be an algorithm that satisfies the regret-upper bound

$$\sum_{t=1}^T \langle g_t, \theta_t \rangle - \langle g_t, \theta^* \rangle \leq R(T, (g_t), \Theta)$$

for some real-valued function  $R$ . Then, for any sequence of sub-differentiable losses  $\ell_t : \Theta \rightarrow \mathbb{R}$ , applying  $\mathcal{A}$  with  $g_t = \nabla \ell_t(\theta_t)$  achieves the regret upper-bound

$$R_T \stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta) \leq R(T, (\nabla \ell_t(\theta_t)), \Theta).$$

This theorems shows that all results (Theorems 2.1, 2.7, 2.8 and 2.9) of Chapter 2 can be transposed to convex loss functions  $\ell_t : \Theta \rightarrow \mathbb{R}$ . Applied with  $g_t = \nabla \ell_t(\theta_t)$ , we get the following regret upper-bounds:

$$R_T \leq 2G_\infty \sqrt{T \log d} \quad (\text{EWA, Thm 2.1})$$

$$R_T \leq D_2 G_2 \sqrt{T} \quad (\text{OGD, Thm. 2.7})$$

$$R_T \leq D G_* \sqrt{2\mu^{-1}T} \quad (\text{FTRL, Thm. 2.8 and OMD, Thm. 2.9})$$

**The EG Algorithm** When the decision space is the simplex  $\Theta = \Delta_d$ , applying the above trick to EWA results in an algorithm called Exponentiated Gradient forecaster (EG). Recall that in this case we denote the decision  $\theta_t = p_t$ .

### 3.2 Faster rate for losses with curvature (ONS, Non-linear Ridge, EWA)

OMD or FTRL are efficient algorithms and their regret bound in Theorem 2.9 is satisfying in general. Yet, the rate  $O(\sqrt{T})$  is sub-optimal for some losses with curvature (e.g., squared, logistic, or log loss) for which logarithmic regret may be achieved. This is the focus of this section.

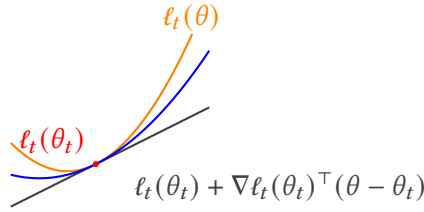


Figure 3.1: The gradient trick approximation may be improved if loss has curvature.

When losses have curvature, logarithmic regret may be achieved. Various assumptions exist to achieve such rates (see e.g., [Van Erven et al. \[2015\]](#)). The most famous one is  $\mu$ -strong convexity, in which case OGD with time varying learning rate  $\eta_t = \mu/t$  achieves a regret of order  $O(\frac{d}{\mu} \log T)$ .

**Exercise 3.1.** Prove a regret upper-bound of order  $O(\frac{d}{\mu} \log T)$  for OGD with time-varying learning rate  $\eta_t = \frac{\mu}{t}$  when the losses  $\ell_t$  are  $\mu$ -strongly convex.

Here, we will focus on a slightly weaker assumption called  $\eta$ -exp-concavity, which assumes that  $\theta \mapsto \exp(-\eta \ell_t(\theta))$  are concave in  $\theta$ . Exp-concavity is related to being strongly convex in the direction of the gradient only as shown by the following lemma.

#### Lemma 3.2

Let  $\ell_t : \Theta \rightarrow \mathbb{R}$  be an  $\eta$ -exp-concave loss function and  $\theta_t \in \Theta$ . Then, for any  $\theta \in \Theta$

$$\ell_t(\theta) \geq \ell_t(\theta_t) + \langle \nabla \ell_t(\theta_t), \theta - \theta_t \rangle + \frac{\gamma}{2} \langle \nabla \ell_t(\theta_t), \theta - \theta_t \rangle^2$$

where  $\gamma \leq \frac{1}{2} \min \left\{ \eta, \frac{1}{4G_*D} \right\}$ , where  $G_*$  and  $D$  are respectively bounds on  $\|\nabla \ell_t(\theta_t)\|_*$  and on the diameter of  $\Theta$ .

#### Lemma 4.2 Hazan et al. [2016]

We introduce below three algorithms, each achieving logarithmic regret under the exp-concavity assumption (or for the squared loss only for the non-linear Ridge regression algorithm thereafter).



### 3.2.1 Online Newton Step (ONS)

The high-level idea of the next algorithm, called Online Newton Step (ONS), introduced by ??, is to include the above quadratic upper-bound of the losses into a follow the regularized leader scheme:

$$\theta_t = \arg \min_{\theta \in \Theta} \left\{ \sum_{s=1}^{t-1} \langle \nabla \ell_s(\theta_s), \theta - \theta_s \rangle + \frac{\gamma}{2} \langle \nabla \ell_s(\theta_s), \theta - \theta_s \rangle^2 + \frac{\lambda}{2} \|\theta\|^2 \right\}. \quad (3.1)$$

Note that the above formulation of ONS as a variant of follow the regularized leader is due to and analyzed by [Orabona \[2019\]](#), and called follow the approximate leader in the original work of ?. Its equivalence with the original formulation of ONS is not clear since here one could take  $\gamma = 0$  in the algorithm which is not the case for the original one but the regret and analysis are similar, and this variant is easier to motivate.

#### Theorem 3.3

[Orabona \[2019\]](#)

Let  $T \geq 1$ . Let  $\Theta \subset \mathbb{R}^d$  be a convex set with diameter  $D > 0$ . Then, for all sequences of  $G$ -Lipschitz,  $\eta$ -exp-concave losses  $\ell_1, \dots, \ell_T : \Theta \rightarrow \mathbb{R}$ , the regret of ONS with hyper-parameters  $\lambda > 0$  and  $\gamma \leq \frac{1}{2} \min \left\{ \eta, \frac{1}{4GD} \right\}$  is upper-bounded as:

$$\forall \theta \in \Theta, \quad \text{Reg}_T(\theta) = \sum_{t=1}^T \ell_t(\theta_t) - \ell_t(\theta) \leq \frac{\lambda}{2} \|\theta\|^2 + \frac{d}{2\gamma} \log \left( 1 + \frac{\gamma T G^2}{d\lambda} \right).$$

Hence, if the losses are  $\eta$ -exp-concave, a logarithmic regret may be guaranteed. Yet, the above theorem assumes bounded domain  $\Theta$  and bounded gradients, with an upper-bound of order at least  $O(dGD \log T)$ . We will see in the next subsection how to remove these assumptions for the particular case of the squared loss.

Futhermore, the complexity of ONS suffers from the projection over  $\Theta$ , which is higher than the one of OMD but still manageable for finite dimensional spaces. Overall, ignoring the projection step, the computation of ONS would require the inversion of a  $d \times d$  matrices and could be done in  $O(d^3)$  operations per round.

*Proof of Theorem 3.3.* For any  $t \geq 1$ , define the matrix  $A_t = \gamma \sum_{s=1}^t \nabla \ell_s(\theta_s) \nabla \ell_s(\theta_s)^\top + \lambda I$  and set for any  $\theta \in \Theta$

$$\psi_t(\theta) = \frac{\gamma}{2} \sum_{s=1}^t \langle \nabla \ell_s(\theta_s), \theta - \theta_s \rangle^2 + \frac{\lambda}{2} \|\theta\|^2 \quad \text{and} \quad \Phi_t(\theta) = \sum_{s=1}^t \langle \nabla \ell_s(\theta_s), \theta - \theta_s \rangle + \psi_t(\theta).$$

Then,  $\psi_t$  and  $\Phi_t$  are 1-strongly convex with respect to the norm  $\|\cdot\|_{A_t}$ . Then, on the one hand, by Cauchy-Schwarz inequality

$$\sum_{t=1}^T \langle \nabla \ell_t(\theta_t), \theta_t - \theta_{t+1} \rangle \leq \frac{1}{2} \sum_{t=1}^T \|\nabla \ell_t(\theta_t)\|_{A_t^{-1}}^2 + \frac{1}{2} \sum_{t=1}^T \|\theta_t - \theta_{t+1}\|_{A_t}^2. \quad (3.2)$$

On the other hand, fixing  $\theta \in \Theta$ , since  $\langle \nabla \ell_t(\theta_t), \theta_{t+1} - \theta_t \rangle = \Phi_t(\theta_{t+1}) - \Phi_{t-1}(\theta_{t+1}) - \frac{\gamma}{2} \langle \ell_t(\theta_t), \theta_{t+1} - \theta_t \rangle^2$ , we have

$$\sum_{t=1}^T \langle \nabla \ell_t(\theta_t), \theta_{t+1} - \theta \rangle = -\Phi_T(\theta) + \psi_T(\theta) + \sum_{t=1}^T \Phi_t(\theta_{t+1}) - \Phi_{t-1}(\theta_{t+1}) - \frac{\gamma}{2} \langle \ell_t(\theta_t), \theta_{t+1} - \theta_t \rangle^2$$

$$\begin{aligned}
&= \cancel{\Phi_T(\theta_{T+1}) - \Phi_T(\theta) - \Phi_0(\theta_1)} + \psi_T(\theta) + \sum_{t=1}^T \Phi_{t-1}(\theta_t) - \Phi_{t-1}(\theta_{t+1}) - \frac{\gamma}{2} \langle \ell_t(\theta_t), \theta_{t+1} - \theta_t \rangle^2 \\
&\leq \psi_T(\theta) - \frac{1}{2} \sum_{t=1}^T \|\theta_{t+1} - \theta_t\|_{A_{t-1}}^2 - \frac{\gamma}{2} \langle \ell_t(\theta_t), \theta_{t+1} - \theta_t \rangle^2 = \psi_T(\theta) - \frac{1}{2} \sum_{t=1}^T \|\theta_{t+1} - \theta_t\|_{A_t}^2. \quad (3.3)
\end{aligned}$$

where the last inequality is because  $\Phi_T(\theta_{T+1}) \leq \Phi_T(\theta)$  and by strong convexity of  $\Phi_{t-1}$  and by the optimality condition of  $\theta_t = \arg \min_{\theta \in \Theta} \Phi_{t-1}(\theta)$  we have  $\Phi_{t-1}(\theta_t) - \Phi_{t-1}(\theta) \leq -\frac{1}{2} \|\theta_t - \theta\|_{A_{t-1}}^2$ . Combining (3.2) and (3.3) into the quadratic upper-bound of Lemma 3.2 yields

$$\sum_{t=1}^T \ell_t(\theta_t) - \ell_t(\theta) \stackrel{\text{Lem. 3.2}}{\leq} \frac{\lambda}{2} \|\theta\|^2 - \Psi_T(\theta) + \sum_{t=1}^T \langle \nabla \ell_t(\theta_t), \theta_t - \theta \rangle \leq \frac{\lambda}{2} \|\theta\|^2 + \frac{1}{2} \sum_{t=1}^T \|\nabla \ell_t(\theta_t)\|_{A_t^{-1}}^2. \quad (3.4)$$

Applying Lemma 11.11 of ? stating that for any full rank matrix  $A$  and any vector  $x \in \mathbb{R}^d$ , then  $x^\top (A + xx^\top)^{-1} x = 1 - \det(A) / \det(A + xx^\top)$ , entails

$$\sum_{t=1}^T \left(1 - \frac{\det(A_{t-1})}{\det(A_t)}\right)^{1-u \leq -\log u} \leq \sum_{t=1}^T \log \frac{\det(A_t)}{\det(A_{t-1})} = \log \frac{\det(A_{T+1})}{\det(A_0)} \leq d \log \left(1 + \frac{\gamma T G^2}{d\lambda}\right),$$

where the last line used that  $\|\nabla \ell_t(\theta_t)\| \leq G$ . Substituting into (3.4) concludes the proof.  $\square$

### 3.2.2 Non-linear Ridge Regression

In the setting of online linear regression, the learner observes an input  $x_t \in \mathbb{R}^d$  before making its prediction  $\theta_t \in \mathbb{R}^d$  and observing an output  $y_t \in \mathbb{R}$ . The loss is of the form  $\ell_t(\theta) = (y_t - \langle \theta, x_t \rangle)^2$ . One may show that for bounded domains, inputs, and outputs, the squared loss is exp-concave and ONS detailed in the previous section may be applied. Yet, these bounded restrictions may be reduced with the non-linear ridge regression forecaster, which is the online counterpart of the well-known Ridge regression estimator in the statistical framework. It produces the estimator as follows:

$$\theta_t = \arg \min_{\theta \in \mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} (y_s - \langle \theta, x_s \rangle)^2 + \lambda \|\theta\|^2 + \langle x_t, \theta \rangle^2 \right\}. \quad (3.5)$$

The last term  $\langle x_t, \theta \rangle^2$  in the above definition is an additional regularization that does not exist in the classical Ridge regression estimator for i.i.d. data. It can be viewed as an additional regularization in the direction of the next input. Since, the output is unknown when choosing  $\theta_t$ , the algorithm replaces it with 0 by adding a loss  $(0 - \langle x_t, \theta \rangle)^2$ .

A key property of this regularization is its dependence on the next input  $x_t$ , making  $\theta_t$  depend on  $x_t$  itself and the prediction  $\langle \theta_t, x_t \rangle$  a non-linear transformation of  $x_t$ . Since the algorithm is compared with respect to linear predictor only of the form  $\theta \mapsto \langle \theta, x_t \rangle$ , we say that the algorithm is *improper*.

The analysis of the following regret upper-bound of the non-linear ridge regression forecaster is standard and can be found in [Cesa-Bianchi and Lugosi \[2006\]](#). It is of order  $O(dB^2 \log(XT))$  for  $\lambda = 1$  when  $|y_t| \leq B$  and  $\|x_t\| \leq X$ .

#### Theorem 3.4

#### Theorem 11.8, Cesa-Bianchi and Lugosi [2006]

Let the non-linear ridge regression (3.5) be run with parameter  $\lambda > 0$ . For all  $T \geq 1$ , for all sequences

$x_1, \dots, x_T$  and all  $y_1, \dots, y_T \in [-B, B]$ , for all  $\theta \in \mathbb{R}^d$

$$\text{Reg}_T(\theta) \leq \lambda \|\theta\|^2 + B^2 \sum_{k=1}^d \log \left( 1 + \frac{\lambda_k(K_T)}{\lambda} \right),$$

where  $K_T = \sum_{t=1}^T x_t x_t^\top$  and  $\lambda_k(K_T)$  denotes the  $k$ -th eigenvalue of  $K_T$ .

Note that compared to ONS, the non-linear ridge regression forecaster works on unbounded domain and will get  $O(dB^2 \log T)$  regret with respect to any fixed  $\theta \in \mathbb{R}^d$ . On such domain gradients of the losses are unbounded, Lemma 3.2 cannot be applied. In ?, we provide uniform regret bounds over  $\mathbb{R}^d$  for this forecaster.

*Proof.* The proof follows the one of ONS (Theorem 3.3). Since it is not used in the prediction, we assume here without loss of generality that  $x_{T+1} = 0$ . First, define the matrix  $A_t \stackrel{\text{def}}{=} \sum_{s=1}^t x_s x_s^\top + \lambda I$  and remark that

$$\theta_t \in \arg \min_{\theta \in \mathbb{R}^d} \Phi_{t-1}(\theta), \quad \text{where} \quad \Phi_{t-1}(\theta) \stackrel{\text{def}}{=} \sum_{s=1}^{t-1} -y_s \langle x_s, \theta \rangle + \psi_t(\theta) \quad \text{and} \quad \psi_t(\theta) \stackrel{\text{def}}{=} \frac{1}{2} \|\theta\|_{A_t}^2.$$

Then,  $\psi_t$  and  $\Phi_{t-1}$  are 1-strongly convex with respect to the norm  $\|\cdot\|_{A_t}$ . Then, on the one hand, by Cauchy-Schwarz inequality

$$\sum_{t=1}^T -y_t \langle x_t, \theta_t - \theta_{t+1} \rangle \leq \frac{1}{2} \sum_{t=1}^T y_t^2 \|x_t\|_{A_t^{-1}}^2 + \frac{1}{2} \sum_{t=1}^T \|\theta_t - \theta_{t+1}\|_{A_t}^2. \quad (3.6)$$

On the other hand, fixing  $\theta \in \Theta$ , since  $-y_t \langle x_t, \theta_{t+1} \rangle = \Phi_t(\theta_{t+1}) - \Phi_{t-1}(\theta_{t+1}) - \frac{1}{2} \langle x_{t+1}, \theta_{t+1} \rangle^2$ , we have

$$\begin{aligned} \sum_{t=1}^T -y_t \langle x_t, \theta_{t+1} - \theta \rangle &= -\Phi_T(\theta) + \frac{\lambda}{2} \|\theta\|^2 + \sum_{t=1}^T \frac{1}{2} \langle x_t, \theta \rangle^2 + \Phi_t(\theta_{t+1}) - \Phi_{t-1}(\theta_{t+1}) - \frac{1}{2} \langle x_{t+1}, \theta_{t+1} \rangle^2 \\ &= \cancel{\Phi_T(\theta_{T+1}) - \Phi_T(\theta) - \Phi_0(\theta_1)} + \frac{\lambda}{2} \|\theta\|^2 + \sum_{t=1}^T \frac{1}{2} \langle x_t, \theta \rangle^2 + \Phi_{t-1}(\theta_t) - \Phi_{t-1}(\theta_{t+1}) - \frac{1}{2} \langle x_{t+1}, \theta_{t+1} \rangle^2 \\ &\leq \frac{\lambda}{2} \|\theta\|^2 + \sum_{t=1}^T \frac{1}{2} \langle x_t, \theta \rangle^2 - \frac{1}{2} \|\theta_{t+1} - \theta_t\|_{A_t}^2 - \frac{1}{2} \langle x_{t+1}, \theta_{t+1} \rangle^2, \end{aligned} \quad (3.7)$$

where the last inequality is because  $\Phi_T(\theta_{T+1}) \leq \Phi_T(\theta)$  and by strong convexity of  $\Phi_{t-1}$  and by the optimality condition of  $\theta_t = \arg \min_{\theta \in \Theta} \Phi_{t-1}(\theta)$  we have  $\Phi_{t-1}(\theta_t) - \Phi_{t-1}(\theta) \leq -\frac{1}{2} \|\theta_t - \theta\|_{A_t}^2$ . Combining (3.6) and (3.7) and using  $x_{T+1} = 0$  and  $\theta_1 = 0$  yields

$$\text{Reg}_T(\theta) = \sum_{t=1}^T -y_t \langle x_t, \theta_t - \theta_{t+1} \rangle - y_t \langle x_t, \theta_{t+1} - \theta \rangle + 2 \langle \theta_t, x_t \rangle^2 - 2 \langle \theta, x_t \rangle^2 \stackrel{(3.6)+(3.7)}{\leq} \lambda \|\theta\|^2 + \sum_{t=1}^T y_t^2 \|x_t\|_{A_t^{-1}}^2. \quad (3.8)$$

The proof is concluded similarly to the one of Theorem 3.3 by applying Lemma 11.11 of ? together with  $|y_t| \leq B$ .  $\square$

### 3.2.3 Variants of Exponential Weights for continuous action spaces

Here, we present simple solutions via discretization or integration of EWA that yield strong theoretical baselines (for the regret) when the action space is continuous but often yields prohibitive complexity.

