

Optimal Multi-Distribution Learning

Zihan Zhang*
Princeton

Wenhao Zhan*
Princeton

Yuxin Chen†
UPenn

Simon S. Du‡
U. Washington

Jason D. Lee*
Princeton

December 2023; Revised: March 2024

Abstract

Multi-distribution learning (MDL), which seeks to learn a shared model that minimizes the worst-case risk across k distinct data distributions, has emerged as a unified framework in response to the evolving demand for robustness, fairness, multi-group collaboration, etc. Achieving data-efficient MDL necessitates adaptive sampling, also called on-demand sampling, throughout the learning process. However, there exist substantial gaps between the state-of-the-art upper and lower bounds on the optimal sample complexity. Focusing on a hypothesis class of Vapnik–Chervonenkis (VC) dimension d , we propose a novel algorithm that yields an ε -optimal randomized hypothesis with a sample complexity on the order of $\frac{d+k}{\varepsilon^2}$ (modulo some logarithmic factor), matching the best-known lower bound. Our algorithmic ideas and theory are further extended to accommodate Rademacher classes. The proposed algorithms are oracle-efficient, which access the hypothesis class solely through an empirical risk minimization oracle. Additionally, we establish the necessity of randomization, revealing a large sample size barrier when only deterministic hypotheses are permitted. These findings resolve three open problems presented in COLT 2023 (i.e., Awasthi et al. (2023, Problems 1, 3 and 4)).

Keywords: multi-distribution learning; on-demand sampling; game dynamics; VC classes; Rademacher classes; oracle efficiency

Contents

1	Introduction	2
2	Problem formulation	6
3	Algorithm	7
4	A glimpse of key technical novelty	9
4.1	Towards sample reuse: uniform concentration and a key quantity	9
4.2	Bounding the key quantity $\ \bar{w}^T\ _1$ by controlling the Hedge trajectory	10
4.2.1	Doubling w_j needs $\tilde{\Omega}(1/\varepsilon^2)$ steps	11
4.2.2	Coping with the segments	12
5	Analysis for VC classes (proof of Theorem 1)	14
6	Necessity of randomization	15

*Department of Electrical and Computer Engineering, Princeton University; email: {zz5478, wenhao.zhan, jasonlee}@princeton.edu.

†Department of Statistics and Data Science, University of Pennsylvania; email: yuxinc@wharton.upenn.edu.

‡Paul G. Allen School of Computer Science and Engineering, University of Washington; email: ssdu@cs.washington.edu.

7	Extension: learning Rademacher classes	18
7.1	Preliminaries: Rademacher complexity	18
7.2	Algorithm and sample complexity	19
8	Extension: oracle-efficient multi-group learning	20
9	Discussion	20
A	Additional figures	22
B	Auxiliary lemmas	23
C	Proofs of auxiliary lemmas for VC classes	26
C.1	Proof of Lemma 1	26
C.2	Proof of Lemma 2	30
C.3	Proof of Lemma 3	31
D	Controlling the Hedge trajectory (proof of Lemma 13)	32
D.1	Outline	33
D.2	Step 1: identifying segments for each distribution in \mathcal{W}_j	33
D.3	Step 2: extracting regular segments from irregular segments	34
D.4	Step 3: bounding the length of segments	39
D.5	Putting all this together	39
D.6	Proof of Lemma 17	41
E	Missing proofs for lower bounds	45
E.1	Restatement and proof of Lemma 4	45
E.2	Statement and proof of Lemma 18	47
E.3	Statement and proof of Lemma 19	48
F	Missing proofs for Rademacher classes	49
F.1	Restatement and proof of Lemma 7	49
F.2	Restatement and proof of Lemma 5	51
F.2.1	Restatement and proof of Lemma 6	52
F.3	Necessity of Assumption 1	53
G	Proof of Theorem 4	54

1 Introduction

Driven by the growing need of robustness, fairness and multi-group collaboration in machine learning practice, the multi-distribution learning (MDL) framework has emerged as a unified solution in response to these evolving demands (Blum et al., 2017; Haghtalab et al., 2022; Mohri et al., 2019; Awasthi et al., 2023). Setting the stage, imagine that we are interested in a collection of k unknown data distributions $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^k$ supported on $\mathcal{X} \times \mathcal{Y}$, where \mathcal{X} (resp. \mathcal{Y}) stands for the instance (resp. label) space. Given a hypothesis class \mathcal{H} and a prescribed loss function¹ $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \rightarrow [-1, 1]$, we are asked to identify a (possibly randomized)

¹For example, for each hypothesis $h \in \mathcal{H}$ and each datapoint $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we employ $\ell(h, (x, y))$ to measure the risk of using hypothesis h to predict y based on x .

hypothesis \hat{h} achieving near-optimal *worst-case* loss across these data distributions, namely,²

$$\max_{1 \leq i \leq k} \mathbb{E}_{(x,y) \sim \mathcal{D}_i, \hat{h}} [\ell(\hat{h}, (x, y))] \leq \min_{h \in \mathcal{H}} \max_{1 \leq i \leq k} \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\ell(h, (x, y))] + \varepsilon \quad (1)$$

with $\varepsilon \in (0, 1]$ a target accuracy level. In light of the unknown nature of these data distributions, the learning process is often coupled with data collection, allowing the learner to sample from $\{\mathcal{D}_i\}_{i=1}^k$. The performance of a learning algorithm is then gauged by its sample complexity — the number of samples required to fulfil (1). Our objective is to design a learning paradigm that achieves the optimal sample complexity.

The MDL framework described above, which can be viewed as an extension of agnostic learning (Valiant, 1984; Blumer et al., 1989) tailored to multiple data distributions, has found a wealth of applications across multiple domains. Here, we highlight a few representative examples, and refer the interested reader to Haghtalab et al. (2022) and the references therein for more extensive discussions.

- *Collaborative and agnostic federated learning.* In the realm of collaborative and agnostic federated learning (Blum et al., 2017; Nguyen and Zakynthinou, 2018; Chen et al., 2018; Mohri et al., 2019; Blum et al., 2021a; Du et al., 2021; Deng et al., 2020; Blum et al., 2021b), a group of k agents, each having access to distinct data sources as characterized by different data distributions $\{\mathcal{D}_i\}_{i=1}^k$, aim to learn a shared prediction model that ideally would achieve low risk for each of their respective data sources. A sample-efficient MDL paradigm would help unleash the potential of collaboration and information sharing in jointly learning a complicated task.
- *Min-max fairness in learning.* The MDL framework is well-suited to scenarios requiring fairness across multiple groups (Dwork et al., 2021; Rothblum and Yona, 2021; Du et al., 2021). For instance, in situations where multiple subpopulations with distinct data distributions exist, a prevailing objective is to ensure that the learned model does not adversely impact any of these subpopulations. One criterion designed to meet this objective, known as “min-max fairness” in the literature (Mohri et al., 2019; Abernethy et al., 2022), plays a pivotal role in mitigating the worst disadvantage experienced by any particular subpopulation.
- *Distributionally robust optimization/learning.* Another context where MDL naturally finds applications is group distributionally robust optimization and learning (DRO/DRL). Group DRO and DRL aim to develop algorithms that offer robust performance guarantees across a finite number of possible distributional models (Sagawa et al., 2019, 2020; Hashimoto et al., 2018; Hu et al., 2018; Xiong et al., 2023; Zhang et al., 2020; Wang et al., 2023; Deng et al., 2020), and have garnered substantial attention recently due to the pervasive need for robustness in modern decision-making (Carmon and Hausler, 2022; Asi et al., 2021; Haghtalab et al., 2022; Kar et al., 2019). When applying MDL to the context of group DRO/DRL, the resultant sample complexity reflects the price that needs to be paid for learning a robust solution.

The MDL framework is also closely related to other topics like multi-source domain adaptation, maximum aggregation, to name just a few (Mansour et al., 2008; Zhao et al., 2020; Bühlmann and Meinshausen, 2015; Guo, 2023).

In stark contrast to single-distribution learning, achieving data-efficient MDL necessitates adaptive sampling throughout the learning process, also known as on-demand sampling (Haghtalab et al., 2022). More specifically, pre-determining a sample-size budget for each distribution beforehand and sampling non-adaptively could result in a loss of sample efficiency, as we lack knowledge about the complexity of learning each distribution before the learning process begins. The question then comes down to how to optimally adapt the online sampling strategy to effectively tackle diverse data distributions.