## Continuous EWA

Let  $\Theta \in \mathbb{R}^d$  be a compact and convex subset of  $\mathbb{R}^d$ . We consider the following continuous variant of EWA, that predicts

$$\theta_t = \frac{\int_{\Theta} \theta e^{-\eta \sum_{s=1}^{t-1} \ell_s(\theta)} d\mu(\theta)}{\int_{\Theta} e^{-\eta \sum_{s=1}^{t-1} \ell_s(\theta')} d\mu(\theta')},$$

where  $\mu$  is the Lebesgue measure on  $\Theta$ .

### Theorem 3.5

### Regret of continuous EWA

Let  $T \geq 1$ . For all sequences of  $\eta$ -exp-concave losses  $\ell_1, \dots, \ell_T$  the continuous EWA forecaster satisfies

$$R_T \stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(\theta_t) - \inf_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta) \leq \frac{1 + d \log(T+1)}{\eta}.$$

Note that if  $\Theta = \Delta_d$ , then  $\Theta$  is of dimension  $d-1$  and not  $d$ . As an exercise, prove a regret upper-bound without the expconcavity assumption of order  $O(\sqrt{T})$ .

*Proof.* The proof starts similarly to the one of Theorem 2.4. Let us denote  $W_t(\theta) = e^{-\eta \sum_{s=1}^t \ell_s(\theta)}$ ,  $W_t = \int_{\Theta} W_t(\theta) d\mu(\theta)$  and  $d\hat{\mu}_t(\theta) = W_t(\theta) d\mu(\theta) / W_t$ . Then,

$$\begin{aligned} W_T &= \int_{\Theta} e^{-\eta \sum_{t=1}^T \ell_t(\theta)} d\mu(\theta) \\ &= W_{T-1} \int_{\Theta} \frac{W_{T-1}(\theta)}{W_{T-1}} e^{-\eta \ell_T(\theta)} d\mu(\theta) \\ &= W_{T-1} \int_{\Theta} e^{-\eta \ell_T(\theta)} d\hat{\mu}_{T-1}(\theta) && \leftarrow \theta_T = \int_{\Theta} \theta d\hat{\mu}_{T-1}(\theta) \\ &\leq W_{T-1} \exp(-\eta \ell_T(\theta_T)) && \leftarrow \eta\text{-exp-concavity} \\ &\leq W_0 \exp\left(-\eta \sum_{t=1}^T \ell_t(\theta_t)\right), && \leftarrow \text{induction} \end{aligned} \tag{3.9}$$

The second part of the proof to lower-bound  $W_T$  is however less straightforward. For simplicity, let us assume that  $\ell_t$  are continuous on  $\Theta$  (do the general case as exercise). Therefore the infimum is a minimum and let  $\theta^* \in \arg \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta)$  and define

$$\Theta_{\varepsilon} \stackrel{\text{def}}{=} \left\{ (1-\varepsilon)\theta^* + \varepsilon\theta, \quad \theta \in \Theta \right\}, \quad \varepsilon \in (0, 1).$$

Note that by convexity of  $\Theta$ ,  $\Theta_{\varepsilon} \subseteq \Theta$ . By expconcavity of  $\ell_t$ , we have for all  $t$  and all  $\theta = (1-\varepsilon)\theta^* + \varepsilon q$

$$e^{-\eta \ell_t(\theta)} \geq (1-\varepsilon)e^{-\eta \ell_t(\theta^*)} + \varepsilon e^{-\eta \ell_t(q)} \geq (1-\varepsilon)e^{-\eta \ell_t(\theta^*)}.$$

Therefore, for all  $\theta \in \Theta_{\varepsilon}$

$$e^{-\eta \sum_{t=1}^T \ell_t(\theta)} \geq (1-\varepsilon)^T e^{-\eta \sum_{t=1}^T \ell_t(\theta^*)}$$

Integrating both parts over  $\Theta_{\varepsilon}$  and using

$$\mu(\Theta_{\varepsilon}) = \int_{\Theta_{\varepsilon}} d\mu(\theta) = \int_{\Theta} d\mu((1-\varepsilon)\theta^* + \varepsilon\theta) \geq \int_{\Theta} \det(\varepsilon I_d) d\mu(\theta) = \int_{\Theta} \varepsilon^d d\mu(\theta) = \varepsilon^d \mu(\Theta)$$

we get

$$W_T \geq \int_{\Theta_\varepsilon} e^{-\eta \sum_{t=1}^T \ell_t(\theta)} d\mu(\theta) \geq \mu(\Theta) \varepsilon^d (1 - \varepsilon)^T e^{-\eta \sum_{t=1}^T \ell_t(\theta^*)}.$$

Combining with (3.9), using  $W_0 = \mu(\Theta)$ , taking the log and reorganizing the terms yields

$$R_T \stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(p_t) - \sum_{t=1}^T \ell_t(\theta^*) \leq \frac{d \log \frac{1}{\varepsilon} + T \log \frac{1}{1-\varepsilon}}{\eta}.$$

Optimizing  $\varepsilon = 1/(T+1)$  concludes the proof since

$$T \log \frac{1}{1-\varepsilon} = T \log \left(1 + \frac{1}{T}\right) \leq 1.$$

□

Though the nice theoretical result, this algorithm is complicated to implement because of the integral. In practice,  $\theta_t$  can be computed by using  $(1/T)$ -discretization grid of  $\Theta$  (bad complexity of order  $T^d$ !) or by using Monte-Carlo methods to approximate the integral of log-concave distributions (polynomial time algorithm). We will see in next lectures efficient algorithms with similar guarantees.

### Example: Portfolio selection

This algorithm was introduced first for the problem of portfolio selection by Cover [1991] and also known as Universal Portfolio in this framework. In the latter, given a initial capital  $\text{Cap}_0$  a trader repeatedly distributes her capital over  $d$  assets with the goal of maximizing the total return. At each round  $t = 1, \dots, T$ , the trader chooses an allocation  $p_t \in \Delta_d$ . Here,  $p_t(i)$  represents the share of capital innvested into asset  $i \in [d]$  at this round. At the end of the round, the returns—the ratios of the closing and opening prices in this round—are revealed in the form of  $x_t \in \mathbb{R}_+^d$  and the trader's capital is updated as

$$\text{Cap}_t = \text{Cap}_{t-1} \langle p_t, x_t \rangle.$$

By Cover, the performance of a strategy that selected portfolios  $(p_t)$  is quantified by comparing the final capital  $\text{Cap}_T = \text{Cap}_0 \prod_{t=1}^T \langle p_t, x_t \rangle$  against

$$\text{Cap}_T^* = \text{Cap}_0 \max_{p \in \Delta_d} \langle p, x_t \rangle,$$

the “idealized” final capital attained by the best “static” strategy constrained to select the same portfolio in all rounds. Due to the multiplicative structure, to maximize the capital it is natural to maximize the ratio  $\text{Cap}_T / \text{Cap}_T^*$ , which is equivalent to minimizing the regret

$$R_T = \sum_{t=1}^T \ell_t(p_t) - \min_{p \in \Delta_d} \sum_{t=1}^T \ell_t(p)$$

where  $\ell_t(p) \stackrel{\text{def}}{=} -\log \langle p, x_t \rangle$ . Noting that the loss is 1-exp-concave, continuous EWA achives the regret upper-bound of Thm 3.5, which yields

$$\text{Cap}_T \geq \frac{\text{Cap}_T^*}{e(T+1)^d},$$

which is optimal in the worst-case.

## Discretized EWA

**Setting:** general compact decision set,  $\beta$ -Hölder loss functions

In this section, we aim at designing a procedure for general compact decision set  $\Theta$ . We will assume for simplicity that  $\Theta \subset \mathbb{R}^d$  with  $\max_{\theta, \theta' \in \Theta} \|\theta - \theta'\| \leq D$ , where  $\|\cdot\|$  denotes the Euclidean norm. If the loss functions  $\ell_t$  are  $\beta$ -Hölder, i.e.,

$$|\ell_t(\theta) - \ell_t(\theta')| \leq c \|\theta - \theta'\|^\beta$$

there exists a simple solution: approximate  $\Theta$  with a finite discretization grid  $\Theta_\varepsilon$  and apply EWA on  $\Theta_\varepsilon$ . If  $\Theta$  or the losses are non-convex, one needs to use the random EWA (see Section 2.1.3) and bound the regret with high-probability. For convenience, we will assume  $\Theta$  and the loss functions  $\ell_t$  to be convex so that the algorithm can play convex combinations of points in  $\Theta_\varepsilon$  and all quantities are deterministic.

### Lemma 3.6

Let  $\varepsilon > 0$ . Let  $\Theta \subset \mathbb{R}^d$  such that  $\max_{\theta, \theta' \in \Theta} \|\theta - \theta'\| \leq D$ . Then, there exists  $\Theta_\varepsilon \subset \Theta$  such that

$$\text{Card}(\Theta_\varepsilon) \lesssim \left(\frac{D}{\varepsilon}\right)^d \quad \text{and} \quad \forall x \in \Theta, \exists x' \in \Theta_\varepsilon \quad \|\theta - \theta'\| \leq \varepsilon,$$

where  $\lesssim$  denotes a rough inequality (up to multiplicative constants and logarithmic terms).

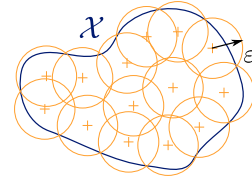
**Remark.** Remark that a set finite  $\Theta_\varepsilon$  which approximate  $\Theta$  at radius  $\varepsilon$ , is called an  $\varepsilon$ -covering of  $\Theta$ . The cardinal of the smallest  $\varepsilon$ -covering is called the covering number of  $\Theta$ . This cardinal is heavily used in theory to analyze the complexity of general spaces  $\Theta$ . It heavily differentiates parametric spaces with covering number of order  $(1/\varepsilon)^d$  with nonparametric spaces (spaces of functions) for which the logarithm of the covering number (or metric entropy) is of order  $(1/\varepsilon)^d$ .

*Proof sketch.* We only provide the high-level idea of the proof. First, by properties of the Lebesgue measure in  $d$ -dimension, denoting  $\mathcal{B}_2(r)$  is the  $\ell_2$ -ball of radius  $r > 0$ , we have

$$\text{Vol}(\mathcal{B}_2(r)) = \frac{\pi^{d/2}}{\Gamma(n/2 + 1)} r^d,$$

where  $\Gamma$  is the Euler's gamma function. Therefore,

$$\text{Vol}(\Theta) \leq \text{Vol}(\mathcal{B}_2(D/2)) = \left(\frac{D}{2\varepsilon}\right)^d \text{Vol}(\mathcal{B}_2(\varepsilon)),$$



and thus approximatively  $\left(\frac{D}{2\varepsilon}\right)^d$  balls of radius  $\varepsilon$  are sufficient to cover  $\Theta$ . □

### Theorem 3.7

### Discretized EWA

Let  $T \geq 1, \varepsilon, D > 0$ . Let  $\Theta$  be a compact convex subset of  $\mathbb{R}^d$  such that  $\max_{\theta, \theta' \in \Theta} \|\theta - \theta'\| \leq D$ . Let  $\Theta_\varepsilon$  be an  $\varepsilon$ -covering of  $\Theta$  with smallest cardinal. Then, for all sequences of  $\beta$ -Hölder convex losses  $\ell_1, \dots, \ell_T : \Theta \rightarrow [0, 1]$ , EWA played on the finite set of action  $\Theta_\varepsilon$  with optimized  $\eta$  satisfies the regret bound

$$R_T \stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta) \lesssim \sqrt{Td \left( \log D + \frac{1}{\beta} \log(cT) \right)}.$$

**Exercise 3.2.** Provide a bound on the expected regret for random EWA when the losses and the decision set are non-convex.

*Proof.* Let  $d = \text{Card}(\Theta_\varepsilon)$ . Let us order the elements of  $\Theta_\varepsilon = \{\theta(1), \dots, \theta(d)\}$ . Therefore, at time  $t \geq 1$ , EWA chooses a weight vector  $p_t \in \Delta_d$  and predict the weighted average  $\theta_t = \sum_{k=1}^d p_t(k)\theta(k) \in \Theta$ . Applying the regret bound of EWA, we get

$$\sum_{t=1}^T \sum_{k=1}^d p_t(k) \ell_t(\theta(k)) - \min_{1 \leq j \leq d} \sum_{t=1}^T \ell_t(\theta(j)) \leq 2\sqrt{T \log d}. \quad (3.10)$$

Let  $\theta^* \in \Theta$  and  $\theta(k^*) \in \Theta_\varepsilon$  such that  $\|\theta^* - \theta(k^*)\| \leq \varepsilon$ . Because the losses are  $\beta$ -Hölder and convex, we have

$$\begin{aligned} R_T &= \sum_{t=1}^T \ell_t(\theta_t) - \ell_t(\theta^*) \\ &\stackrel{\text{Convexity}}{\leq} \sum_{t=1}^T \sum_{k=1}^d p_t(k) \ell_t(\theta(k)) - \ell_t(\theta^*) \quad \leftarrow \theta_t = \sum_{k=1}^d p_t(k)\theta(k) \\ &\leq \sum_{t=1}^T \sum_{k=1}^d p_t(k) \ell_t(\theta(k^*)) - \ell_t(\theta(k^*)) + \sum_{t=1}^T |\ell_t(\theta(k^*)) - \ell_t(\theta^*)| \\ &\stackrel{(3.10)}{\leq} 2\sqrt{T \log d} + T \max_{1 \leq t \leq T} |\ell_t(\theta^*) - \ell_t(\theta(k^*))| \\ &\stackrel{\beta\text{-Hölder}}{\leq} 2\sqrt{T \log d} + cT\varepsilon^\beta \\ &\stackrel{\text{Lem. 3.6}}{\lesssim} \sqrt{Td \log \left( \frac{D}{\varepsilon} \right)} + cT\varepsilon^\beta. \end{aligned}$$

Optimizing  $\varepsilon^\beta = 1/cT$ , hence  $\varepsilon = (cT)^{-1/\beta}$ , we get

$$R_T \lesssim \sqrt{Td \left( \log D + \frac{1}{\beta} \log(cT) \right)}.$$

□

Though this algorithm is theoretically convenient since it can deal with general compact sets  $\Theta$  and general loss functions (which can be non-convex and non-differentiable). It suffers two considerable drawbacks:

- *computational complexity*: the algorithm needs to consider a discretization space of cardinal  $(X/\varepsilon)^d$  which is of order  $O(T^{d/\beta})$ . This is prohibitive in practice.
- *bad regret dependence on the dimension*: the regret bound is of order  $O(\sqrt{dT \log T})$ . We will see how to have no dependence on  $d$  when  $\Theta$  is bounded in  $\ell_2$ -norm.

**Example 3.1** (Prediction with expert advice (continued)). *In prediction with expert advice, a sequence of observations  $y_1, \dots, y_T \in [0, 1]$  is to be predicted with the help of  $d$  expert advice  $x_t(k) \in [0, 1]$  for  $1 \leq k \leq d$ . The learner predicts  $\hat{y}_t = \sum_{k=1}^d p_t(k)x_t(k)$  and suffers a loss  $\ell(\hat{y}_t, y_t)$ . If the loss function is convex and Lipschitz in its first argument we can apply Theorem ?? with  $\ell_t : p \mapsto \ell(p \cdot x_t, y_t)$ . For instance,*

with the absolute loss,  $G = 1$  and EG satisfies a bounded regret with respect to any fixed convex combination of experts:

$$\sum_{t=1}^T |\hat{y}_t - y_t| - \min_{p \in \Theta} \sum_{t=1}^T |p \cdot x_t - y_t| \leq 2\sqrt{T \log d}.$$

Hence, on the long run we perform as good as the best convex combination of the experts which may outperform the best expert. This may leads to much better performance than a simple EWA on the experts if

$$\min_{p \in \Theta} \sum_{t=1}^T |p \cdot x_t - y_t| \ll \min_{k \in [d]} \sum_{t=1}^T |x_t(k) - y_t|.$$

**Convex hull of finite point set** It is worth pointing out that the simplex decision set  $\Delta_d$  can be generalized with any convex hull of a finite point set  $S = \{\theta(1), \dots, \theta(d)\}$ :

$$\text{Conv}(S) = \left\{ \sum_{i=1}^d p_i \theta(i) : \forall i, p_i > 0 \text{ and } \sum_{i=1}^d p_i = 1 \right\}.$$

Transforming the loss functions, EWA can be applied to compete with such sets as shown by the theorem bellow.

### Theorem 3.8

Let  $T \geq 1$ . Let  $\Theta \subset \mathbb{R}^d$  be a convex set and  $S = \{\theta(1), \dots, \theta(d)\} \in \Theta^d$  with diameter  $D_1 \geq \max_{i,j} \|\theta(i) - \theta(j)\|_1$ . Let  $\ell_1, \dots, \ell_T : \Theta \rightarrow \mathbb{R}$  be an arbitrary sequence of convex differentiable losses with bounded gradient  $\max_{\theta \in \Theta} \|\nabla \ell_t(\theta)\|_\infty \leq G_\infty$ . Then, EWA applied with  $g_t = \nabla \tilde{\ell}_t$  where  $\tilde{\ell}_t : p \mapsto \ell_t \left( \sum_{i=1}^d p(i) \theta(i) \right)$  achieves the regret bound

$$R_T \stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta \in \text{Conv}(S)} \sum_{t=1}^T \ell_t(\theta) \leq 2G_\infty D_1 \sqrt{T \log d},$$

where  $\theta_t = \sum_{k=1}^d p_t(k) \theta(k)$

Such a trick can be used for instance to compete with the  $\ell_1$ -balls using  $S = \{\theta \in \mathbb{R}^d : \|\theta\|_1 = R, \|x\|_0 = 1\}$ . Since  $\ell_p$ -balls are contained into the  $\ell_1$ -ball (of possibly larger radius depending on  $p$ ) this can also be used to compete against any  $\ell_p$ -ball for  $p \geq 1$ . This trick was introduced by [Kivinen and Warmuth \[1997\]](#) for the EG $\pm$  forecaster.