Inadequacy of prior results. The sample complexity of MDL has been explored in a strand of recent works under various settings. Consider first the case where the hypothesis class \mathcal{H} comprises a *finite* number

²Here, the expectation on the left-hand side of (1) is taken over the randomness of both the datapoints (x, y) and the (randomized) hypothesis \hat{h} .

of hypotheses. If we sample non-adaptively and draw the same number of samples from each individual distribution \mathcal{D}_i , then this results in a total sample size exceeding the order of $\frac{k \log(|\mathcal{H}|)}{\varepsilon^2}$ (given that learning each distribution requires a sample size at least on the order of $\frac{\log(|\mathcal{H}|)}{\varepsilon^2}$). Fortunately, this sample size budget can be significantly reduced with the aid of adaptive sampling. In particular, the state-of-the-art approach, proposed by Haghtalab et al. (2022), accomplishes the objective (1) with probability at least $1 - \delta$ using $O\left(\frac{\log(|\mathcal{H}|) + k \log(k/\delta)}{\varepsilon^2}\right)$ samples. In comparison to agnostic learning on a single distribution, it only incurs an extra *additive cost* of $k \log(k/\delta)/\varepsilon^2$ as opposed to a multiplicative factor in k , thus underscoring the importance of adaptive sampling.

A more challenging scenario arises when \mathcal{H} has a finite Vapnik–Chervonenkis (VC) dimension d . The sample complexity for VC classes has only been settled for the realizable case (Blum et al., 2017; Chen et al., 2018; Nguyen and Zakyntinou, 2018), a special scenario where the loss function takes the form of $\ell(h, (x, y)) = \mathbb{1}\{h(x) \neq y\}$ and it is feasible to achieve zero mean loss. For the general non-realizable case, the best-known lower bound for such VC classes is (Haghtalab et al., 2022)³

$$\tilde{\Omega}\left(\frac{d+k}{\varepsilon^2}\right), \quad (2)$$

which serves as a theoretical benchmark. By first collecting $\tilde{O}\left(\frac{dk}{\varepsilon}\right)$ samples to help construct a cover of \mathcal{H} with reasonable resolution, Haghtalab et al. (2022) established a sample complexity upper bound of

$$(\text{Haghtalab et al., 2022}) \quad \tilde{O}\left(\frac{d+k}{\varepsilon^2} + \frac{dk}{\varepsilon}\right). \quad (3a)$$

Nevertheless, the term dk/ε in (3a) fails to match the lower bound (2); put another way, this term might result in a potentially large burn-in cost, as the optimality of this approach is only guaranteed (up to log factors) when the total sample size already exceeds an enormous threshold on the order of $\frac{d^2 k^2}{d+k}$. In an effort to alleviate this dk/ε factor, Awasthi et al. (2023) put forward an alternative algorithm — which utilizes an oracle to learn on a single distribution and obviates the need for computing an epsilon-net of \mathcal{H} — yielding a sample complexity of

$$(\text{Awasthi et al., 2023}) \quad \tilde{O}\left(\frac{d}{\varepsilon^4} + \frac{k}{\varepsilon^2}\right). \quad (3b)$$

However, this result (3b) might fall short of optimality as well, given that the scaling d/ε^4 is off by a factor of $1/\varepsilon^2$ compared with the lower bound (2). A more comprehensive list of past results can be found in Table 1.

Given the apparent gap between the state-of-the-art lower bound (2) and achievability bounds (3), a natural question arises:

Question: *Is it plausible to design a multi-distribution learning algorithm with a sample complexity of $\tilde{O}\left(\frac{d+k}{\varepsilon^2}\right)$ for VC classes, thereby matching the established lower bound (2)?*

Notably, this question has been posed as an open problem during the Annual Conference on Learning Theory (COLT) 2023; see Awasthi et al. (2023, Problem 1).

A glimpse of our main contributions. The present paper delivers some encouraging news: we come up with a new MDL algorithm that successfully resolves the aforementioned open problem in the affirmative. Specifically, focusing on a hypothesis class with VC dimension d and a collection of k data distributions, our main findings can be summarized as follows.

Theorem 1. *There exists an algorithm (see Algorithm 1 for details) such that: with probability exceeding $1 - \delta$, the randomized hypothesis h^{final} returned by this algorithm achieves*

$$\max_{1 \leq i \leq k} \mathbb{E}_{(x,y) \sim \mathcal{D}_i, h^{\text{final}}} [\ell(h^{\text{final}}, (x, y))] \leq \min_{h \in \mathcal{H}} \max_{1 \leq i \leq k} \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\ell(h, (x, y))] + \varepsilon,$$

³Here and throughout, $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ hide all logarithmic factors in $(k, d, \frac{1}{\varepsilon}, \frac{1}{\delta})$.

Paper	Sample complexity bound
Haghtalab et al. (2022)	$\frac{\log(\mathcal{H})+k}{\varepsilon^2}$
Haghtalab et al. (2022)	$\frac{d+k}{\varepsilon^2} + \frac{dk}{\varepsilon}$
Awasthi et al. (2023)	$\frac{d}{\varepsilon^4} + \frac{k}{\varepsilon^2}$
Peng (2023)	$\frac{d+k}{\varepsilon^2} \left(\frac{k}{\varepsilon}\right)^{o(1)}$
our work (Theorem 1)	$\frac{d+k}{\varepsilon^2}$
lower bound: Haghtalab et al. (2022)	$\frac{d+k}{\varepsilon^2}$

Table 1: Sample complexity bounds of MDL with k data distributions and a hypothesis class of VC dimension d . Here, we only report the polynomial dependency and hide all logarithmic dependency on $(k, d, \frac{1}{\varepsilon}, \frac{1}{\delta})$.

provided that the total sample size exceeds

$$\frac{d+k}{\varepsilon^2} \text{poly log} \left(k, d, \frac{1}{\varepsilon}, \frac{1}{\delta} \right). \quad (4)$$

The polylog factor in (4) will be specified momentarily. In a nutshell, we develop the first algorithm that provably achieves a sample complexity matching the lower bound (2) modulo logarithmic factors. Following the game dynamics template adopted in previous methods — namely, viewing MDL as a game between the learner (who selects the best hypothesis) and the adversary (who chooses the most challenging mixture of distributions) — our algorithm is built upon a novel and meticulously designed sampling scheme that deviates significantly from previous methods. Further, we extend our algorithm and theory to accommodate Rademacher classes, establishing a similar sample complexity bound when the weighted Rademacher complexity of the hypothesis class on n points is upper bounded by $O(\sqrt{\frac{d \log(n)}{n}})$.

Additionally, we solve two other open problems posed by Awasthi et al. (2023):

- *Oracle-efficient solutions.* An algorithm is said to be oracle-efficient if it only accesses \mathcal{H} through an empirical risk minimization (ERM) oracle (Dudík et al., 2020). Awasthi et al. (2023, Problem 4) then asked what the sample complexity of MDL is when confined to oracle-efficient paradigms. Encouragingly, our algorithm (i.e., Algorithm 1) adheres to the oracle-efficient criterion, thus uncovering that the sample complexity of MDL remains unchanged when restricted to oracle-efficient algorithms.
- *Necessity of randomization.* Both our algorithm and the most sample-efficient methods preceding our work produce randomized hypotheses. As discussed around Awasthi et al. (2023, Problem 3), a natural question concerns characterization of the sample complexity when restricting the final output to deterministic hypotheses from \mathcal{H} . Our result (see Theorem 2) delivers a negative message: under mild conditions, for any MDL algorithm, there exists a hard problem instance such that it requires at least $\Omega(dk/\varepsilon^2)$ samples to find a deterministic hypothesis $h \in \mathcal{H}$ that attains ε -accuracy. This constitutes an enormous sample complexity gap between what is achievable under randomized hypotheses and what is achievable using deterministic hypotheses.

Concurrent work. We shall mention that a concurrent work Peng (2023), posted around the same time as our work, also studied the MDL problem and significantly improved upon the prior results. More specifically, Peng (2023) established a sample complexity of $O\left(\frac{(d+k)\log(d/\delta)}{\varepsilon^2} \left(\frac{k}{\varepsilon}\right)^{o(1)}\right)$, which is optimal up to some sub-polynomial factor in k/ε ; in comparison, our sample complexity is optimal up to polylogarithmic factor. Additionally, it is worth noting that the algorithm therein relies upon a certain recursive structure to eliminate the non-optimal hypothesis, thus incurring exponential computational cost even when an ERM oracle is available.