## 4 Adversarial Bandits

In previous chapters, we considered the full-information feedback and the bandit feedback with stochastic loss functions. In *full information with finite decision set*  $\Theta = [K] \stackrel{\text{def}}{=} \{1, \dots, K\}$ , we saw the Random Exponentially Weighted Average (EWA) forecaster. It is defined as

$$p_t(k) = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_s(k)}}{\sum_{j=1}^K e^{-\eta \sum_{s=1}^{t-1} \ell_s(j)}}. \quad (\text{EWA})$$

and draws  $\theta_t = k$  with probability  $p_t(k)$ . If  $-\eta \ell_t(j) \leq 1$  (see the proof of EWA in first lecture), it satisfies the upper-bound:

$$\sum_{t=1}^T p_t \cdot \ell_t - \min_{1 \leq j \leq K} \sum_{t=1}^T \ell_t(j) \leq \eta \sum_{t=1}^T \sum_{k=1}^K p_t(k) \ell_t(k)^2 + \frac{\log K}{\eta}. \quad (*)$$

Since the decision  $\theta_t$  is random, we assume that  $\ell_t$  cannot depend on  $\theta_t$  but may depend on past information  $\sigma(p_1, \ell_1, x_1, \dots, x_{t-1}, p_t)$ . The above bound can be converted into a bound on the expected regret for well-calibrated learning rate  $\eta$

$$\mathbb{E}[R_T] = \mathbb{E} \left[ \sum_{t=1}^T \ell_t(\theta_t) - \min_{k \in [K]} \sum_{t=1}^T \ell_t(k) \right] \leq 2\sqrt{T \log K}.$$

In this chapter, we will see adversarial bandits: that is bandit feedback (only  $\ell_t(\theta_t)$  is observed at the end of round  $t$  by the player) with an adversarial sequence of loss function  $\ell_t$  (i.e., no stochastic assumptions). Note that we turn back to losses instead of rewards but we will come back to rewards whenever it makes the proof easier. Remember that the lower-bound on the regret in the worst-case is of order  $O(\sqrt{TK})$ .

### 4.1 Adversarial multi-armed bandits

We consider Setting 1.1 with bandit feedback, finite decision space  $\Theta = [K] \stackrel{\text{def}}{=} \{1, \dots, K\}$  and adversarial losses. To emphasize that the action is in  $[K]$ , we denote by  $k_t$  the action chosen by the player (instead of  $\theta_t$ ). We do not assume the loss functions  $\ell_t$  to be linear nor convex (the decision space is not). Similarly to Random EWA the chosen action  $k_t \in [K]$  is sampled randomly from a distribution  $p_t$  chosen at round  $t$  by the player. We will provide an algorithm called Exp3 inspired by EWA.

#### 4.1.1 Pseudo-regret bound

Let us denote the regret with respect to action  $k \in [K]$  by

$$R_T(k) \stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(k_t) - \sum_{t=1}^T \ell_t(k).$$

Instead of minimizing the *expected regret*  $\mathbb{E}[R_T] = \mathbb{E}[\max_k R_T(k)]$ , we will start with an easier objective, the *pseudo-regret* defined as

$$\bar{R}_T \stackrel{\text{def}}{=} \max_{k \in [K]} \mathbb{E}[R_T(k)] = \max_{k \in [K]} \mathbb{E} \left[ \sum_{t=1}^T \ell_t(k_t) - \sum_{t=1}^T \ell_t(k) \right]. \quad (\text{pseudo regret})$$

It is worth pointing out that the expectations are taken with respect to the randomness of the algorithm: the decisions  $k_t$  are random. We can distinguish two types of adversaries:

- *oblivious adversary*: all the loss functions  $\ell_1, \dots, \ell_T$  are chosen in advance before the game starts and do not depend on the past player decisions  $k_1, \dots, k_T$ . In this case, the losses  $\ell_t(k)$  are determinist and there is thus equality:  $\bar{R}_T = \mathbb{E}[R_T]$ .
- *adaptive adversary*: the loss function  $\ell_t$  at round  $t \geq 1$  may depend on past information  $\sigma(k_1, \dots, k_{t-1})$ . It is thus random. By Jensen's inequality  $\max_{k \in [K]} \mathbb{E}[R_T(k)] \leq \mathbb{E}[\max_{k \in [K]} R_T(k)]$  and thus  $\bar{R}_T \leq \mathbb{E}[R_T]$ .

**The EXP3 algorithm** Ideally, we would like to reuse our algorithm EWA that assigned weights

$$\forall k \in [K], \quad p_t(k) = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_s(k)}}{\sum_{j=1}^K e^{-\eta \sum_{s=1}^{t-1} \ell_s(j)}}. \quad (\text{EWA})$$

Unfortunately this is not possible since the player does not observe  $\ell_t(k)$  for  $k \neq k_t$ . The high-level idea of Exp3 is to replace  $\ell_t(k)$  with an unbiased estimate that is observed by the player. A first idea would be to use  $\ell_t(k)$  if we observe it and 0 otherwise:

$$g_t(k) = \begin{cases} \ell_t(k) & \text{if } k = k_t \quad \leftarrow \text{i.e., decision } k \text{ is observed} \\ 0 & \text{otherwise} \end{cases}.$$

However, this estimate is biased:

$$\mathbb{E}_{k_t \sim p_t} [g_t(k_t)] = p_t(k_t) \ell_t(k_t) \neq \ell_t(k_t).$$

In other words, the actions that are less likely to be chosen by the algorithm (small weight  $p_t(k)$ ) are more likely to be unobserved and incur 0 loss. We need to correct this phenomenon. Therefore we choose

$$g_t(k) = \frac{\ell_t(k)}{p_t(k)} \mathbb{1}\{k = k_t\}, \quad (4.1)$$

which leads to the algorithm EXP3 detailed below.

**EXP3**

Parameter:  $\eta > 0$

Initialize:  $p_1 = (\frac{1}{K}, \dots, \frac{1}{K})$

For  $t = 1, \dots, T$

- draw  $k_t \sim p_t$ ; incur loss  $\ell_t(k_t)$  and observe  $\ell_t(k_t) \in [0, 1]$ ;
- update for all  $k \in \{1, \dots, K\}$

$$p_{t+1}(k) = \frac{e^{-\eta \sum_{s=1}^t g_s(k)}}{\sum_{j=1}^K e^{-\eta \sum_{s=1}^t g_s(j)}}, \quad \text{where } g_s(k) = \frac{\ell_s(k)}{p_s(k)} \mathbb{1}\{k = k_s\}$$

Then applying the Inequality (\*) for EWA with the substituted losses  $g_t$ , we get the following theorem.

**Theorem 4.1**

Let  $T \geq 1$ . The pseudo-regret of EXP3 run with  $\eta = \sqrt{\frac{\log K}{KT}}$  is upper-bounded as:

$$\bar{R}_T \leq 2\sqrt{KT \log K}.$$

*Proof.* Apply EWA to the estimated losses  $g_t(j)$  that are completely observed (nonnegative but not bounded), we get from Inequality (\*) and taking the expectation:

$$\mathbb{E} \left[ \sum_{t=1}^T \langle p_t, g_t \rangle - \min_{j \in [K]} \sum_{t=1}^T g_t(j) \right] \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^T \mathbb{E} [\langle p_t, g_t \rangle^2]. \quad (4.2)$$

Now we compute the expectations. Denote by  $\mathcal{F}_{t-1} \stackrel{\text{def}}{=} \sigma(p_1, \ell_1, k_1, \dots, k_{t-1}, p_t, \ell_t)$  the past information available at round  $t$  for the adversary (which cannot use the randomness of  $k_t$  but can use  $p_t$ ). Note that  $\ell_t$  and  $p_t$  are  $\mathcal{F}_{t-1}$ -measurable by assumption. We have

$$\forall j \in [K] \quad \mathbb{E} [g_t(j) | \mathcal{F}_{t-1}] = \mathbb{E} \left[ \frac{\ell_t(j)}{p_t(j)} \mathbb{1}\{j = k_t\} | \mathcal{F}_{t-1} \right] = \sum_{k=1}^K p_t(k) \frac{\ell_t(j)}{p_t(j)} \mathbb{1}\{j = k\} = \ell_t(j)$$

thus the estimated losses are unbiased  $\mathbb{E}[g_t(j)] = \mathbb{E}[\ell_t(j)]$  and

$$\begin{aligned} \mathbb{E} [\langle p_t, g_t \rangle] &= \mathbb{E} \left[ \sum_{j=1}^K p_t(j) g_t(j) \right] = \mathbb{E} \left[ \sum_{j=1}^K p_t(j) \mathbb{E} [g_t(j) | \mathcal{F}_{t-1}] \right] \\ &= \mathbb{E} \left[ \sum_{j=1}^K p_t(j) \ell_t(j) \right] = \mathbb{E} [\mathbb{E} [\ell_t(k_t) | \mathcal{F}_{t-1}]] = \mathbb{E} [\ell_t(k_t)]. \end{aligned}$$

Therefore, we can lower-bound the left-hand side:

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=1}^T \langle p_t, g_t \rangle - \min_{j \in [K]} \sum_{t=1}^T g_t(j) \right] &\geq \max_{j \in [K]} \mathbb{E} \left[ \sum_{t=1}^T \langle p_t, g_t \rangle - \sum_{t=1}^T g_t(j) \right] \\ &= \max_{j \in [K]} \mathbb{E} \left[ \sum_{t=1}^T \ell_t(k_t) - \sum_{t=1}^T \ell_t(j) \right] = \bar{R}_T. \end{aligned}$$

On the other hand, the expectation of the right-hand side satisfies

$$\begin{aligned} \mathbb{E} [\langle p_t, g_t \rangle^2] &= \mathbb{E} \left[ \sum_{j=1}^K p_t(j) g_t(j)^2 \right] = \mathbb{E} \left[ \sum_{j=1}^K p_t(j) \mathbb{E} [g_t(j)^2 | \mathcal{F}_{t-1}] \right] \\ &= \mathbb{E} \left[ \sum_{j=1}^K \sum_{k=1}^K p_t(j) p_t(k) \left( \frac{\ell_t(j)}{p_t(j)} \mathbb{1}\{j = k\} \right)^2 \right] \\ &= \mathbb{E} \left[ \sum_{j=1}^K \sum_{k=1}^K p_t(k) \frac{\ell_t(j)^2}{p_t(j)} \mathbb{1}\{j = k\} \right] \end{aligned}$$

$$= \mathbb{E} \left[ \sum_{j=1}^K \ell_t(j)^2 \right] \leq K.$$

Substituting into Inequality (4.2) yields

$$\bar{R}_T \leq \frac{\log K}{\eta} + \eta KT.$$

and optimizing  $\eta = \sqrt{KT/(\log K)}$  concludes.  $\square$

The issue with the above regret bound is that it bounds the pseudo-regret and not the expected regret. This is because we have

$$\mathbb{E} \left[ \min_j \sum_{t=1}^T g_t(j) \right] \leq \min_j \mathbb{E} \left[ \sum_{t=1}^T g_t(j) \right] = \min_{j \in [K]} \mathbb{E} \left[ \sum_{t=1}^T \ell_t(j) \right]$$

but not

$$\mathbb{E} \left[ \min_j \sum_{t=1}^T g_t(j) \right] \not\leq \mathbb{E} \left[ \min_j \sum_{t=1}^T \ell_t(j) \right]. \quad (4.3)$$

Hence, controlling the cumulative loss against the best estimated action only controls the pseudo regret and not the true regret.

#### 4.1.2 High probability bound on the regret

**Gains versus losses** In this part, we will switch the analysis from losses  $\ell_t(k)$  to gains  $g_t(k) = 1 - \ell_t(k) \in [0, 1]$  because the core idea of the next algorithm is easier to see with gains. Remark that the loss and gain versions are symmetric via the transformation  $g_t(k) = 1 - \ell_t(k)$ . The regret in terms of gains is defined as

$$R_T \stackrel{\text{def}}{=} \max_{k \in [K]} \sum_{t=1}^T g_t(k) - \sum_{t=1}^T g_t(k_t).$$

Using EWA with full information from (\*), if  $\eta g_t(k) \leq 1$ , we also have for gains the inequality

$$\max_{1 \leq j \leq K} \sum_{t=1}^T g_t(j) - \sum_{t=1}^T \langle p_t, g_t \rangle \leq \eta \sum_{t=1}^T p_t \cdot g_t^2 + \frac{\log K}{\eta}, \quad \text{where} \quad p_t(k) = \frac{e^{\eta \sum_{s=1}^{t-1} g_s(k)}}{\sum_{j=1}^K e^{\eta \sum_{s=1}^{t-1} g_s(j)}}. \quad (4.4)$$

**High-level idea of EXP3.P** The high-level idea of the next algorithm is to ensure that the estimators  $\hat{g}_t(k)$  of the gains satisfy

$$\mathbb{E} \left[ \max_j \sum_{t=1}^T \hat{g}_t(j) \right] \geq \mathbb{E} \left[ \max_j \sum_{t=1}^T g_t(j) \right] \quad (4.5)$$

so that controlling the performance with respect to the estimated gains (left-hand side) also controls the performance with respect to the true gains (right-hand side). This was not the case of the estimators used for EXP3 (see (4.3)). To ensure (4.5), we add a bias term  $\beta$  to the estimators  $\hat{g}_t(k)$  as follows:

$$\hat{g}_t(k) \stackrel{\text{def}}{=} \frac{g_t(k) \mathbb{1}\{k = k_t\} + \beta}{p_t(k)} \quad (4.6)$$

In contrary to (4.1), the estimator is indeed biased

$$\mathbb{E}[\widehat{g}_t(k)|\mathcal{F}_{t-1}] = g_t(k) + \frac{\beta}{p_t(k)}, \quad (4.7)$$

where we recall that  $\mathcal{F}_{t-1} \stackrel{\text{def}}{=} \sigma(p_1, k_1, g_1, \dots, k_{t-1}, p_t, g_t)$  contains the information up to time  $t$  available to the environment. We have the following Lemma:

**Lemma 4.2**

For any  $\delta > 0$ , with probability  $1 - \delta$  and  $\beta \in (0, 1)$ ,

$$\sum_{t=1}^T \widehat{g}_t(j) \geq \sum_{t=1}^T g_t(j) - \frac{\log(1/\delta)}{\beta}.$$

*Proof.* Let  $\beta \in (0, 1)$ , from Markov's inequality, we have

$$\begin{aligned} \mathbb{P}\left(\sum_{t=1}^T \widehat{g}_t(j) \geq \sum_{t=1}^T g_t(j) - \frac{\log(1/\delta)}{\beta}\right) &= \mathbb{P}\left(\exp\left(\beta \sum_{t=1}^T (g_t(j) - \widehat{g}_t(j))\right) \geq \delta^{-1}\right) \\ &\leq \delta \mathbb{E}\left[\exp\left(\beta \sum_{t=1}^T (g_t(j) - \widehat{g}_t(j))\right)\right]. \end{aligned}$$

It only remains to upper-bound the expectation in the right-hand side by 1, which we do now. Since  $\beta \in (0, 1)$  and  $\widehat{g}_t(j) \geq \beta/p_t(j)$ , we have  $\beta(g_t(j) - \widehat{g}_t(j) + \beta/p_t(j)) \leq 1$ . Therefore, we can use the inequality  $e^x \leq 1 + x + x^2$  for  $x \leq 1$ , which entails

$$\begin{aligned} \mathbb{E}\left[\exp\left(\beta(g_t(j) - \widehat{g}_t(j))\right)\middle|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\exp\left(\beta\left(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\right)\right)\middle|\mathcal{F}_{t-1}\right] \exp\left(-\frac{\beta^2}{p_t(j)}\right) \\ &\leq \mathbb{E}\left[\left(1 + \beta\left(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\right) + \beta^2\left(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\right)^2\right)\middle|\mathcal{F}_{t-1}\right] e^{-\frac{\beta^2}{p_t(j)}} \\ &\stackrel{(4.7)}{=} \left(1 + \beta^2 \mathbb{E}\left[\left(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\right)^2\middle|\mathcal{F}_{t-1}\right]\right) e^{-\frac{\beta^2}{p_t(j)}} \end{aligned}$$

where the last equality is by (4.7) and because  $p_t(j)$  is  $\mathcal{F}_{t-1}$ -measurable. Now,

$$\begin{aligned} \mathbb{E}\left[\left(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\right)^2\middle|\mathcal{F}_{t-1}\right] &= \text{Var}\left(\widehat{g}_t(j)\middle|\mathcal{F}_{t-1}\right) = \text{Var}\left(\frac{g_t(j)\mathbb{1}\{j=k_t\}}{p_t(j)}\middle|\mathcal{F}_{t-1}\right) \\ &\leq \mathbb{E}\left[\left(\frac{g_t(j)\mathbb{1}\{j=k_t\}}{p_t(j)}\right)^2\middle|\mathcal{F}_{t-1}\right] \leq \mathbb{E}\left[\frac{\mathbb{1}\{j=k_t\}}{p_t(j)^2}\middle|\mathcal{F}_{t-1}\right] = \sum_{k=1}^K \frac{p_t(k)\mathbb{1}\{j=k\}}{p_t(j)^2} = \frac{1}{p_t(j)}. \end{aligned}$$

Substituting into the previous inequality and using  $1 + x \leq e^x$ , it yields

$$\mathbb{E}\left[\exp\left(\beta(g_t(j) - \widehat{g}_t(j))\right)\middle|\mathcal{F}_{t-1}\right] \leq \left(1 + \frac{\beta^2}{p_t(j)}\right) e^{-\beta^2/p_t(j)} \leq 1.$$

The proof is concluded by induction

$$\begin{aligned} \mathbb{E} \left[ \exp \left( \beta \sum_{t=1}^T (g_t(j) - \widehat{g}_t(j)) \right) \right] &= \mathbb{E} \left[ \underbrace{\mathbb{E} \left[ \exp \left( \beta (g_T(j) - \widehat{g}_T(j)) \right) \middle| \mathcal{F}_{T-1} \right]}_{\leq 1} \exp \left( \beta \sum_{t=1}^{T-1} (g_t(j) - \widehat{g}_t(j)) \right) \right] \\ &\leq \mathbb{E} \left[ \exp \left( \beta \sum_{t=1}^{T-1} (g_t(j) - \widehat{g}_t(j)) \right) \right] \leq \dots \leq 1. \end{aligned}$$

□

The issue with the estimators  $\widehat{g}_t(j) \in (0, +\infty)$  defined in Equation (4.6) is that they might be unbounded if the weights  $p_t(j)$  are close to zero. The condition  $\eta \widehat{g}_t(j) \leq 1$  which appeared in the proof of EWA cannot hold for any  $\eta > 0$ . Remark that this was not a problem for EXP3 with the preceding choice (4.1) because  $-\eta g_t(j) \leq 1$  (see the proof of EWA for details).

The next algorithm called EXP3.P, is close to EXP3 but ensures the weights do not vanish to zero by adding an exploration parameter  $\gamma > 0$ .

**EXP3.P**

Parameters:  $\eta > 0, \beta \in (0, 1), \gamma \in (0, 1)$

Initialize:  $p_1 = (\frac{1}{K}, \dots, \frac{1}{K})$

For  $t = 1, \dots, T$

- draw  $k_t \sim p_t$ ; receive gain  $g_t(k_t) = 1 - \ell_t(k_t)$  and observe  $g_t(k_t) \in [0, 1]$ ;
- update for all  $k \in \{1, \dots, K\}$

$$p_{t+1}(k) = (1 - \gamma) \frac{e^{\eta \sum_{s=1}^t \widehat{g}_s(k)}}{\sum_{j=1}^K e^{\eta \sum_{s=1}^t \widehat{g}_s(j)}} + \frac{\gamma}{K},$$

$$\text{where } \widehat{g}_s(k) = \frac{g_s(k) \mathbb{1}_{\{k=k_s\}} + \beta}{p_s(k)}.$$

The weights  $p_t(k)$  of EXP3.P are necessary larger than  $\gamma/K$  and thus  $|\eta g_t(j)| \leq 1$  as soon as  $\eta(1 + \beta)K/\gamma \leq 1$ . We get the following high-probability bound on the regret.