Notation. Throughout this paper, we denote $[N] := \{1, \dots, N\}$ for any positive integer N . Let $\text{conv}(\mathcal{A})$ represent the convex hull of a set \mathcal{A} , and denote by $\Delta(n)$ the n -dimensional simplex for any positive integer

n . For two vectors $v = [v_i]_{1 \leq i \leq n}$ and $v' = [v'_i]_{1 \leq i \leq n}$ with the same dimension, we overload the notation by using $\max\{v, v'\} = [\max\{v_i, v'_i\}]_{1 \leq i \leq n}$ to denote the coordinate-wise maximum of v and v' . Also we say $v \leq v'$ iff $v_i \leq v'_i$ for all $i \in [n]$. For any random variable X , we use $\mathbb{V}[X]$ to denote its variance, i.e., $\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$. For any two distributions P and Q supported on \mathcal{X} , the Kullback-Leibler (KL) divergence from Q to P is defined and denoted by

$$\text{KL}(P \parallel Q) := \mathbb{E}_Q \left[\frac{dP}{dQ} \log \frac{dP}{dQ} \right]. \quad (5)$$

2 Problem formulation

This section formulates the multi-distribution learning problem. We assume throughout that each datapoint takes the form of $(x, y) \in \mathcal{X} \times \mathcal{Y}$, with \mathcal{X} (resp. \mathcal{Y}) the instance space (resp. label space).

Learning from multiple distributions. The problem setting encompasses several elements below.

- *Hypothesis class.* Suppose we are interested in a hypothesis class \mathcal{H} , comprising a set of candidate functions from the instance space \mathcal{X} to the label space \mathcal{Y} . Overloading the notation, we use h_π to represent a *randomized hypothesis* associated with a probability distribution $\pi \in \Delta(\mathcal{H})$, meaning that a hypothesis h from \mathcal{H} is randomly selected according to distribution π . Additionally, the VC dimension (Vapnik et al., 1994) of \mathcal{H} is assumed to be

$$\text{VC-dim}(\mathcal{H}) = d. \quad (6)$$

- *Loss function.* Suppose we are given a loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \rightarrow [-1, 1]$, so that $\ell(h, (x, y))$ quantifies the risk of using hypothesis $h \in \mathcal{H}$ to make prediction on a datapoint $(x, y) \in \mathcal{X} \times \mathcal{Y}$ (i.e., predicting y based on x). One example is the 0-1 loss function $\ell(h, (x, y)) = \mathbb{1}\{h(x) \neq y\}$, which is often used to measure the misclassification error.
- *(Multiple) data distributions.* Suppose that there are k data distributions of interest supported on $\mathcal{X} \times \mathcal{Y}$, denoted by $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_k\}$. We are permitted to draw independent samples from each of these data distributions.

Given a target accuracy level $\varepsilon \in (0, 1)$, our objective is to identify a (randomized) hypothesis, represented by h_π with $\pi \in \Delta(\mathcal{H})$, such that

$$\max_{1 \leq i \leq k} \mathbb{E}_{(x, y) \sim \mathcal{D}_i, h_\pi \sim \pi} [\ell(h_\pi, (x, y))] \leq \min_{h \in \mathcal{H}} \max_{1 \leq i \leq k} \mathbb{E}_{(x, y) \sim \mathcal{D}_i} [\ell(h, (x, y))] + \varepsilon. \quad (7)$$

Sampling and learning processes. In order to achieve the aforementioned goal (7), we need to draw samples from the available data distributions in \mathcal{D} , and the current paper focuses on sampling in an online fashion. More precisely, the learning process proceeds as follows: in each step τ ,

- the learner selects $i_\tau \in [k]$ based on the previous samples;
- the learner draws an *independent* sample (x_τ, y_τ) from the data distribution \mathcal{D}_{i_τ} .

The sample complexity of a learning algorithm thus refers to the total number of samples drawn from \mathcal{D} throughout the learning process. A desirable learning algorithm would yield an ε -optimal (randomized) hypothesis (i.e., a hypothesis that achieves (7)) using as few samples as possible.

3 Algorithm

In this section, we present our proposed algorithm. Before proceeding, we find it convenient to introduce some notation concerning the loss under mixed distributions. Specifically, for any distribution $w = [w_i]_{1 \leq i \leq k} \in \Delta(k)$ and any hypothesis $h \in \mathcal{H}$, the risk over the mixture $\sum_{i \in [k]} w_i \mathcal{D}_i$ of data distributions is denoted by:

$$L(h, w) := \sum_{i=1}^k w_i \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\ell(h, (x, y))]; \quad (8a)$$

similarly, the risk of a randomized hypothesis h_π (associated with $\pi \in \Delta(\mathcal{H})$) over $\sum_{i \in [k]} w_i \mathcal{D}_i$ is given by

$$L(h_\pi, w) := \sum_{i=1}^k w_i \mathbb{E}_{(x,y) \sim \mathcal{D}_i, h \sim \pi} [\ell(h_\pi, (x, y))] = \mathbb{E}_{h \sim \pi} [L(h, w)]. \quad (8b)$$

Algorithm 1: Hedge for Multi-distribution Learning (MDL – Hedge)

1 **input:** k data distributions $\{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_k\}$, hypothesis class \mathcal{H} , target accuracy level ε , target success rate $1 - \delta$.
2 **hyper-parameters:** stepsize $\eta = \frac{1}{100}\varepsilon$, number of rounds $T = \frac{20000 \log(\frac{k}{\delta})}{\varepsilon^2}$, auxiliary accuracy level $\varepsilon_1 = \frac{1}{100}\varepsilon$, auxiliary sub-sample-size $T_1 := \frac{4000(k \log(k/\varepsilon_1) + d \log(kd/\varepsilon_1) + \log(1/\delta))}{\varepsilon_1^2}$.
3 **initialization:** for all $i \in [k]$, set $W_i^1 = 1$, $\hat{w}_i^0 = 0$ and $n_i^0 = 0$; $\mathcal{S} = \emptyset$.
4 **for** $t = 1, 2, \dots, T$ **do**
5 set $w^t = [w_i^t]_{1 \leq i \leq k}$ and $\hat{w}^t = [\hat{w}_i^t]_{1 \leq i \leq k}$, with $w_i^t \leftarrow \frac{W_i^t}{\sum_j W_j^t}$ and $\hat{w}_i^t \leftarrow \hat{w}_i^{t-1}$ for all $i \in [k]$.
6 /* recompute \hat{w}^t & draw new samples for \mathcal{S} only if w^t changes sufficiently. */
7 **if** there exists $j \in [k]$ such that $w_j^t \geq 2\hat{w}_j^{t-1}$ **then**
8 $\hat{w}_i^t \leftarrow \max\{w_i^t, \hat{w}_i^{t-1}\}$ for all $i \in [k]$;
9 **for** $i = 1, \dots, k$ **do**
10 $n_i^t \leftarrow \lceil T_1 \hat{w}_i^t \rceil$;
11 draw $n_i^t - n_i^{t-1}$ independent samples from \mathcal{D}_i , and add these samples to \mathcal{S} .
12 /* estimate the near-optimal hypothesis for weighted data distributions. */
13 compute $h^t \leftarrow \arg \min_{h \in \mathcal{H}} \hat{L}^t(h, w^t)$, where
14
$$\hat{L}^t(h, w^t) := \sum_{i=1}^k \frac{w_i^t}{n_i^t} \cdot \sum_{j=1}^{n_i^t} \ell(h, (x_{i,j}, y_{i,j})) \quad (9)$$

15 with $(x_{i,j}, y_{i,j})$ being the j -th datapoint from \mathcal{D}_i in \mathcal{S} .
16 /* estimate the loss vector and execute weighted updates. */
17 $\bar{w}_i^t \leftarrow \max_{1 \leq r \leq t} w_i^r$ for all $i \in [k]$.
18 **for** $i = 1, \dots, k$ **do**
19 draw $\lceil k \bar{w}_i^t \rceil$ independent samples — denoted by $\{(x_{i,j}^t, y_{i,j}^t)\}_{j=1}^{\lceil k \bar{w}_i^t \rceil}$ — from \mathcal{D}_i , and set
20
$$\hat{r}_i^t = \frac{1}{\lceil k \bar{w}_i^t \rceil} \sum_{j=1}^{\lceil k \bar{w}_i^t \rceil} \ell(h^t, (x_{i,j}^t, y_{i,j}^t));$$

21 update the weight as $W_i^{t+1} = W_i^t \exp(\eta \hat{r}_i^t)$. // Hedge updates.
22 **output:** a randomized hypothesis h^{final} uniformly distributed over $\{h^t\}_{t=1}^T$.

Following the game dynamics proposed in previous works (Awasthi et al., 2023; Haghtalab et al., 2022), our algorithm alternates between computing the most favorable hypothesis (performed by the learner) and estimating the most challenging mixture of data distributions (performed by the adversary), with the aid of

no-regret learning algorithms (Roughgarden, 2016; Shalev-Shwartz, 2012). More specifically, in each round t , our algorithm performs the following two steps:

- (a) Given a mixture of data distributions $\mathcal{D}^{(t)} = \sum_{i \in [k]} w_i^t \mathcal{D}_i$ (with $w^t = [w_i^t]_{i \in [k]} \in \Delta(k)$), we construct a dataset to compute a hypothesis h^t that nearly minimizes the loss under $\mathcal{D}^{(t)}$, namely,

$$h^t \approx \arg \min_{h \in \mathcal{H}} L(h, w^t). \quad (10)$$

This is accomplished by calling an empirical risk minimization oracle.