**Theorem 4.3**

For well-chosen parameters  $\gamma \in (0, 1)$ ,  $\beta \in (0, 1)$  and  $\eta > 0$  satisfying  $\eta(1 + \beta)K/\gamma \leq 1$ , for any  $\delta > 0$ , the EXP3.P algorithm achieves

$$R_T \leq 6\sqrt{TK \log K} + \sqrt{\frac{TK}{\log K}} \log(1/\delta).$$

with probability at least  $1 - \delta$ .

Remark that the above bound leads to a bound on the expected regret, with the choice  $\delta = 1/T$  it yields

$$\mathbb{E}[R_T] \leq 6\sqrt{TK \log K} + \sqrt{\frac{TK}{\log K}} \log(T) + 1$$

The logarithmic dependency on  $T$  can even be removed using  $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq \varepsilon) d\varepsilon$ .

*Proof of Theorem 4.3.* Defining the weights that would assign EXP3,

$$q_t(j) \stackrel{\text{def}}{=} \frac{e^{\eta \sum_{s=1}^{t-1} \widehat{g}_s(j)}}{\sum_{k=1}^K e^{\eta \sum_{s=1}^{t-1} \widehat{g}_s(k)}},$$

we get from Inequality (4.4) applied with  $\widehat{g}_t(j)$ ,

$$\max_{j \in [K]} \sum_{t=1}^T \widehat{g}_t(j) \leq \sum_{t=1}^T q_t \cdot \widehat{g}_t + \eta \sum_{t=1}^T q_t \cdot \widehat{g}_t^2 + \frac{\log K}{\eta}.$$

where we used  $\eta \widehat{g}_t(j) \leq 1$  because  $\eta(1 + \beta)K/\gamma \leq 1$ . Now, we use that  $p_t \stackrel{\text{def}}{=} (1 - \gamma)q_t + \gamma/K$ , which entails  $q_t = (p_t - \gamma/K)/(1 - \gamma) \leq p_t/(1 - \gamma)$ . Substituting into the above inequality

$$(1 - \gamma) \max_{j \in [K]} \sum_{t=1}^T \widehat{g}_t(j) \leq \sum_{t=1}^T p_t \cdot \widehat{g}_t + \eta \sum_{t=1}^T p_t \cdot \widehat{g}_t^2 + \frac{\log K}{\eta}. \quad (4.8)$$

But by definition of  $\widehat{g}_t$ ,

$$p_t \cdot \widehat{g}_t = \sum_{j=1}^K p_t(j) \widehat{g}_t(j) = \sum_{j=1}^K (g_t(j) \mathbb{1}\{j = k_t\} + \beta) = g_t(k_t) + K\beta.$$

and since  $p_t(j) \widehat{g}_t(j) \leq (1 + \beta)$ ,

$$\sum_{t=1}^T p_t \cdot \widehat{g}_t^2 \leq (1 + \beta) \sum_{j=1}^K \sum_{t=1}^T \widehat{g}_t(j) \leq K(1 + \beta) \max_{j \in [K]} \sum_{t=1}^T \widehat{g}_t(j) \leq \frac{\gamma}{\eta} \max_{j \in [K]} \sum_{t=1}^T \widehat{g}_t(j).$$

Therefore, substituting into Inequality (4.8) gives

$$(1 - \gamma) \max_{j \in [K]} \sum_{t=1}^T \widehat{g}_t(j) \leq \sum_{t=1}^T g_t(k_t) + K\beta T + \gamma \max_{j \in [K]} \sum_{t=1}^T \widehat{g}_t(j) + \frac{\log K}{\eta},$$

where we used  $(1 + \beta)K \leq \gamma/\eta$ . Reorganizing, we get

$$(1 - 2\gamma) \max_{j \in [K]} \sum_{t=1}^T \widehat{g}_t(j) \leq \sum_{t=1}^T g_t(k_t) + K\beta T + \frac{\log K}{\eta}.$$

Using Lemma 4.2 together with a union bound (to have it for all  $j \in [K]$ ), we have with probability  $1 - \delta$

$$(1 - 2\gamma) \left( \max_{j \in [K]} \sum_{t=1}^T g_t(j) - \frac{\log(K/\delta)}{\beta} \right) \leq \sum_{t=1}^T g_t(k_t) + K\beta T + \frac{\log K}{\eta},$$

and thus reorganizing and choosing  $\gamma \stackrel{\text{def}}{=} 2\eta K \geq \eta(1 + \beta)K$ ,

$$\max_{j \in [K]} \sum_{t=1}^T g_t(j) - \sum_{t=1}^T g_t(k_t) \leq K\beta T + \frac{\log K}{\eta} + \frac{\log(K/\delta)}{\beta} + 4\eta K T.$$

The proof is concluded by optimizing  $\eta \stackrel{\text{def}}{=} (1/2)\sqrt{(\log K)/KT}$  and  $\beta \stackrel{\text{def}}{=} \sqrt{(\log K)/(KT)}$ .  $\square$

## 4.2 Contextual adversarial multi-armed bandits

Here, we consider the adversarial bandit framework in which the learner has access to some external information (context) before making its decision. We turn back to the loss version of the game.

### 4.2.1 Adversarial multi-armed bandits with experts

We now consider prediction with expert advice in the bandit framework. The setting is the same as the one described in Figure 1.1, but at the beginning of each round  $t \geq 1$ , some experts  $i = 1, \dots, N$  propose recommendations  $h_t(i) \in [K]$ . These recommendations may be random and may depend on past actions  $k_s$ ,  $s \leq t-1$  and past observations  $\ell_s(k_s)$ . The loss of each expert is given by the loss of the chosen decision  $\ell_t(h_t(i))$  but only  $\ell_t(k_t)$  is observed by the learner. The goal of the learner is then to be competitive with the best expert on a long run. To do so, it minimizes the pseudo-regret

$$R_T^{\text{exp}} \stackrel{\text{def}}{=} \max_{i=1, \dots, N} \mathbb{E} \left[ \sum_{t=1}^T \ell_t(k_t) - \sum_{t=1}^T \ell_t(h_t(i)) \right]$$

with respect to the experts. In order to bound the pseudo-regret, one could consider experts as the set of arms and use EXP3. This would give a bound of order  $\sqrt{TN \log N}$ . However it does not take into account the information on the reward of all experts that choose the same action  $h_t(i) = k_t$ .

#### EXP4

Parameter:  $\eta > 0$

Initialize:  $q_1 = (\frac{1}{N}, \dots, \frac{1}{N})$ .

For each round  $t = 1, \dots, n$

1. Get expert advice  $h_t(1), \dots, h_t(N) \in [K]$
2. Draw an expert  $i_t$  with probability distribution  $q_t \in \Delta_N$
3. Choose decision  $k_t = h_t(i_t)$
4. Compute the estimated loss for each decision

$$g_t(k) = \frac{\ell_t(k)}{p_t(k)} \mathbb{1}_{\{k = k_t\}},$$

where  $p_t \stackrel{\text{def}}{=} \sum_{i=1}^N q_t(i) \delta_{\ell_t(i)} \in \Delta_K$ .

5. Compute the estimated loss of the experts component-wise  $g_t(h_t(i))$
6. Update the probability distribution over the experts component-wise

$$q_{t+1}(i) = \frac{\exp \left( -\eta \sum_{s=1}^t g_s(h_s(i)) \right)}{\sum_{j=1}^N \exp \left( \eta \sum_{s=1}^t g_s(h_s(j)) \right)}, \quad \forall 1 \leq i \leq N.$$

#### Theorem 4.4

EXP4 with  $\eta = \sqrt{\log N / (KT)}$  satisfies  $R_T^{\text{exp}} \leq 2\sqrt{TK \log N}$ .

Similarly to the variant EXP3.P, we can define a variant EXP4.P to bound the regret with high probability (and thus the expected regret). Furthermore, the above algorithm (and theorem) can be extended to the case where expert advice are distributions  $h_t(i) \in \Delta_K$ . The algorithm is the same by sampling  $k_t$  according to  $h_t(i_t)$  and assigning to expert  $i$  the loss  $\sum_{k=1}^K h_t(i)(k)g_t(k)$ .



*Proof.* We can apply the analysis of EXP to a learner using distribution  $q_t$  over  $N$  actions (here experts) with (full-information) losses  $g_t(h_t(i))$  for  $i \in \{1, \dots, N\}$ . We get from Inequality (\*)

$$\mathbb{E} \left[ \sum_{t=1}^T \sum_{i=1}^N q_t(i) \cdot g_t(h_t(i)) - \min_{1 \leq j \leq N} \sum_{t=1}^T g_t(h_t(j)) \right] \leq \eta \sum_{t=1}^T \sum_{i=1}^N \mathbb{E} \left[ q_t(i) g_t(h_t(i))^2 \right] + \frac{\log N}{\eta}. \quad (4.9)$$

Remark that  $k_t = h_t(i)$  with probability  $q_t(i)$  so that,  $k_t$  follows the distribution  $p_t \stackrel{\text{def}}{=} \sum_{i=1}^N q_t(i) \delta_{h_t(i)}$  knowing the past information  $\mathcal{F}_{t-1} \stackrel{\text{def}}{=} \sigma(h_1, i_1, k_1, \dots, i_{t-1}, k_{t-1}, h_t)$ . Now, similarly to the proof of EXP3, we compute the expectations. We have for all  $k \in [K] \stackrel{\text{def}}{=} \{1, \dots, K\}$

$$\mathbb{E} \left[ g_t(k) \middle| \mathcal{F}_{t-1} \right] = \mathbb{E} \left[ \frac{\ell_t(k)}{p_t(k)} \mathbb{1}\{k = k_t\} \middle| \mathcal{F}_{t-1} \right] = \sum_{j=1}^K p_t(j) \frac{\ell_t(k)}{p_t(k)} \mathbb{1}\{k = j\} = \ell_t(k),$$

and thus for all  $i \in \{1, \dots, N\}$

$$\mathbb{E} \left[ g_t(h_t(i)) \middle| \mathcal{F}_{t-1} \right] = \ell_t(h_t(i)), \quad (4.10)$$

and

$$\begin{aligned} \mathbb{E} \left[ \sum_{i=1}^N q_t(i) \cdot g_t(h_t(i)) \middle| \mathcal{F}_{t-1} \right] &= \sum_{i=1}^N q_t(i) \mathbb{E} \left[ g_t(h_t(i)) \middle| \mathcal{F}_{t-1} \right] = \sum_{i=1}^N q_t(i) \ell_t(h_t(i)) \\ &= \mathbb{E} \left[ \ell_t(h_t(i)) \middle| \mathcal{F}_{t-1} \right] = \mathbb{E} \left[ \ell_t(k_t) \middle| \mathcal{F}_{t-1} \right]. \end{aligned} \quad (4.11)$$

Furthermore,

$$\mathbb{E} \left[ g_t(h_t(i))^2 \middle| \mathcal{F}_{t-1} \right] = \sum_{k=1}^K p_t(k) \left( \frac{\ell_t(h_t(i))}{p_t(h_t(i))} \right)^2 \mathbb{1}\{k = h_t(i)\} = \frac{\ell_t(h_t(i))^2}{p_t(h_t(i))} \leq \frac{1}{p_t(h_t(i))},$$

and

$$\sum_{i=1}^N q_t(i) \mathbb{E} \left[ g_t(h_t(i))^2 \middle| \mathcal{F}_{t-1} \right] \leq \sum_{i=1}^N \frac{q_t(i)}{p_t(h_t(i))} = \mathbb{E} \left[ \frac{1}{p_t(h_t(i))} \middle| \mathcal{F}_{t-1} \right] = \mathbb{E} \left[ \frac{1}{p_t(k_t)} \middle| \mathcal{F}_{t-1} \right] = \sum_{k=1}^K \frac{p_t(k)}{p_t(k)} = K. \quad (4.12)$$

Substituting (4.10), (4.11), and (4.12) into Inequality (4.9) and lower-bounding the expected regret with the pseudo-regret, we get

$$\begin{aligned} \bar{R}_T^{\text{exp}} &\stackrel{\text{def}}{=} \max_{1 \leq i \leq N} \mathbb{E} \left[ \sum_{t=1}^T \ell_t(k_t) - \ell_t(h_t(i)) \right] \\ &\stackrel{(4.10), (4.11)}{=} \max_{1 \leq i \leq N} \mathbb{E} \left[ \sum_{t=1}^T \sum_{i=1}^N q_t(i) g_t(h_t(i)) - g_t(h_t(i)) \right] \\ &\stackrel{\text{Jensen}}{\leq} \mathbb{E} \left[ \sum_{t=1}^T \sum_{i=1}^N q_t(i) g_t(h_t(i)) - \min_{1 \leq i \leq N} g_t(h_t(i)) \right] \\ &\stackrel{(4.9)}{\leq} \eta \sum_{t=1}^T \sum_{i=1}^N \mathbb{E} \left[ q_t(i) g_t(h_t(i))^2 \right] + \frac{\log N}{\eta} \\ &\stackrel{(4.12)}{\leq} \eta K T + \frac{\log N}{\eta}. \end{aligned}$$

Optimizing  $\eta$  concludes the proof.  $\square$

### 4.2.2 Adversarial multi-armed bandits with side information

A natural extension of the previous setting is by adding side (or contextual) information: this is called contextual bandits. It arises in most applications such as recommendation systems or online advertisement. The side information can then be the cookies of a new user to which we need to recommend a product.

Assume that for each time step  $t \geq 1$ , before doing its prediction  $k_t$  the learner observes a context  $x_t$  in a finite set  $\mathcal{X}$  of contexts. The learner must then learn the best mapping  $g : \mathcal{X} \rightarrow [K]$  and is evaluated with the contextual pseudo-regret:

$$R_T^{\mathcal{X}} \stackrel{\text{def}}{=} \max_{g: \mathcal{X} \rightarrow [K]} \mathbb{E} \left[ \sum_{t=1}^T \ell_t(k_t) - \sum_{t=1}^T \ell_t(g(x_t)) \right].$$

Similarly, to the stochastic setting, if  $\mathcal{X}$  is finite, a simple algorithm consists in running a different copy  $\text{EXP3}(c)$  of EXP3 for each context  $c \in \mathcal{X}$ . We denote by  $\mathcal{X}$ -EXP3 this algorithm. At each time step  $t \geq 1$ , the learner uses  $\text{EXP}(x_t)$  to make the prediction. The following theorem follows from Theorem 4.1.

#### Theorem 4.5

The contextual pseudo-regret of  $\mathcal{X}$ -EXP3 is upper-bounded as:

$$R_T^{\mathcal{X}} \leq 2\sqrt{T|\mathcal{X}|K \log K}.$$

*Proof.* Applying the proof of the pseudo-regret bound of EXP3 for each instance  $x \in \mathcal{X}$ :

$$\max_{j \in [K]} \mathbb{E} \left[ \sum_{t=1}^T (\ell_t(k_t) - \ell_t(j)) \mathbb{1}\{x_t = x\} \right] \leq 2\sqrt{K(\log K)T_x},$$

where  $T_x = \sum_{t=1}^T \mathbb{1}\{x_t = x\}$ . Summing over  $x \in \mathcal{X}$ ,

$$\sum_{x \in \mathcal{X}} \max_{j \in [K]} \mathbb{E} \left[ \sum_{t=1}^T (\ell_t(k_t) - \ell_t(j)) \mathbb{1}\{x_t = x\} \right] \leq 2 \sum_{x \in \mathcal{X}} \sqrt{K(\log K)T_x} \stackrel{\text{Jensen}}{\leq} 2\sqrt{|\mathcal{X}|K(\log K)T}$$

where the last inequality is by using the concavity of the square root together with  $\sum_{s \in \mathcal{X}} T_s = T$ . The proof is concluded by remarking that the left-hand side is the contextual pseudo-regret.  $\square$

Similarly to the classical lower-bound  $O(\sqrt{TK})$ , a lower-bound of order  $\sqrt{|\mathcal{X}|KT}$  holds under the assumption that a significant proportion of the contexts are used at least a constant fraction of the  $T$  rounds. The above bound is nice but the dependency  $|\mathcal{X}|$  might be annoying if  $\mathcal{X}$  is large.

**Exercise 4.1.** Generalize the above algorithm and upper-bound when the context-space is continuous and the loss functions are  $\beta$ -Hölder in the contexts.

### Competing against the best context set

In some cases, one may want to combine bandit algorithms. For example, we could have in hand different context set  $\mathcal{X}$ . For each of these sets  $\mathcal{X}$ , we can bound the pseudo-regret  $R_T^{\mathcal{X}}$  using Theorem 4.5 with  $\mathcal{X}$ -EXP3 of Section 4.2.2, but we would like to find the best set  $\mathcal{X}$ . To do so, we may want to

combine with EXP4 different instances of  $\mathcal{X}$ -EXP3, each using its own context set  $\mathcal{X}$ . We can then combine the bounds of Theorem 4.4 and Theorem 4.5 to ensure we are competing with the best possible context set  $\mathcal{X}$ . In this case, each instance of  $\mathcal{X}$ -EXP3 does not observe their own choice of action but the action chosen by EXP4 which follows a different distribution. The bound of Theorem 4.4 is valid but the regrets of the experts cannot be bounded using Theorem 4.5. It is however possible to use a variant of EXP4 to combine bandit algorithms by adding an exploration parameter. We then lose however in the rate of the regret bound which is then of order

$$\max_{\mathcal{X}} R_T^{\mathcal{X}} \leq O\left(T^{2/3} (\max_{\mathcal{X}} |\mathcal{X}| K \log K)^{1/3} \sqrt{\log M}\right)$$

where  $M$  is the number of context sets  $\mathcal{X}$ . We refer to Section 4.2.1 of [Bubeck et al. \[2012\]](#) for more details on this application.

### 4.3 Online convex optimization with bandit feedback

In the previous sections, we saw how to deal with bandit feedback when the decision set is finite  $\Theta = [K]$ . The goal of this section is to extend to bandit feedback the general framework of online convex optimization of Figure 1.1 for any compact and convex decision space  $\Theta \subset \mathbb{R}^d$ . Here, at each round  $t \geq 1$ , the learner picks  $\theta_t \in \Theta$  and observes  $\ell_t(\theta)$ . When gradients are observed, one may choose  $\theta_t$  by following online gradient descent (OGD):

$$\theta_{t+1} = \Pi_{\Theta}(\theta_t - \eta \nabla \ell_t(\theta_t)). \quad (\text{OGD})$$

Theorem 2.7 showed that the regret of OGD for well-chosen learning rate  $\eta > 0$  is upper-bounded as:

$$R_T = \sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta) \leq DG\sqrt{T},$$

where  $D \geq \max_{\theta, \theta' \in \Theta} \|\theta - \theta'\|$  and  $G \geq \max_{\theta \in \Theta} \|\nabla \ell_t(\theta)\|$ . Here, we aim at answering the following question:

“How to adapt OGD to the bandit setting, when  $\ell_t(\theta_t)$  is observed but not  $\nabla \ell_t(\theta_t)$ ?”

Similarly to EXP3, the idea is to replace the gradient in OGD with estimators. That is to try to find a random variable  $g_t$  that satisfies  $\mathbb{E}[g_t] \approx \nabla \ell_t(\theta_t)$  and which can be computed with the observation at hand.

#### 4.3.1 Estimating gradients from value observations

**One-dimensional example.** Let us first, consider the one-dimensional case, let  $\ell : \mathbb{R} \rightarrow \mathbb{R}$  be a one-dimensional differentiable loss function. Is it possible to design an estimator of  $\ell'(\theta)$  by evaluating  $\ell$  in a single point? It turns out that this is the case. Indeed, the derivative in  $\theta$  can be written as

$$\ell'(\theta) = \lim_{\delta \rightarrow 0} \frac{\ell(\theta + \delta) - \ell(\theta - \delta)}{2\delta}.$$

Thus, denoting  $\xi$  a Rademacher random variable (which equals 1 with probability 1/2 and -1 otherwise) and defining  $g(\theta) = \frac{1}{\delta} \xi \ell(\theta + \xi \delta)$ , we have

$$\mathbb{E}[g(\theta)] = \frac{1}{2} \times \frac{\ell(\theta + \delta)}{\delta} + \frac{1}{2} \times \frac{-\ell(\theta - \delta)}{\delta} \approx \ell'(\theta)$$

if  $\delta$  is sufficiently small. Therefore, for a small value of  $\delta$ ,  $g(\theta)$  is approximatively an unbiased estimator of  $\ell'(\theta)$ . Note that  $g(\theta)$  can be computed by evaluating the function  $\ell$  a single time at a random point close to  $\theta$ . On the other side, one may compute

$$\text{Var}(g(\theta)) = \left( \frac{\ell(\theta + \delta) + \ell(\theta - \delta)}{2\delta} \right)^2 \approx \frac{\ell(\theta)^2}{\delta^2},$$

which may explode as  $\delta \rightarrow 0$ . Thus,  $\delta$  will need to be chosen carefully to optimize a bias-variance trade-off.

**Multi-dimensional case.** We now formally extend the above point-wise estimator to the multi-dimensional case. To do so, we first need to define  $\widehat{\ell}_t$  a smoothed version of the loss by: for all  $\theta \in \Theta$

$$\widehat{\ell}_t(\theta) = \mathbb{E}_v[\ell_t(\theta + \delta v)],$$

where  $v \sim \text{Unif}(\mathbb{B})$  is a uniform random-variable over the Euclidean unit ball  $\mathbb{B} := \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ . The following lemma follows from Lipschitzness of  $\ell_t$ , and states that  $\ell_t$  is a  $\delta G$ -approximation of  $\widehat{\ell}_t$ .

**Lemma 4.6**

Let  $\delta > 0$ . Then,  $|\widehat{\ell}_t(\theta) - \ell_t(\theta)| \leq \delta G$  for all  $\theta \in \mathbb{R}^d$ .

Now, define the estimator

$$g_t = \frac{d}{\delta} \ell_t(\theta_t + \delta u_t) u_t \quad \text{with } u_t \sim \text{Unif}(\mathbb{S}) \quad (4.13)$$

where  $\mathbb{S} = \{x \in \mathbb{R}^d : \|x\| = 1\}$  is the Euclidean unit sphere. Then, the following lemma states that  $g_t$  is an unbiased estimator of  $\widehat{\ell}_t$ .

**Lemma 4.7**

Let  $\widehat{\ell}_t(\theta_t) = \mathbb{E}_v[\ell_t(\theta_t + \delta v)]$  with  $v \sim \text{Unif}(\mathbb{B})$  and  $g_t = \frac{d}{\delta} \ell_t(\theta_t + \delta u_t) u_t$  with  $u_t \sim \text{Unif}(\mathbb{S})$ . Then,

$$\mathbb{E}_u[g_t] = \nabla \widehat{\ell}_t(\theta_t).$$

*Proof.* The proof is a consequence of Stokes' theorem that states that for any continuously differentiable function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ ,

$$\int_{\mathbb{B}} \nabla f(x) dx = \int_{\mathbb{S}} f(x) x dx,$$

applied with  $f : x \mapsto \ell_t(\theta_t + \delta x)$  and  $\nabla f(x) = \delta \nabla \ell_t(\theta_t + \delta x)$ . Denoting by  $\mu$  the Lebesgue measure, we then have

$$\mathbb{E}_u[g_t] = \frac{1}{\mu(\mathbb{S})} \int_{\mathbb{S}} \frac{d}{\delta} \ell_t(\theta_t + \delta u) u du = \frac{d}{\mu(\mathbb{S})} \int_{\mathbb{B}} \nabla \ell_t(\theta_t + \delta v) dv = \frac{d\mu(\mathbb{B})}{\mu(\mathbb{S})} \mathbb{E}_v[\nabla \ell_t(\theta_t + \delta v)] = \frac{d\mu(\mathbb{B})}{\mu(\mathbb{S})} \nabla \widehat{\ell}_t(\theta_t),$$

where the last inequality is by interchanging the derivative and the integral. The proof is then completed by noting that  $d\mu(\mathbb{B}) = \mu(\mathbb{S})$ .  $\square$

### 4.3.2 OGD with bandit feedback

Noting that the above estimator  $g_t$  can be computed with a single evaluation of  $\ell_t$  at the random point  $\theta_t + \delta u_t$ , we may define the Online Gradient Descent algorithm with bandit feedback.

**Online Gradient Descent with bandit feedback (OGD without gradient)**

Parameters:  $\eta > 0$   $\delta > 0$

Initialize:  $\theta_1 \in \Theta_\delta$  arbitrarily chosen in  $\Theta_\delta := \{\theta \in \Theta \text{ s.t. } \forall u \in \mathbb{S}, \theta + \delta u \in \Theta\}$

For  $t = 1, \dots, T$

- Draw  $u_t$  uniformly at random in the unit sphere.
- Set  $\widehat{\theta}_t := \theta_t + \delta u_t$  a random perturbation of the current point  $\theta_t$
- Play  $\widehat{\theta}_t$
- Incur and observe loss  $\ell_t(\widehat{\theta}_t) \in [-1, 1]$
- Estimate the gradient with

$$g_t := \frac{d}{\delta} \ell_t(\widehat{\theta}_t) u_t$$

- Update

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \eta g_t) .$$

where  $\Pi_{\Theta_\delta}$  is the Euclidean projection onto  $\Theta_\delta$ .

We have the following theorem.

**Theorem 4.8**

If the losses  $\ell_t$  are  $G$ -Lipschitz and take values in  $[-B, B]$ , then OGD without gradient with parameters  $\delta = \min \left\{ D, \frac{1}{2} \sqrt{\frac{dBD}{G}} T^{-1/4} \right\}$  and  $\eta = \delta D / (dB\sqrt{T})$  satisfies the expected regret bound

$$\mathbb{E} \left[ \sum_{t=1}^T \ell_t(\widehat{\theta}_t) - \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta) \right] \leq 2dB\sqrt{T} + 4\sqrt{dBGDT}^{3/4} .$$

The above regret bound is sub-optimal and more sophisticated algorithms achieve  $O(\sqrt{T})$  but with worse dependency on  $d$  and higher computational cost.

*Proof.* We define

$$\theta^* = \arg \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta) \quad \text{and} \quad \theta_\delta^* = \Pi_{\Theta_\delta}(\theta^*) .$$

Then, by definition of  $\Theta_\delta := \{\theta \in \Theta \text{ s.t. } \theta + \delta u \in \Theta \text{ for all } u \in \mathbb{S}\}$ , we have  $\|\theta^* - \theta_\delta^*\| \leq \delta$  (left as exercise). Thus, if the losses are  $G$ -Lipschitz

$$\begin{aligned} R_T &:= \mathbb{E} \left[ \sum_{t=1}^T \ell_t(\widehat{\theta}_t) - \sum_{t=1}^T \ell_t(\theta^*) \right] \\ &\leq \mathbb{E} \left[ \sum_{t=1}^T \ell_t(\widehat{\theta}_t) - \sum_{t=1}^T \ell_t(\theta_\delta^*) \right] + \delta TG \quad \leftarrow \text{because } \|\theta^* - \theta_\delta^*\| \leq \delta \end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E} \left[ \sum_{t=1}^T \ell_t(\theta_t) - \sum_{t=1}^T \ell_t(\theta_\delta^*) \right] + 2\delta TG \quad \leftarrow \text{because } \|\widehat{\theta}_t - \theta_t\| \leq \delta \\
&\leq \mathbb{E} \left[ \sum_{t=1}^T \widehat{\ell}_t(\theta_t) - \sum_{t=1}^T \widehat{\ell}_t(\theta_\delta^*) \right] + 4\delta TG \quad \leftarrow \text{because } |\widehat{\ell}_t(\theta_t) - \ell_t(\theta_t)| \leq \sup_{\|v\| \leq 1} |\ell_t(\theta_t + \delta v) - \ell_t(\theta_t)| \leq \delta G
\end{aligned} \tag{4.14}$$

where  $\widehat{\ell}_t(\theta) = \mathbb{E}_v[\ell_t(\theta + \delta v)]$  with  $v \sim \text{Unif}(\mathbb{B})$  are the smoothed versions of the losses.

Now, recall that the algorithm runs OGD with  $g_t$  in place of the gradients:

$$\theta_{t+1} \leftarrow \Pi_{\Theta_\delta}(\theta_t - \eta g_t)$$

Defining the pseudo-loss  $h_t(\theta) = \widehat{\ell}_t(\theta) + (g_t - \nabla \widehat{\ell}_t(\theta_t))^\top \theta$ , we can see that

$$\nabla h_t(\theta_t) = \nabla \widehat{\ell}_t(\theta_t) + g_t - \nabla \widehat{\ell}_t(\theta_t) = g_t.$$

Therefore, the algorithm actually runs OGD on the losses  $h_t$  and thus satisfies the OGD regret bound (see Theorem 2.7)

$$\sum_{t=1}^T h_t(\theta_t) - \sum_{t=1}^T h_t(\theta_\delta^*) \leq \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|g_t\|^2.$$

Furthermore, by construction of the gradient estimator, we have  $\mathbb{E}_{u_t}[g_t] = \nabla \widehat{\ell}_t(\theta_t)$ , which yields

$$\mathbb{E}_{u_t}[h_t(\theta_t)] = \widehat{\ell}_t(\theta_t) \quad \text{and} \quad \mathbb{E}_{u_t}[h_t(\theta_\delta^*)] = \widehat{\ell}_t(\theta_\delta^*).$$

Thus taking the expectation in the previous regret bound entails

$$\mathbb{E} \left[ \sum_{t=1}^T \widehat{\ell}_t(\theta_t) - \sum_{t=1}^T \widehat{\ell}_t(\theta_\delta^*) \right] = \mathbb{E} \left[ \sum_{t=1}^T h_t(\theta_t) - \sum_{t=1}^T h_t(\theta_\delta^*) \right] \leq \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathbb{E}[\|g_t\|^2]. \tag{4.15}$$

Combining the two bounds (4.14) and (4.15) that we have proved, we get

$$R_T \leq \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathbb{E}[\|g_t\|^2] + 4\delta TG.$$

Then, since  $|\ell_t(\theta)| \leq B$  for all  $\theta \in \Theta$ ,

$$\|g_t\|^2 = \left( \frac{d}{\delta} \ell_t(\widehat{\theta}_t) \right)^2 \leq \frac{d^2 B^2}{\delta^2}.$$

This finally yields the regret

$$R_T \leq \frac{D^2}{2\eta} + \frac{\eta d^2 B^2 T}{2\delta^2} + 4\delta GT = \frac{dBD}{\delta} \sqrt{T} + 4\delta GT \leq 2dB\sqrt{T} + 4\sqrt{dBGDT}^{3/4}$$

for the choices of  $\delta$  and  $\eta$ . □

**Coordinate gradient descent.** In some situations, the full gradient  $\nabla \ell_t(\theta_t)$  is too costly to be computed at each step, and one only observes the gradient of a single coordinate  $k_t \sim \text{Unif}(d)$ . In this case, one may use OGD (or OMD, or EG) with the gradient estimator

$$g_t = d \nabla \ell_t(\theta_t)_{k_t} e_{k_t},$$

where  $\{e_k\}$  are the elements of the canonical basis in  $\mathbb{R}^d$ . Here,

$$\mathbb{E}_{k_t}[g_t] = \sum_{k=1}^d d \nabla \ell_t(\theta_t)_k e_k \mathbb{P}(k_t = k) = \nabla \ell_t(\theta_t),$$

and

$$\mathbb{E}_{k_t}[\|g_t\|^2] = \sum_{k=1}^d d^2 \nabla \ell_t(\theta_t)_k^2 \|e_k\|^2 \mathbb{P}(k_t = k) = d \|\nabla \ell_t(\theta_t)\|^2 \leq dG^2.$$

Then, following the proof of OGD, for the update rule  $\theta_{t+1} = \Pi_{\Theta}(\theta_t - \eta g_t)$ , we get the regret bound in expectation

$$\mathbb{E}[R_T] = \mathbb{E} \left[ \sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta) \right] \leq \frac{\eta d^2 G^2 T}{2} + \frac{D^2}{2\eta} = GD\sqrt{dT},$$

where we optimized over  $\eta = D/G(dT)^{-1/2}$  in the last equality. The regret bound of OGD is deteriorated by a factor  $\sqrt{d}$  compared to OGD with full gradient observation.

## 5 Stochastic Multi-armed Bandits

During the last few chapters, we have reviewed the framework for comprehensive information. We designed algorithms to minimize regret for the different decision spaces  $\Theta$  and loss assumptions  $f_t$ . Most of the algorithms were based on variations in the exponentially weighted average forecaster or online gradient descent. We also found some links with game theory, including the Blackwell approach, two-player zero-sum games, and calibration.

In this chapter, we consider the bandit setting, when the player only observes the performance of  $f_t(\theta_t)$  but not  $f_t(\theta)$  for  $\theta \neq \theta_t$ . We will start by providing fundamental results for stochastic bandits with finitely many actions, also called  $K$ -armed bandits which basically corresponds to  $\Theta = \{1, \dots, K\}$  and i.i.d. loss functions  $f_t$ . For a thorough introduction to stochastic bandits we refer the interested student to the monographs [Bubeck et al. \[2012\]](#) or [Lattimore and Szepesvári \[2020\]](#).

### 5.1 Setting: stochastic bandit with finitely many actions

We state here the setting of stochastic bandits with finitely many actions (also called multi-armed bandit) and fix notations that we will use.

A multi-armed bandit problem is a sequential decision problem defined by a finite set of actions  $\Theta \stackrel{\text{def}}{=} \{1, \dots, K\}$  also called *arms*. We assume that there are  $K$  unknown sequences  $X_{i,1}, X_{i,2}, \dots$  of rewards in  $[0, 1]$  associated with each arm  $i = 1, \dots, K$ . At each round, the player makes a decision by pulling an arm  $k_t$  in  $\Theta$  and observes the corresponding reward<sup>1</sup>  $X_{k_t,t}$ . The objective of the player is to minimize his cumulative regret:

$$R_T \stackrel{\text{def}}{=} \max_{k=1,\dots,K} \sum_{t=1}^T X_{k,t} - \sum_{t=1}^T X_{k_t,t}.$$

In stochastic bandits, we generally assume the sequences to be i.i.d. Each arm  $k = 1, \dots, K$  is associated an unknown probability distribution  $\nu_k$  over  $[0, 1]$  and  $X_{k,t} \sim \nu_k$ . We also denote

$$\mu_k \stackrel{\text{def}}{=} \mathbb{E}[X_{k,t}], \quad \text{and} \quad \mu^* \in \arg \max_{k=1,\dots,K} \{\mu_k\}.$$

The player aims at finding the arm with the highest mean reward  $\mu_k$  as quickly as possible. The setting is summarized in Setting 5.1. Note that we retrieve the setting of online optimization (Setting 1.1) with the notation  $X_{k,t} \stackrel{\text{def}}{=} 1 - f_t(k)$  with i.i.d. loss functions.

---

<sup>1</sup>In the bandit community, it is more common to consider rewards rather than losses.



*Unknown parameters:*  $K$  probability distributions  $\nu_1, \dots, \nu_K$  on  $[0, 1]$

At each time step  $t = 1, \dots, T$

- the player chooses an action  $k_t \in \Theta = \{1, \dots, K\}$ ;
- given  $k_t$ , the environment draws the reward  $X_{k_t, t} \sim \nu_{k_t}$ ;
- the player only observes the feedback  $X_{k_t, t}$ .

#### Setting 5.1: Setting of stochastic bandit with finitely many actions

Multi-armed bandits have several concrete historical applications in a variety of fields, including ad placement, clinical trials, source routing or game AI. The name bandit refers to the “slot machine” in casinos, and the bandit problem corresponds to a player that inserts coins into different machines and tries to maximize his payoff. The finite arms bandit settings we consider are simple enough to be analyzed and the algorithms can often be generalized to more realistic settings including for example contextual bandits.

**Remark.** Assume that all arms  $\nu_k \sim \mathcal{B}(1/2)$  for  $k = 1, \dots, K$ . Then,  $\mathbb{E}[X_{k,t}] = 1/2$  and  $\mathbb{E}[X_{k_t, t}] = 1/2$ . But because of fluctuations of random walks, the expected magnitude of the maximum rewards is of order

$$\mathbb{E} \left[ \max_{k=1, \dots, K} \sum_{n=1}^T X_{k,n} \right] \approx \sqrt{T \log K}.$$

Therefore, in this case though all arms are optimal, the expected regret is of order  $\sqrt{T \log K}$ . We will thus consider a more quantity in the stochastic framework called the pseudo-regret which corresponds to competing with the best action in expectation, rather than the optimal action on the sequence of realized rewards.

#### Definition 5.1

The pseudo-regret is defined as

$$\bar{R}_T \stackrel{\text{def}}{=} T\mu^* - \mathbb{E} \left[ \sum_{t=1}^T \mu_{k_t} \right],$$

where we recall  $\mu_k = \mathbb{E}[X_{k,t}]$ .

#### Pseudo-regret

Remark that the pseudo-regret is upper-bounded by the expected regret  $\bar{R}_T \leq \mathbb{E}[R_T]$ . It is thus harder to design algorithms for the true regret but we will focus here on the pseudo-regret.