- (b) Given hypothesis h^t , we compute an updated weight vector $w^{t+1} \in \Delta(k)$ — and hence an updated mixed distribution $\mathcal{D}^{(t+1)} = \sum_{i \in [k]} w_i^{t+1} \mathcal{D}_i$. The weight updates are carried out using the celebrated Hedge algorithm (Freund and Schapire, 1997) designed for online adversarial learning,⁴ in an attempt to achieve low regret even when the loss vectors are adversarially chosen. More precisely, we run

$$w_i^{t+1} \propto w_i^t \exp(\eta \hat{r}_i^t), \quad i \in [k], \quad (11)$$

where the loss vector $\hat{r}^t = [\hat{r}_i^t]_{i \in [k]}$ contains the empirical loss of h^t under each data distribution, i.e.,

$$\hat{r}_i^t \approx \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\ell(h^t, (x, y))], \quad i \in [k],$$

computed over another set of data samples.

At the end of the algorithm, we output a randomized hypothesis h^{final} that is uniformly distributed over the hypothesis iterates $\{h^t\}_{1 \leq t \leq T}$ over all T rounds, following common practice in online adversarial learning.

While the above paradigm has been adopted in past works (Awasthi et al., 2023; Haghtalab et al., 2022), the resulting sample complexity depends heavily upon how data samples are collected and utilized throughout the learning process. For instance, Awasthi et al. (2023, Algorithm 1) draws *fresh data* at each step of every round, in order to ensure reliable estimation of the loss function of interest through elementary concentration inequalities. This strategy, however, becomes wasteful over time, constituting the main source of its sample sub-optimality.

In order to make the best use of data, we propose the following key strategies.

- *Sample reuse in Step (a).* In stark contrast to Awasthi et al. (2023, Algorithm 1) that draws new samples for estimating each h^t , we propose to reuse all samples collected in Step (a) up to the t -th round to assist in computing h^t . As will be made precise in lines 6-11 of Algorithm 1, we shall maintain a *growing dataset* \mathcal{S} for conducting Step (a) throughout, ensuring that there are n_i^t samples drawn from distribution \mathcal{D}_i in the t -th round. These datapoints are employed to construct an empirical loss estimator $\hat{L}^t(h, w^t)$ for each $h \in \mathcal{H}$ in each round t , with the aim of achieving uniform convergence $|\hat{L}^t(h, w^t) - L(h, w^t)| \leq O(\varepsilon)$ over all $h \in \mathcal{H}$. More detailed explanations are provided in Section 4.1.
- *Weighted sampling for Step (b).* As shown in line 14 of Algorithm 1, in each round t , we sample each \mathcal{D}_i a couple of times to compute the empirical estimator for $\mathbb{E}_{(x,y) \in \mathcal{D}_i} [\ell(h^t, (x, y))]$, where the number of samples depends upon the running weights $\{w_i^\tau\}$. More precisely, we collect $\lceil k \bar{w}_i^t \rceil$ fresh samples from each \mathcal{D}_i , where $\bar{w}_i^t := \max_{1 \leq \tau \leq t} w_i^\tau$ is the maximum weight assigned to \mathcal{D}_i up to now. Informally speaking, this strategy ensures reduced variance of the estimators and ultimately allows for an improved bound for $\sum_{i=1}^k \max_{1 \leq t \leq T} w_i^t$. The interested reader is referred to Section 4.2 and Lemma 17 for more detailed explanations.

The whole procedure can be found in Algorithm 1.

⁴Note that the Hedge algorithm is closely related to Exponentiated Gradient Descent, Multiplicative Weights Update, Online Mirror Descent, etc (Arora et al., 2012; Shalev-Shwartz, 2012; Hazan, 2022).

4 A glimpse of key technical novelty

In this section, we highlight two technical novelty that empowers our analysis: (i) uniform convergence of the weighted sampling estimator that allows for sample reuse (see Section 4.1), and (ii) tight control of certain ℓ_1/ℓ_∞ norm of the iterates $\{w^t\}_{1 \leq t \leq T}$ that dictates the sample efficiency (see Section 4.2). Given that Haghtalab et al. (2022) already established near-optimal upper bounds when $k = O(1/\varepsilon)$ (cf. the 2nd row in Table 1), our analysis should focus on the regime where $k \geq 100/\varepsilon$.