**Useful notation** In the following, we will denote by  $\hat{\mu}_k(s)$  the empirical mean of rewards obtained after pulling arm  $k$   $s$  times. Let us also denote for all arms  $k = 1, \dots, K$  by

$$\Delta_k \stackrel{\text{def}}{=} \mu^* - \mu_k \quad \text{and} \quad N_k(t) \stackrel{\text{def}}{=} \sum_{s=1}^t \mathbb{1}_{k_s=k},$$

respectively the suboptimal gap of arm  $k$  and the number of times the arm  $k$  was pulled by the player before time  $t$ . Then, the pseudo-regret can be rewritten

$$\bar{R}_T = \left( \sum_{k=1}^K \mathbb{E}[N_k(t)] \right) \mu^* - \mathbb{E} \left( \sum_{k=1}^K N_k(t) \mu_k \right) = \sum_{k=1}^K \Delta_k \mathbb{E}[N_k(t)]. \quad (5.1)$$

## Hoeffding's inequality and subGaussian random variable

We recall here a concentration result on the sum of independent random variables that will be useful for the analysis. Let us first recall the definition of a subGaussian random variable.

### Definition 5.2

### SubGaussian r.v.

We say that a real-valued random variable  $X$  is  $\sigma^2$ -subGaussian if

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E} \left[ e^{\lambda(X - \mathbb{E}[X])} \right] \leq e^{\frac{\lambda^2 \sigma^2}{2}}.$$

We are now ready to state our concentration inequality.

### Lemma 5.1

### Hoeffding's lemma

Any random variable supported on  $[a, b]$  is  $(b - a)^2/4$ -subGaussian.

*Proof.* By replacing  $X$  with  $(X - \mathbb{E}[X])/(b - a)$ , without loss of generality, we can assume that  $\mathbb{E}[X] = 0$  and  $b - a = 1$ . By convexity of  $x \mapsto \exp(\lambda x)$ , we have taking the expectation

$$\mathbb{E}[e^{\lambda X}] \leq \frac{b - \mathbb{E}[X]}{b - a} e^{\lambda a} + \frac{\mathbb{E}[X] - a}{b - a} e^{\lambda b} = b e^{\lambda(b-1)} + (1 - b) e^{\lambda b} = e^{\lambda b} (1 - b + b e^{-\lambda}) = e^{f(\lambda)},$$

where  $f(\lambda) = \lambda b + \log(1 - b + b e^{-\lambda})$ . It remains to show that  $f(\lambda) \leq \lambda^2/8$ . Computing the derivatives:

$$f'(\lambda) = b - \frac{b e^{-\lambda}}{1 - b + b e^{-\lambda}} \quad \text{and} \quad f''(\lambda) = \frac{(1 - b) b e^{-\lambda}}{(1 - b + b e^{-\lambda})^2} \leq \frac{1}{4}$$

where the last inequality is because  $\sqrt{xy} \leq \frac{x^2 + y^2}{2}$  with  $x = \sqrt{1 - b}$  and  $y = b e^{-\lambda}$ . Now, since  $f(0) = 0$  and  $f'(0) = 0$ , from Taylor's theorem, there exists  $0 \leq t \leq 1$  such that

$$f(\lambda) = f(0) + t f'(0) + \frac{\lambda^2}{2} f''(t\lambda) \leq \frac{\lambda^2}{8}.$$

This concludes the proof. □

### Proposition 5.2

### Hoeffding's Inequality

If  $X_1, \dots, X_n$  are independent random variables almost surely in  $[a, b]$  then for all  $\delta \in (0, 1)$  we have

$$\mathbb{P} \left\{ \sum_{k=1}^n X_k - \mathbb{E} \left[ \sum_{k=1}^n X_k \right] \geq (b - a) \sqrt{\frac{n}{2} \log \frac{1}{\delta}} \right\} \leq \delta,$$

or equivalently for all  $\varepsilon > 0$

$$\mathbb{P} \left\{ \sum_{k=1}^n X_k - \mathbb{E} \left[ \sum_{k=1}^n X_k \right] \geq \varepsilon \right\} \leq \exp \left( - \frac{2\varepsilon^2}{n(b - a)^2} \right).$$

*Proof.*

$$\begin{aligned}
\mathbb{P} \left\{ \sum_{t=1}^n X_k - \mathbb{E}[X_k] \geq \varepsilon \right\} &= \mathbb{P} \left\{ e^{\lambda \left( \sum_{t=1}^n X_k - \mathbb{E}[X_k] \right)} \geq e^{\lambda \varepsilon} \right\} \stackrel{\text{Markov's ineq.}}{\leq} e^{-\lambda \varepsilon} \mathbb{E} \left[ e^{\lambda \left( \sum_{t=1}^n X_k - \mathbb{E}[X_k] \right)} \right] \\
&\stackrel{\text{independence}}{=} e^{-\lambda \varepsilon} \prod_{t=1}^n \mathbb{E} \left[ e^{\lambda (X_k - \mathbb{E}[X_k])} \right] \stackrel{\text{Hoeffding's lem.}}{\leq} e^{-\lambda \varepsilon + \frac{n\lambda^2}{8(b-a)^2}}.
\end{aligned}$$

Minimizing  $-\lambda \varepsilon + \frac{n\lambda^2}{8(b-a)^2}$  in  $\lambda$  yields  $\lambda = 4(b-a)^2 \varepsilon / n$ . Substituting in the previous upper-bound proves the second inequality. The first inequality is obtained by solving  $\delta = \exp(-2\varepsilon^2 / (n(b-a)^2))$  in  $\varepsilon$ .  $\square$

## 5.2 Explore-Then-Commit (ETC)

Contrary to the full information we examined earlier, the player only observes the rewards of the chosen actions. He must therefore make a trade-off between exploration and exploitation. The first bandit algorithm that we consider is Explore Then Commit (ETC). It consists in first performing an exploration phase of  $mK$  length in which each arm is pulled  $m \geq 1$  times. Then it exploits by pulling the arm with the best empirical reward for the remaining rounds. It is summarized in Algorithm 5.1.

*Parameter:*  $m \geq 1$ .

### 1. Exploration

- For rounds  $t = 1, \dots, mK$  explore by drawing each arm  $m$  times.
- Compute for each arm  $k$  its empirical mean of rewards obtained by pulling arm  $k$   $m$  times

$$\hat{\mu}_k(m) = \frac{1}{m} \sum_{s=1}^{Km} X_{k,t} \mathbb{1}\{k_t = k\}.$$

**2. Exploitation:** keep playing the best arm  $\arg \max_k \hat{\mu}_k(m)$  for the remaining rounds  $t = mK + 1, \dots, T$ .

**Algorithm 5.1:** Explore-Then-Commit (ETC)

### Theorem 5.3

If  $1 \leq m \leq T/K$  then

$$\bar{R}_T \leq m \sum_{k=1}^K \Delta_k + (T - mK) \sum_{k=1}^K \Delta_k \exp(-m\Delta_k^2).$$

Thm 6.1, ?

*Proof.* Assume without loss of generality that the first arm is optimal, i.e.,  $\mu_1 = \mu^*$  and  $\Delta_1 = 0$ . From (5.1), we have

$$\bar{R}_T = \sum_{k=1}^K \Delta_k \mathbb{E}[N_k(t)].$$

Let  $k \geq 2$  be a suboptimal arm. Then, the arm  $k$  is selected  $m$  times during the exploration phase, and  $T - mK$  times during the exploitation if  $k$  is selected, which implies  $\hat{\mu}_k(m) \geq \hat{\mu}_1(m)$ . Therefore,

$$\mathbb{E}[N_k(t)] \leq m + (T - mK) \mathbb{P}(\hat{\mu}_k(m) \geq \hat{\mu}_1(m))$$

Now, we can use Hoeffding's inequality to control the probability in the right-hand side. Indeed  $\hat{\mu}_k(m)$  and  $\mu_1$  are respectively the empirical averages of  $m$  i.i.d. random variables in  $[0, 1]$  of mean  $\mu_k$  and

$\mu_1 = \mu^*$ . Therefore,

$$\begin{aligned} \mathbb{P}(\widehat{\mu}_k(m) - \widehat{\mu}_1(m) \geq 0) &= \mathbb{P}(\widehat{\mu}_k(m) - \widehat{\mu}_1(m) - \mu_k + \mu_1 \geq -\mu_k + \mu_1) \\ &= \mathbb{P}(\widehat{\mu}_k(m) - \widehat{\mu}_1(m) - \mu_k + \mu_1 \geq \Delta_k) \\ &= \mathbb{P}(m\widehat{\mu}_k(m) - m\widehat{\mu}_1(m) - m\mu_k + m\mu_1 \geq m\Delta_k) \\ &\leq \exp(-m\Delta_k^2). \end{aligned}$$

This implies

$$\bar{R}_T \leq m \sum_{k=1}^K \Delta_k + (T - mK) \sum_{k=1}^K \Delta_k \exp(-m\Delta_k^2).$$

□

The bound in Theorem 5.3 illustrates the trade-off between exploration and exploitation. If  $m$  is large, the exploration is too long and the first term  $m \sum_{k=1}^K \Delta_k$  yields a large regret. On the other hand, for small  $m$ , there is a large probability to choose a suboptimal arm during the exploitation and the other term might lead to a large regret. The question is which value of  $m$  should we choose?

To have an idea, we will consider the case  $K = 2$ , in which case the bound is

$$\bar{R}_T \leq m\Delta_2 + T\Delta_2 \exp(-m\Delta_2^2).$$

#### Corollary 5.4

If  $K = 2$  and  $m = \max\{1, \lceil \log(T\Delta_2^2)/\Delta_2^2 \rceil\}$ , then

$$\bar{R}_T \leq \Delta_2 + \frac{1 + \log(T\Delta_2^2)}{\Delta_2}.$$

The above bound is of order  $O((\log T)/\Delta_2)$ . Such bounds are called distribution-dependent because they heavily depend on the distributions  $\nu_k$  via  $\Delta_k$ . If  $\Delta_2 \rightarrow 0$ , it explodes. However, we also have from (5.1) that  $\bar{R}_T \leq \Delta_2 T$ . Therefore, in the worst case for any value of  $\Delta_2$ , Corollary 5.4 yields to the worst-case bound

$$\bar{R}_T \leq \min \left\{ T\Delta_2, \Delta_2 + \frac{1 + \log(T\Delta_2^2)}{\Delta_2} \right\} \lesssim \sqrt{T \log T}. \quad (5.2)$$

The above bound is close to be optimal. Yet, the issue is that the parameter  $m$  depends on  $\Delta_2$  and  $T$ . If the dependence on  $T$  can be dealt with a doubling-trick it is harder to optimize it in  $\Delta_2$ . Furthermore, when there are more than two arms, one might want to explore differently the arms. The upper-confidence-bound algorithm that we will see next does not have these issues.

**Exercise 5.1.** Show that it is possible to achieve the worst-case bound on the pseudo-regret of order  $O(T^{2/3})$  by optimizing  $m$  independently of  $\Delta$  (only with  $T$ ).

**Exercise 5.2.** Generalize the results of Theorem 5.3 and 5.4 when the rewards are not-bounded but  $\sigma^2$ -sub-Gaussian, i.e., for all  $\lambda > 0$

$$\mathbb{E} \left[ \exp(\lambda(X_{k,t} - \mathbb{E}[X_{k,t}])) \right] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right).$$

### 5.3 Upper-Confidence-Bound (UCB)

The UCB (Upper-Confidence-Bound) algorithm is a very popular bandit algorithm that has several advantages over ETC. It does not rely on an initial exploration phase but explores on the fly as rewards are observed. It explores and exploits sequentially throughout the experience. Unlike ETC, it does not require knowledge of gaps and behaves well when there are more than two arms.

To perform exploration and face uncertainty, the UCB algorithm is based on the *optimism principle*.

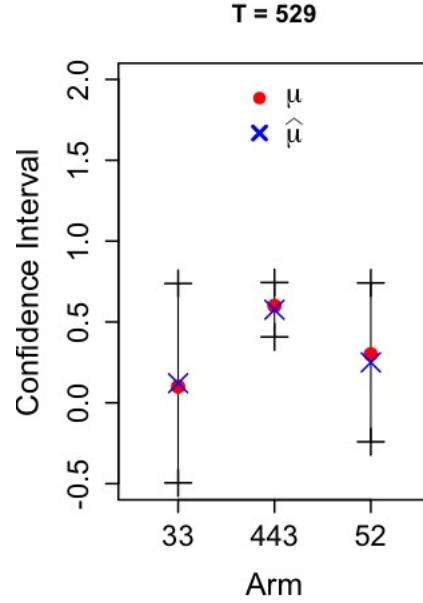
For each arm  $k$ , it builds a confidence interval on its expected reward based on past observation

$$I_k(t) = [LCB_k(k), UCB_k(t)]$$

where  $LCB$  is the Lower-Confidence-Bound and  $UCB$  is the Upper-Confidence-Bound. Then it is *optimistic* acting as if the best possible rewards are the real rewards and chooses the next arm accordingly

$$k_t \in \arg \max_{k \in \{1, \dots, K\}} UCB_k(t).$$

In other words, it pulls the arm with the higher upper-confidence-bound. An example of how UCB works with three arms of means  $\mu_1 = 0.1$ ,  $\mu_2 = 0.6$  and  $\mu_3 = 0.3$  is plotted in the Figure on the right. The best arm is pulled more often (see x-axis for number of times arms are selected) and his confidence interval is smaller.



The only question is how to design the upper-confidence-bounds. This is based on Hoeffding's inequality. Since the rewards are i.i.d. the distribution of  $\hat{\mu}_k(s)$  is equal to the distribution of

$$\frac{1}{s} \sum_{s'=1}^s X_{k,s'},$$

with mean  $\mu_k$ . Therefore, from Hoeffding's inequality, we have for all arms  $k \in \{1, \dots, K\}$ , for all  $s \geq 1$  and all  $\delta \in (0, 1)$

$$\mathbb{P} \left\{ \mu_k \geq \hat{\mu}_k(s) + \sqrt{\frac{\log \frac{1}{\delta}}{2s}} \right\} \leq \delta. \quad (5.3)$$

where  $\hat{\mu}_k(t)$  is the empirical reward of arm  $k$  after pulling it  $t$  times. Therefore, it is reasonable to choose the upper-confidence bound

$$UCB_t(k) = \begin{cases} \infty & \text{if } N_k(t-1) = 0 \\ \hat{\mu}_k(N_k(t-1)) + \sqrt{\frac{2 \log t}{N_k(t-1)}} & \text{otherwise} \end{cases}$$

The UCB algorithm is described in Algorithm 5.2.

**Initialization** For rounds  $t = 1, \dots, K$  pull arm  $k_t = t$

**For**  $t = K + 1, \dots, T$ , choose

$$k_t \in \arg \max_{k \in \{1, \dots, K\}} \left\{ \hat{\mu}_k(N_k(t-1)) + \sqrt{\frac{2 \log t}{N_k(t-1)}} \right\},$$

and get reward  $X_{k_t, t}$ .

**Algorithm 5.2:** Upper-Confidence-Bound (UCB)

**Theorem 5.5**

If the distributions  $v_k$  have supports all included in  $[0, 1]$  then for all  $k$  such that  $\Delta_k > 0$

$$\mathbb{E}[N_k(T)] \leq \frac{8 \log T}{\Delta_k^2} + 2.$$

In particular, this implies that the pseudo-regret of UCB is upper-bounded as

$$\bar{R}_T \leq 2K + \sum_{k: \Delta_k > 0} \frac{8 \log T}{\Delta_k}.$$

**Remark.** Let us make some remarks about the about upper-bound on the pseudo-regret.

- UCB has a regret bound of order

$$\bar{R}_T \lesssim \frac{K \log T}{\Delta},$$

where  $\Delta = \min_{i: \Delta_i > 0} \Delta_i$ . Once again, using that the regret incurred from pulling arm  $k$  cannot be larger than  $T \Delta_k$ , this distribution-dependent upper-bound can be transformed into a distribution-free bound of order  $\bar{R}_T \lesssim \sqrt{TK \log T}$ . We leave this proof as an exercise.

- This bound is close to optimal since the lower bound is of order  $O(\sqrt{KT})$ . There exists modification of UCB to get rid of the extra logarithmic term. For instance, the MOSS algorithm (Minimax Optimal Strategy in the Stochastic Case) achieves

$$\bar{R}_T \lesssim \min \left\{ \sqrt{TK}, \frac{K}{\Delta} \log \frac{T \Delta^2}{K} \right\},$$

however it depends on the smallest gap  $\Delta$  only and not on all gaps  $\Delta_i$ .

- The assumption that the rewards are independent between arms can be relaxed.
- The assumption that the rewards are in  $[0, 1]$  can be relaxed to a sub-Gaussian assumption.
- While a bound on the pseudo-regret is interesting, one would actually want a bound with high probability on

$$\hat{R}_T \stackrel{\text{def}}{=} T \mu^* - \sum_{t=1}^T \mu_{k_t, t}.$$

Using Hoeffding's inequality to control  $\hat{R}_T$  with  $\bar{R}_T = \mathbb{E}[\hat{R}_T]$  would yield an additional term of order  $\sqrt{T}$  due to fluctuations which would dominate  $O(K \log T / \Delta)$ . Obtaining a bound of order  $O(K \log T / \Delta)$  on  $\hat{R}_T$  is a challenging problem and not achieved by UCB. Some strategies using the knowledge of  $T$  can satisfy it.

*Proof.* Without loss of generality let us assume that the first arm is optimal, i.e.,  $\mu_1 = \mu^*$  and  $\Delta_1 = 0$ . We show below that if  $k_t = k$ , then at least one of the following three inequalities must be satisfied

$$\mu^* > \widehat{\mu}_1(N_1(t-1)) + \sqrt{\frac{2 \log t}{N_1(t-1)}} \quad \leftarrow \mu^* \text{ larger than UCB} \quad (\text{i})$$

$$\mu_k < \widehat{\mu}_k(N_k(t-1)) - \sqrt{\frac{2 \log t}{N_k(t-1)}} \quad \leftarrow \mu_k \text{ smaller than LCB} \quad (\text{ii})$$

$$N_k(t-1) \leq \frac{8 \log t}{\Delta_k^2} \quad \leftarrow k \text{ not played enough yet} \quad (\text{iii})$$

Indeed, otherwise assume that the three inequalities are all false than

$$\begin{aligned} \widehat{\mu}_1(N_1(t-1)) + \sqrt{\frac{2 \log t}{N_1(t-1)}} &\geq \mu^* && \leftarrow \text{not (i)} \\ &\geq \mu_k + \Delta_k && \leftarrow \text{Def of } \Delta_k \\ &> \mu_k + 2\sqrt{\frac{2 \log t}{N_k(t-1)}} && \leftarrow \text{not (iii)} \\ &\geq \widehat{\mu}_k(N_k(t-1)) + \sqrt{\frac{2 \log t}{N_k(t-1)}} && \leftarrow \text{not (ii)}. \end{aligned}$$

This contradicts the fact that  $k_t = k$  (see Algorithm 5.2). Therefore, denoting  $u = \lfloor \frac{8 \log T}{\Delta_k^2} \rfloor$ , we have

$$\begin{aligned} \mathbb{E}[N_k(T)] &= \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{k_t=k}] = u + \sum_{t=u+1}^T \mathbb{P}\{k_t = k \text{ and (iii) is false}\} \\ &= u + \sum_{t=u+1}^T \left( \mathbb{P}\{(\text{i}) \text{ or } (\text{ii})\} \right) \\ &\leq u + \sum_{t=u+1}^T \left( \mathbb{P}\{(\text{i})\} + \mathbb{P}\{(\text{ii})\} \right). \end{aligned} \quad (5.4)$$

Therefore, it suffices to control the probabilities of (i) and (ii), which we do now. At round  $t \geq 1$ ,

$$\begin{aligned} \mathbb{P}\{(\text{i})\} &\leq \mathbb{P}\left\{ \exists s \in \{1, \dots, t-1\}, \text{ such that } \mu^* > \widehat{\mu}_1(s) + \sqrt{\frac{2 \log t}{s}} \right\} \\ &\leq \sum_{s=1}^t \mathbb{P}\left\{ \mu^* > \widehat{\mu}_1(s) + \sqrt{\frac{\log(1/t^{-4})}{2s}} \right\} \\ &\stackrel{(5.3)}{\leq} \sum_{s=1}^t t^{-4} = t^{-3}. \end{aligned}$$

By symmetry, the same applies for  $\mathbb{P}\{(\text{ii})\} \leq t^{-3}$ . Combining into (5.4), it concludes the proof of the first inequality

$$\mathbb{E}[N_k(T)] \leq \frac{8 \log T}{\Delta_k^2} + 2 \sum_{t=u+1}^T t^{-3} \leq \frac{8 \log T}{\Delta_k^2} + 2.$$

The upper-bound on the pseudo-regret follows from (5.1).  $\square$

## 5.4 Other algorithms

Other algorithms exist in the literature. The best known are  $\varepsilon$ -greedy and Thompson sampling.

### 5.4.1 $\varepsilon$ -greedy

The idea of  $\varepsilon$ -greedy is very simple: first choose a parameter  $\varepsilon \in (0, 1)$ , then at each round, select the arm with the highest empirical mean with probability  $\varepsilon$  (i.e., be greedy), and explore by playing a random arm with probability  $\varepsilon$ . It works quite well in practice and is used in many application because of its simple implementation (in particular in reinforcement learning). Choosing  $\varepsilon \approx K/(\Delta^2 T)$  yields to an upper-bound of order  $K \log T/\Delta^2$ . However it requires the knowledge of  $\Delta$ .

### 5.4.2 Thompson Sampling

Thomson sampling was the first algorithm proposed for bandits by Thomson in 1933. It assumes a uniform prior over the expected rewards  $\mu_i \in (0, 1)$ , then at each round  $t \geq 1$ , for each arm  $\pi_{k,t}$ , it

- computes  $\widehat{v}_{k,t}$  the posterior distribution of the rewards of arm  $k$  given the rewards observed so far;
- samples  $\theta_{k,t} \sim \widehat{v}_{k,t}$  independently;
- selects  $k_t \in \arg \max_{k \in \{1, \dots, K\}} \theta_{k,t}$ .

Thomson sampling has a similar upper-bound of order  $O(K \log T/\Delta)$  than the one achieved by UCB. An advantage over UCB is the possibility of incorporating easily prior knowledge on the arms.

## 5.5 Lower bounds for multi-armed bandit

In this section, we essentially state that the regret bound of UCB

$$\bar{R}_T \lesssim \min \left\{ \sqrt{KT \log T}, \sum_{k: \Delta_k > 0} \frac{\log T}{\Delta_k} \right\}$$

is close to optimal regret for multi-armed bandit.

### 5.5.1 Distribution-free lower bound

The next theorem shows that the previous results are not improvable (up to log factors).

#### Theorem 5.6

#### Lower bound

For any forecaster, there exists distributions  $v_1, \dots, v_K$  such that

$$\bar{R}_T \gtrsim \sqrt{KT}.$$

The complete proof can be found in [Bubeck et al. \[2012\]](#). We only present here the high-level idea of the proof. We design the adversary as follows: it generates i.i.d. Bernoulli rewards such that  $\mathbb{E}[X_{k,t}] = \frac{1}{2}$  for all  $k \in \{1, \dots, K\}$  except for one arm  $k^*$  where  $\mathbb{E}[X_{k^*,t}] = \frac{1}{2} + \varepsilon$ .



- Fact 1: to distinguish between a Bernoulli of parameter  $1/2$  and a Bernoulli of parameter  $1/2 + \varepsilon$ , one needs  $1/\varepsilon^2$  samples. This result can be proved formally by using Pinsker's inequality. The intuition goes as follows. From the Central Limit Theorem (or the distribution of a Binomial random variable), after  $T_k$  observations of an arm  $k$ , one can estimate its mean with an error of order  $1/\sqrt{T_k}$ . In other words, to estimate it with an error smaller than  $\varepsilon$ , one needs  $T_k \approx \varepsilon^{-2}$  observations.
- Fact 2: at least one arm is sampled less than  $T/K$  times.

Assume that this arm is  $k^*$ , then the learner cannot distinguish it with other arms as soon as  $T_{k^*} \leq T/K \leq \varepsilon^{-2}$ , which corresponds to  $\varepsilon \leq \sqrt{K/T}$ . Choosing  $\varepsilon = \sqrt{K/T}$ , the pseudo-regret is then at least  $(1 - 1/K)T\varepsilon \approx T\varepsilon \approx \sqrt{KT}$ .

### 5.5.2 Distribution-dependent lower bound

Here, we show that the distribution dependent upper bound is not also optimal in the case of Bernoulli rewards.

A caveat with distribution dependent lower bounds is that for any distribution, there exists an algorithm with no-regret. For instance, consider a distribution  $\nu_1, \dots, \nu_K$  such that  $\nu_1$  is optimal (i.e.,  $\mu_1 = \max_k \mu_k$ ), then the algorithm that pull always the first arm will have zero regret. Yet such an algorithm will incur linear regret for some other distributions.

Hence, the following theorem states that any algorithm that incur sublinear regret for all distributions, achieves at best a pseudo regret of the same order of the one satisfied by UCB. The proof can be found in [Bubeck et al. \[2012\]](#).

#### Theorem 5.7

Thm 2.2. [Bubeck et al. \[2012\]](#)

Consider a strategy such that  $\mathbb{E}[N_k(T)] = O(T^a)$  for any Bernoulli distributions, all suboptimal arms  $k$  and some  $a > 0$ . Then, for any Bernoulli distributions with means  $\mu_k$ , we have

$$\liminf_{T \rightarrow \infty} \frac{\bar{R}_T}{\log T} \geq \sum_{k: \Delta_k > 0} \frac{\mu^*(1 - \mu^*)}{\Delta_k}.$$

Note that the only difference with UCB is the factor  $\mu^*(1 - \mu^*)$  which corresponds to the variance of the best arm. In the case of Bernoulli noise, the KL-UCB algorithm can take advantage of the knowledge that the rewards are Bernoulli to close this gap.

## 6 Stochastic Contextual bandits

During last chapter, we considered the finite-armed bandit setting and saw several algorithms (ETC, UCB,  $\varepsilon$ -greedy, Thomson sampling) that achieve sublinear pseudo regret. UCB achieves for instance

$$\bar{R}_T \lesssim \min \left\{ \sum_{k: \Delta_k > 0}^K \frac{\log T}{\Delta_k}, \sqrt{TK \log T} \right\},$$

where  $\Delta_k \stackrel{\text{def}}{=} \mu^* - \mu_k$  is the suboptimality gap of arm  $k$ . The first bound is distribution dependent (it depends on the gaps  $\Delta_k$ ) and is of order  $O(\log T)$  while the second bound is distribution free but is of order  $O(\sqrt{T})$ . In this chapter, we consider the more practical setting of contextual bandits, in which the learner observes a context  $c_t \in C$  before choosing the action  $k_t$ .

In most applications, before choosing an action  $k_t$  the player observes some context  $c_t \in C$ .

For instance, consider a bandit problem in which the player needs to display ads on his website. At each new visitor, the player chooses an add to display and observes if the visitor click on it. The reward is one if there is a click and 0 otherwise. In this case, the player can see the cookie of the visitor before choosing the ad. A first step towards contextual bandits, is to consider continuous sets of actions  $\mathcal{X}$ , which may correspond to mapping between context and arms.

### 6.1 Continuous stochastic bandits

Let first generalize the finite-armed bandit setting to continuous set of arms in Setting 6.1.

*Unknown parameters:*  $v(\theta)$ , for each  $\theta \in [0, 1]^d$ , a probability distribution on  $[0, 1]$  with expectation  $\mu(\theta) \in [0, 1]$ .

At each time step  $t = 1, \dots, T$

- the player chooses an action  $\theta_t \in \Theta \subseteq [0, 1]^d$ ;
- given  $\theta_t$ , the environment draws the reward  $Y_t \sim v(\theta_t)$  independently from the past;
- the player only observes the feedback  $Y_t$ .

The player wants to minimize its pseudo-regret defined as

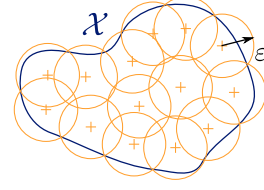
$$\bar{R}_T \stackrel{\text{def}}{=} T\mu^* - \mathbb{E} \left[ \sum_{t=1}^T \mu(\theta_t) \right],$$

where  $\mu^* = \sup_{\theta \in \Theta} \mu(\theta)$ .

Setting 6.1: Setting of stochastic bandit with continuous set of actions

Similarly to what we did in the full-information setting with EWA, if the expectation function  $\mu$  is  $\beta$ -Hölder: i.e., there exists  $c > 0$

$$\forall \theta, \theta' \in \mathcal{X} \quad |\mu(\theta) - \mu(\theta')| \leq c \|\theta - \theta'\|^\beta,$$



then we may discretize the action space  $\Theta$  and run any discrete bandit algorithm (UCB,  $\epsilon$ -greedy, ...).

### Theorem 6.1

Let  $\beta > 0$  and  $\epsilon > 0$ . Assume that  $\mu$  is  $\beta$ -Hölder. If UCB is run on an  $\epsilon$ -covering of minimal cardinal of  $\Theta \subset [0, 1]^d$ , then it satisfies

$$\bar{R}_T \lesssim T\epsilon^\beta + \sqrt{\frac{T \log(T)}{\epsilon^d}}.$$

In particular for  $\epsilon \approx \left(\frac{\log T}{T}\right)^{\frac{1}{2\beta+d}}$ , we have  $\bar{R}_T \lesssim T \left(\frac{\log T}{T}\right)^{\frac{\beta}{2\beta+d}}$ .

*Proof.* An optimal  $\epsilon$ -covering of  $[0, 1]^d$  has cardinal of order  $K \approx \epsilon^{-d}$ . Let  $x^* \in \arg \max_{\theta \in \Theta} \mu(\theta)$  (we assume that it exists) and  $\tilde{\theta}^*$  its  $\epsilon$ -approximation, then the distribution-free upper-bound of UCB yields

$$\bar{R}_T \lesssim T(\mu(\theta^*) - \mu(\tilde{\theta}^*)) + \sqrt{KT \log T} \approx cT\epsilon^\beta + \sqrt{\epsilon^{-d} T \log T}.$$

The second part of the theorem is obtained by optimizing  $\epsilon$ . □

To build the discretization, both  $\beta$  and  $T$  need to be known in advance. The horizon  $T$  can be calibrated online through a “doubling trick” (left as exercise). The parameter  $\beta$  may be tuned through bandit with experts (or bandits where arms are bandit algorithms) that we may see in next lecture (see Exp4 algorithm).

Note that the per-round complexity of such an algorithm is of order  $\epsilon^{-d} \approx T^{\frac{d}{2\beta+d}}$ . Quite surprisingly it does not explodes with the dimension  $d$  and is always smaller than  $T$ . This is due to the fact that the higher the dimension  $d$  is, the worse will be the regret bound, and the cruder needs the discretization to be.

### 6.1.1 Contextual bandits through discretization

No we consider the following contextual bandit setting in which the player has a finite decision set  $\Theta = \{1, \dots, K\}$  but observes a context  $x_t \in \mathcal{X}$  before choosing his action.

**Finite set of contexts** If the set of context is finite  $\mathcal{X} \stackrel{\text{def}}{=} \{1, \dots, |\mathcal{X}|\}$  we can denote

$$\bar{R}_T(c) \stackrel{\text{def}}{=} \mathbb{E} \left[ \sum_{t=1}^T (\mu^*(x) - \mu(k_t, x)) \mathbb{1}_{x_t=x} \right]$$

the pseudo-regret due to context  $x \in \mathcal{X}$ . Then applying a separate instance of UCB (or any bandit algorithm) for each context  $x \in \mathcal{X}$ , we get by using the distribution-free upper-bound of UCB

$$\bar{R}_T(x) \lesssim \sqrt{T_x K \log T_x}, \quad \text{where} \quad T_x \stackrel{\text{def}}{=} \sum_{t=1}^T \mathbb{1}_{x_t=x}.$$

*Unknown parameters:*  $v(k, x)$ , for each arm  $k \in \{1, \dots, K\}$  and context  $x \in \mathcal{X}$ , a probability distribution on  $[0, 1]$  with expectation  $\mu(k, x) \in [0, 1]$ .

At each time step  $t = 1, \dots, T$

- the environment chooses  $x_t \in \mathcal{X}$  and reveals it to the player;
- the player chooses an action  $k_t \in \{1, \dots, K\}$ ;
- given  $k_t$ , the environment draws the reward  $Y_t \sim v(k_t, x_t)$  independently from the past;
- the player only observes the feedback  $Y_t$ .

The player wants to minimize its pseudo-regret defined as

$$\bar{R}_T \stackrel{\text{def}}{=} \mathbb{E} \left[ \sum_{t=1}^T \mu^*(x_t) - \sum_{t=1}^T Y_t \right],$$

where  $\mu(k, x) = \mathbb{E}_{Y \sim v(k, x)}[Y]$  and  $\mu^*(x) = \max_{k=1, \dots, K} \mu(k, x)$ .

### Setting 6.2: Setting of contextual stochastic bandit

Note that because  $T_x$  are not known in advance it is important that the bound of UCB is anytime (i.e., that UCB does not need to know the horizon). The total pseudo-regret of UCB is then obtained by summing over all contexts

$$\bar{R}_T = \sum_{x \in \mathcal{X}} \bar{R}_T(x) \lesssim \sum_{x \in \mathcal{X}} \sqrt{T_x K \log T} \leq \sqrt{|\mathcal{X}| T K \log T},$$

where the last inequality is by Jensen's inequality using the concavity of the square root and  $\sum_{x \in \mathcal{X}} T_x = T$ .

**Continuous set of contexts** If the set of context is continuous  $\mathcal{X} \subset [0, 1]^d$ , one needs again to make assumption on the distributions  $v(k, x)$  which needs to vary smoothly in  $x$ . Doing so, one may discretize the set of context with an  $\varepsilon$ -covering of  $\mathcal{X}$  of size  $N \approx \varepsilon^{-d}$  and run an independent instance of UCB in each of the  $N$  bins.

#### Theorem 6.2

Let  $\beta > 0$  and  $\varepsilon > 0$ . Assume that  $x \mapsto \mu(k, x)$  is  $\beta$ -Hölder for all  $k \in \mathcal{X}$ . If UCB is independently run in each bin of an optimal  $\varepsilon$ -covering of  $\mathcal{X}$ , then

$$\bar{R}_T \lesssim T \varepsilon^\beta + \sqrt{\frac{K T \log T}{\varepsilon^d}}.$$

In particular for  $\varepsilon$  well-optimized, we have  $\bar{R}_T \lesssim T \left( \frac{K \log T}{T} \right)^{\frac{\beta}{2\beta+d}}.$

Remark that in all these regret bounds, the suboptimal  $\log T$  term can be removed by using MOSS (a minimax optimal variant of UCB).

**Better rates using distribution-dependent bound?** In the above results, we used the distribution-free regret bound of UCB. Because, if the function  $\mu(\cdot, x)$  varies smoothly with  $x$ , there should be some

context with zero suboptimality gaps. Yet, it is possible to get better rates by assuming the following  $\alpha$ -margin assumption. It controls the suboptimality gap with high probability: the contexts  $x_t$  are i.i.d. and satisfy for all  $\delta \in (0, 1)$

$$\mathbb{P}\left\{\min_{k:\Delta(k,x_t)>0}\Delta(k,x_t)<\delta\right\}\leq\varpi\delta^\alpha \quad (6.1)$$

where  $\Delta(k, x_t) \stackrel{\text{def}}{=} \mu^*(x) - \mu(k, x)$  and  $\varpi$  is some constant. Note that the larger the value of  $\alpha$  is the easier is the problem.

### Theorem 6.3

Let  $\alpha \in (0, 1)$ ,  $\beta > 0$  and  $\varepsilon > 0$ . Assume that  $c \mapsto \mu(k, x)$  is  $\beta$ -Hölder for all  $k \in \mathcal{X}$  and that the  $\alpha$ -margin assumption (6.1) holds. Running a bandit algorithm (similar to UCB) independently in each bin of an optimal  $\varepsilon$ -covering of  $\mathcal{X}$ , we get

$$\bar{R}_T \lesssim T \left( \frac{K \log K}{T} \right)^{\frac{\beta(\alpha+1)}{2\beta+d}},$$

for optimized  $\varepsilon$ .

### Theorem 4.1, Perchet and Rigollet [2013]

The proof (for another algorithm than UCB) may be found in Perchet and Rigollet [2013]. We see that the factor  $\alpha$  improves the rate of convergence with respect to the previous rate.

## 6.1.2 Stochastic Linear bandits

Contextual bandits that we just saw generalizes multi-armed bandits by allowing contexts. However, the dimension of the context space significantly worsen the regret rate from  $\sqrt{T}$  to  $T^{\frac{d+1}{d+2}}$  for Lipschitz rewards for instance ( $\beta$ -Hölder with  $\beta = 1$ ). In this part, we will see *Stochastic linear bandits*, in which we assume the rewards to have a linear structure. This includes rich classes of models and allows better regret of order  $O(\sqrt{T})$ .

Unknown parameter:  $\mu^* \in \mathbb{R}^d$ .

At each time step  $t = 1, \dots, T$

- the environment chooses  $\Theta_t \subseteq \mathbb{R}^d$  the decision set;
- the player chooses an action  $\theta_t \in \Theta_t$ ;
- given  $\theta_t$ , the environment draws the reward

$$Y_t = \theta_t \cdot \mu^* + \varepsilon_t$$

where  $\varepsilon_t$  is i.i.d. 1-subgaussian noise.

- the player only observes the feedback  $Y_t$ .

The player wants to minimize its pseudo-regret defined as

$$\bar{R}_T \stackrel{\text{def}}{=} \mathbb{E} \left[ \sum_{t=1}^T \max_{\theta \in \Theta_t} \theta \cdot \mu^* - \sum_{t=1}^T Y_t \right].$$

Setting 6.3: Setting of stochastic linear bandit

The setting of stochastic linear bandits is described in Setting 6.3. For simplicity, the noise  $\varepsilon_t$  is assumed

to be i.i.d. and 1-subgaussian noise: i.e.,  $\mathbb{E}[\varepsilon_t] = 0$  and

$$\forall \lambda > 0, \quad \mathbb{E}[\exp(\lambda \varepsilon_t)] \leq \exp(\lambda^2/2)$$

almost surely. Note that we could consider  $\sigma^2$ -subgaussian noise, or make it depend on the past  $\mathcal{F}_t = \sigma(x_1, \varepsilon_1, \dots, x_t, \varepsilon_t)$  with  $\mathbb{E}[\varepsilon_t | \mathcal{F}_t] = 0$ .

**Particular cases: why is this setting interesting?** Different choices of decision sets  $\mathcal{X}_t$  lead to different settings of stochastic bandits:

- *Finite-armed bandit*: if  $\Theta_t = (e_1, \dots, e_d)$  where  $e_i$  are the unit vectors in  $\mathbb{R}^d$  and  $\mu^* = (\mu_1, \dots, \mu_d)$ , we recover the setting of finite-armed bandit.
- *Contextual linear bandit*: we can recover a particular case of Setting 6.2, if  $x_t \in \mathcal{X}$  is a context observed by the player and the reward function  $\mu$  is of the form

$$\mu(\theta, x) = \psi(\theta, x) \cdot \mu^*, \quad \forall (\theta, x) \in [K] \times \mathcal{X},$$

for some unknown parameter  $\mu^* \in \mathbb{R}^d$  and *feature map*  $\psi : [K] \times \mathcal{X} \rightarrow \mathbb{R}^d$ . For example, assume that you are a website which wants to display ads to visitors. The context  $x_t$  can be the cookie of the visitor containing information about what he likes, the actions are the possible ads to be displayed and the reward tells if there is a click. If the possible interests of the visitor are grouped in finite categories (such as traveling), so are the ads (in groups of products, such as flight tickets), the feature maps could contained all the combinations between interests and groups of products. The unknown vector  $\theta^*$  would be tell which interests and groups of products are positively correlated. Of course the feature map could be created using any methods (such as deep-learning or splines).

- *Combinatorial bandit*: if  $\Theta_t \subseteq \{0, 1\}^d$  yields to combinatorial bandit problems. For instance, the decision set corresponds to possible paths in a graph, the vector  $\mu^*$  assigns to each edge a reward corresponding to its cost and the goal is to find the smallest path with smallest cost.

**Algorithm: LinUCB** As we saw earlier with UCB, the “optimism principle” is a good option for bandit problems to explore. The LinUCB algorithm is based on the same principle:

1. Build confidence set that contain  $\mu^*$ :  $\mu^* \in C_t$  with high probability
2. Build confidence upper-bound on the rewards: for all  $\theta \in \Theta_t$

$$\text{UCB}_t(\theta) = \max_{\mu \in C_t} \theta \cdot \mu \tag{6.2}$$

3. Be optimistic: act as if the best possible rewards were the true rewards

$$\theta_t \in \arg \max_{\theta \in \Theta_t} \text{UCB}_t(\theta). \tag{6.3}$$

Therefore the only remaining question is how to build the confidence set  $C_t \subseteq \mathbb{R}^d$ ? They should contain  $\mu^*$  with high probability but be as small as possible. Given the observed rewards the key is thus to estimate the parameter  $\mu^*$ . Denoting by  $I_d$  the  $d \times d$  identity matrix and picking  $\lambda > 0$ , we can estimate  $\mu^*$  with *regularized least square*

$$\hat{\mu}_t \stackrel{\text{def}}{=} \arg \min_{\mu \in \mathbb{R}^d} \left\{ \sum_{s=1}^t (Y_s - \theta_s \cdot \mu)^2 + \lambda \|\mu\|^2 \right\} = V_t^{-1} \sum_{s=1}^t \theta_s Y_s,$$

where  $V_t \stackrel{\text{def}}{=} \lambda I_d + \sum_{s=1}^t \theta_s \theta_s^\top$ . We have the following result whose proof can be found in [Lattimore and Szepesvári \[2020\]](#).

**Lemma 6.4**

**Theorem 20.2, Lattimore and Szepesvári [2020]**

Let  $\delta \in (0, 1)$ . Then, with probability at least  $1 - \delta$ , if  $\max_{\theta \in \Theta_t} \|\theta\|_2 \leq 1$ , for all  $t \geq 1$

$$\|\hat{\mu}_t - \mu^*\|_{V_t} \leq \sqrt{\lambda} \|\mu^*\| + \sqrt{2 \log(1/\delta) + d \log \left(1 + \frac{T}{\lambda}\right)} \stackrel{\text{def}}{=} \beta(\delta),$$

where  $\|\mu\|_{V_t}^2 = \mu^\top V_t \mu$ .

The above lemma, states that with probability  $1 - \delta$ , for all  $t \geq 1$ ,

$$\mu^* \in C_t, \quad \text{where} \quad C_t \stackrel{\text{def}}{=} \left\{ \theta \in \mathbb{R}^d : \|\mu - \hat{\mu}_{t-1}\|_{V_{t-1}} \leq \beta(\delta/T) \right\}. \quad (6.4)$$

*Proof of Lemma 6.4.* The proof relies on Laplace's method on super-martingales which is a standard argument to provide confidence bounds on a self-normalized sum of conditionally centered random vectors. We have

$$\hat{\mu}_t = V_t^{-1} \sum_{s=1}^t \theta_s Y_s = V_t^{-1} \sum_{s=1}^t \theta_s (\theta_s^\top \mu^* + \varepsilon_s) = V_t^{-1} (V_t - \lambda I_d) \mu^* + M_t = \mu^* - \lambda V_t^{-1} \mu^* + V_t^{-1} M_t,$$

where we introduced  $M_t = \sum_{s=1}^t \theta_s \varepsilon_s$ , which is a martingale with respect to  $\mathcal{F}_t = \sigma(\varepsilon_1, \dots, \varepsilon_t)$ . Therefore, by triangle inequality

$$\|V_t^{1/2} (\hat{\mu}_t - \mu^*)\| = \|\lambda V_t^{-1/2} \mu^* + V_t^{-1/2} M_t\| \leq \lambda \|V_t^{-1/2} \mu^*\| + \|V_t^{-1/2} M_t\|.$$

On the one hand, given that all eigenvalues of the symmetric matrix  $V_t$  are larger than  $\lambda$ , all eigenvalues of  $V_t^{-1/2}$  are smaller than  $1/\sqrt{\lambda}$  and thus

$$\lambda \|V_t^{-1/2} \mu^*\| \leq \lambda \frac{1}{\sqrt{\lambda}} \|\mu^*\| = \sqrt{\lambda} \|\mu^*\|.$$

We now prove, on the other hand, that with probability at least  $1 - \delta$

$$\|V_t^{-1/2} M_t\| \leq \sqrt{2 \log \frac{1}{\delta} + d \log \frac{1}{\lambda} + \log \det(V_t)}.$$

Upper-bounding  $\log \det(V_t) \leq d \log(\lambda + t)$  (since all the eigenvalues of  $V_t$  are smaller than  $\lambda + t$ ) will then conclude the proof of the Theorem.

*Step 1: Introducing super-martingales.* For all  $v \in \mathbb{R}^d$ , we consider

$$S_{t,v} = \exp \left( v^\top M_t - \frac{1}{2} v^\top V_t v \right)$$

and now show that it is an  $\mathcal{F}_t$ -super-martingale. First, note that since the common distribution of the  $\varepsilon_1, \dots, \varepsilon_t$  is 1-sub-Gaussian, the for all  $\mathcal{F}_{t-1}$ -measurable random variable  $v_{t-1}$

$$\mathbb{E} \left[ e^{v_{t-1}^\top \varepsilon_t} \middle| \mathcal{F}_{t-1} \right] \leq e^{\frac{v_{t-1}^2}{2}}.$$

Now,

$$\mathbb{E}[S_{t,v}|\mathcal{F}_{t-1}] = S_{t-1,v}\mathbb{E}\left[\exp\left(v^\top\theta_t\varepsilon_t - \frac{1}{2}v^\top\theta_t\theta_t^\top v\right)\middle|\mathcal{F}_{t-1}\right] \leq S_{t-1,v}.$$

Note that rewriting  $S_{t,v}$  in its vertex form is, with  $m = V_t^{-1}M_t$ :

$$S_{t,v} = \exp\left(\frac{1}{2}(v-m)^\top V_t(v-m)\right) \times \exp\left(\frac{1}{2}\|V_t^{-1/2}M_t\|^2\right).$$

*Step 2: Laplace's method—integrating  $S_{t,v}$  over  $v \in \mathbb{R}^d$ .* The basic observation behind this method is that (given the vertex form)  $S_{t,v}$  is maximal at  $v = m = V_t^{-1}M_t$  and then equals  $\exp\left(\frac{1}{2}\|V_t^{-1/2}M_t\|^2\right)$ , which is (a transformation of) the quantity to control. Now, because the exp function quickly vanishes, the integral over  $v \in \mathbb{R}^d$  is close to its maximum. We therefore consider

$$\bar{S}_t = \int_{\mathbb{R}^d} S_{t,v} dv.$$

We will make repeated uses of the fact that the Gaussian density function

$$v \mapsto \frac{1}{\sqrt{\det(2\pi C)}} \exp\left((v-m)^\top C^{-1}(v-m)\right),$$

where  $m \in \mathbb{R}^d$  and  $C$  is a symmetric positive definite matrix, integrate to 1 over  $\mathbb{R}^d$ . This gives us the first rewriting

$$\bar{S}_t = \sqrt{\det(2\pi V_t^{-1})} \exp\left(\frac{1}{2}\|V_t^{-1/2}M_t\|^2\right).$$

Second, by the Fubini-Tonelli theorem and the super-martingale property

$$\mathbb{E}[S_{t,v}] \leq \mathbb{E}[S_{0,v}] = \exp(-\lambda\|v\|^2/2),$$

we also have

$$\mathbb{E}[\bar{S}_t] \leq \int_{\mathbb{R}^d} \exp(-\lambda\|v\|^2/2) dv = \sqrt{\det(2\pi\lambda^{-1}I_d)}.$$

Combining the two statements, we proved

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\|V_t^{-1/2}M_t\|^2\right)\right] \leq \sqrt{\frac{\det(V_t)}{\lambda^d}}.$$

*Step 3: Markov-Chernov bound.* For  $u > 0$ ,

$$\begin{aligned} \mathbb{P}\left(\|V_t^{-1/2}M_t\| > u\right) &= \mathbb{P}\left(\frac{\|V_t^{-1/2}M_t\|^2}{2} > \frac{u^2}{2}\right) \\ &\leq \exp\left(-\frac{1}{2}u^2\right)\mathbb{E}\left[\exp\left(\frac{1}{2}\|V_t^{-1/2}M_t\|^2\right)\right] \leq \exp\left(-\frac{u^2}{2} + \frac{1}{2}\log\frac{\det(V_t)}{\lambda^d}\right) = \delta, \end{aligned}$$

for the claimed choice

$$u = \sqrt{2\log\frac{1}{\delta} + d\log\frac{1}{\lambda} + \log\det(V_t)}.$$

□



**Theorem 6.5****Corollary 19.2, ?**

Let  $T \geq 1$  and  $\mu^* \in \mathbb{R}^d$ . Assume that for all  $\theta \in \cup_{t=1}^T \Theta_t$ ,  $|\theta \cdot \mu^*| \leq 1$ ,  $\|\mu^*\| \leq 1$  and  $\|\theta_t\| \leq 1$ , then LinUCB with  $C_t$  defined in (6.4) satisfies the pseudo-regret bound

$$\bar{R}_T \leq \square_\lambda d \sqrt{T} \log(T),$$

where  $\square_\lambda$  is a constant that may depend on  $\lambda$ .

*Proof.* Let  $\delta = 1/T$ . By Lemma 6.4, with probability  $1 - 1/T$ ,

$$\forall t \geq 1, \quad \mu^* \in C_t. \quad (6.5)$$

*Step 1: Small instantaneous regrets under the event (6.5).* Assume that (6.5) holds. Let

$$\theta_t^* \stackrel{\text{def}}{=} \max_{\theta \in \Theta_t} \theta \cdot \mu^* \quad \text{and} \quad r_t \stackrel{\text{def}}{=} (\theta_t^* - \theta_t) \cdot \mu^*$$

be respectively the optimal decision and the instantaneous regret at round  $t$ . We also define

$$\tilde{\mu}_t \in \arg \max_{\mu \in C_t} \{\theta_t \cdot \mu\}.$$

Since  $\mu^* \in C_t$ , we have

$$\theta_t^* \cdot \mu^* \leq \max_{\mu \in C_t} \{\theta_t^* \cdot \mu\} \stackrel{(6.2)}{=} UCB_t(\theta_t^*) \stackrel{(6.3)}{\leq} UCB_t(\theta_t) = \max_{\mu \in C_t} \{\theta_t \cdot \mu\} = \theta_t \cdot \tilde{\mu}_t,$$

which entails because  $\mu^*$  and  $\tilde{\mu}_t$  belong to  $C_t$ ,

$$r_t \stackrel{\text{def}}{=} (\theta_t^* - \theta_t) \cdot \mu^* \leq \theta_t \cdot (\tilde{\mu}_t - \mu^*) \stackrel{\text{Cauchy-Schwarz}}{\leq} \|\theta_t\|_{V_{t-1}^{-1}} \|\tilde{\mu}_t - \mu^*\|_{V_{t-1}} \leq 2 \|\theta_t\|_{V_{t-1}^{-1}} \beta(1/T^2).$$

Therefore, summing over  $t = 1, \dots, T$  and using  $r_t \leq 2$ , we have

$$\begin{aligned} R_T &\stackrel{\text{def}}{=} \sum_{t=1}^T r_t \leq \sqrt{T \sum_{t=1}^T r_t^2} \quad \leftarrow \text{Jensen's inequality} \\ &\leq 2 \sqrt{T \sum_{t=1}^T \min \left\{ \|\theta_t\|_{V_{t-1}^{-1}}^2 \beta(1/T^2)^2, 1 \right\}} \\ &\leq 2\beta(1/T^2) \sqrt{T \sum_{t=1}^T \min \left\{ \|\theta_t\|_{V_{t-1}^{-1}}^2, 1 \right\}} \quad \leftarrow \beta_T(1/T^2) \geq 1 \\ &\leq 2\beta(1/T^2) \sqrt{T \sum_{t=1}^T \log \left( 1 + \|\theta_t\|_{V_{t-1}^{-1}}^2 \right)} \quad \leftarrow \min\{u, 1\} \leq 2 \log(1+u). \end{aligned}$$

But, we have

$$1 + \|\theta_t\|_{V_{t-1}^{-1}}^2 = \det \left( 1 + \|\theta_t\|_{V_{t-1}^{-1}}^2 \right)$$

$$\begin{aligned}
&= \det \left( V_{t-1}^{-1} (V_{t-1} + V_{t-1}^{1/2} \|\theta_t\|_{V_{t-1}^{-1}}^2 V_{t-1}^{1/2}) \right) \leftarrow \text{using } \det(I + AB) = \det(I + BA) \\
&= \det \left( V_{t-1}^{-1} (V_{t-1} + \theta_t \theta_t^\top) \right) \leftarrow \|\theta_t\|_{V_{t-1}^{-1}}^2 = V_{t-1}^{-1/2} \theta_t \theta_t^\top V_{t-1}^{-1/2} \\
&= \det(V_t) \leftarrow V_t = V_{t-1} + \theta_t \theta_t^\top \\
&= \frac{\det(V_t)}{\det(V_{t-1})} \leftarrow \det(A^{-1}B) = \frac{\det(B)}{\det(A)}.
\end{aligned}$$

Substituting into the regret bound, the sum telescopes and it entails

$$R_T \leq 2\beta(1/T^2) \sqrt{T \log \left( \frac{\det(V_T)}{\det(V_0)} \right)}.$$

Then, using  $V_0 \stackrel{\text{def}}{=} \lambda I_d$  and since  $V_T = \lambda I_d + \sum_{t=1}^T \theta_t \theta_t^\top$  with  $\|\theta_t\| \leq 1$ , all eigenvalues of  $V_T$  lie in  $[\lambda, \lambda + T]$  which yields

$$\det(V_0) = \lambda^d \quad \text{and} \quad \det(V_T) \leq (\lambda + T)^d.$$

Plugging back into the previous upper-bound and using that  $\beta(1/T^2) \leq \square_\lambda \sqrt{d \log T}$

$$R_T \leq 2\sqrt{dT\beta(1/T) \log \left( 1 + \frac{T}{\lambda} \right)} \leq \square_\lambda d\sqrt{T} \log T.$$

*Part 2: without the event (6.4)* We because  $r_t \leq 2$ , almost surely  $R_T \leq 2T$ , and we have

$$\begin{aligned}
\bar{R}_T &= \mathbb{E}[R_T] \leq \mathbb{E} \left[ R_T \mid \text{Event (6.4)} \right] \mathbb{P}\{\text{Event (6.4)}\} + 2T(1 - \mathbb{P}\{\text{Event (6.4)}\}) \\
&\leq \square d\sqrt{T} \log T + 2.
\end{aligned}$$

This concludes the proof. □

**Better regret with assumptions** It is worth pointing out that if we make additional assumptions, it is possible to improve the regret bound  $O(d\sqrt{T} \log T)$ . A first setting corresponds to the case where the set of available actions at time  $t$  is fixed and finite; i.e., the learner needs to choose  $\theta_t \in \Theta$  where  $|\Theta| = K$ . Then, it is possible to achieve

$$R_T \leq \square \sqrt{Td \log(TK)},$$

which improves the previous bound by a factor  $\sqrt{d}/\log(K)$  and improves the classical bound of UCB  $O(\sqrt{TK \log T})$  by a factor  $K/\sqrt{d}$ . These improvements can be significant when  $K \gg d \gg 1$ . We refer the curious reader to [Lattimore and Szepesvári, 2020, Chapter 22].

Another possible improvement when  $d \gg 1$  is to assume that  $\mu^*$  is  $m_0$ -sparse (i.e., most of its components are zero). Then under assumptions, one can get a regret of order  $\tilde{O}(\sqrt{dm_0 T})$ .

## 6.2 Other possible extensions of bandits

Note that there exist many different extensions of stochastic bandits to make it more realistic or with improved regret.

- *Bandit with delays*: For instance, consider the example of the website which wants to display ads. The website does not observe if there is no click, he needs to fix some time after which he consider that the visitor will not click, and if the visitor stays long on the webpage, the website may need to display ads to other visitors before getting the rewards. There is thus delayed feedback the website needs to deal with.
- *Non stationarity*
- *Combinatorial bandits*
- *Dueling bandits*
- ...

We refer the interested student to the monograph [Lattimore and Szepesvári \[2020\]](#) for more information on these settings. Next week, we will deal with adversarial bandits.

# Bibliography

- Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.
- Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.
- Thomas M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991.
- Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.
- Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.
- Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.
- Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, pages 693–721, 2013.
- Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.
- Tim Van Erven, Peter D Grünwald, Nishant A Mehta, Mark D Reid, and Robert C Williamson. Fast rates in statistical and online learning. *The Journal of Machine Learning Research*, 16(1):1793–1861, 2015.
- Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.