4.1 Towards sample reuse: uniform concentration and a key quantity

Recall that in Algorithm 1, we invoke the empirical risk estimator $\hat{L}^t(h, w^t)$ as an estimate of the true risk of hypothesis h over the weighted distribution specified by w^t (cf. (9)). In order to facilitate sample reuse when constructing such risk estimators across all iterations, it is desirable to establish uniform concentration results to control the errors of such risk estimators throughout the execution of the algorithm. Towards this end, our analysis strategy proceeds as follows.

Step 1: concentration for any fixed set of parameters. Consider any given set of integers $n = \{n_i\}_{i=1}^k$ and any given vector $w \in \Delta(k)$. Suppose, for each $i \in [k]$, we have n_i i.i.d. samples drawn from \mathcal{D}_i — denoted by $\{(x_{i,j}, y_{i,j})\}_{j=1}^{n_i}$ — and let us look at the empirical risk estimator,

$$\hat{L}_n(h, w) := \sum_{i=1}^k w_i \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(h, (x_{i,j}, y_{i,j})), \quad (12)$$

which is a sum of independent random variables. Evidently, for a given hypothesis h , the variance of $\hat{L}_n(h, w)$ is upper bounded by

$$\text{Var}(\hat{L}_n(h, w)) \leq \sum_{i=1}^k \frac{w_i^2}{n_i} \leq \left(\sum_{i=1}^k w_i \right) \frac{1}{\min_i n_i / w_i} = \frac{1}{\min_i n_i / w_i}.$$

Assuming that the central limit theorem is applicable, one can derive

$$\mathbb{P} \left\{ |\hat{L}_n(h, w) - L(h, w)| \geq \varepsilon \right\} \lesssim \exp \left(- \frac{\varepsilon^2}{2 \text{Var}(\hat{L}_n(h, w))} \right) \lesssim \exp \left(- \frac{\varepsilon^2}{2} \min_i \frac{n_i}{w_i} \right).$$

Armed with this result, we can extend it to accommodate all $h \in \mathcal{H}$ through the union bound. For a VC class with $\text{VC-dim}(\mathcal{H}) = d$, the celebrated Sauer–Shelah lemma (Wainwright, 2019, Proposition 4.18) tells us that the set of hypotheses can be effectively compressed into a subset with cardinality no larger than $\exp(\tilde{O}(d))$. Taking the union bound then yields

$$\mathbb{P} \left(\max_{h \in \mathcal{H}} |\hat{L}_n(h, w) - L(h, w)| \geq \varepsilon \right) \lesssim \exp \left(\tilde{O}(d) - \frac{\varepsilon^2}{2} \min_i \frac{n_i}{w_i} \right)$$

Step 2: uniform concentration. Next, we would like to extend the above result to establish uniform concentration over all n and w of interest. Towards this, we shall invoke the union bound as well as the standard epsilon-net arguments. Let the set $\mathcal{X} \subseteq \Delta(k)$ be a proper discretization of $\Delta(k)$, with cardinality $\exp(\tilde{O}(k))$. In addition, given the trivial upper bound $n_i \leq T_1$ for all $i \in [k]$, we know that there exist at most $T_1^k = \exp(\tilde{O}(k))$ possible combinations of $\{n_i\}_{i \in [k]}$. We can then apply the union bound to show that

$$\mathbb{P} \left\{ \exists w \in \mathcal{X} \text{ and feasible } n \text{ s.t. } |\hat{L}_n(h, w) - L(h, w)| \geq \varepsilon \right\} \lesssim \exp \left(\tilde{O}(k) + \tilde{O}(d) - \frac{\varepsilon^2}{2} \min_i \frac{n_i}{w_i} \right). \quad (13)$$

When the discretized set \mathcal{X} is chosen to have sufficient resolution, we can straightforwardly employ the standard covering argument to extend the above inequality to accommodate all $w \in \Delta(k)$.

Key takeaways. The above arguments reveal the following high-probability property: *whenever* we collect $n = \{n_i\}_{i=1}^k$ samples in the learning process, we could obtain ε -approximation $\hat{L}_n(h, w)$ (see (12)) of $L(h, w)$ for all $h \in \mathcal{H}$ and all $w \in \Delta(k)$ with high probability, provided that

$$\min_i \frac{n_i}{w_i} \gtrsim \tilde{O}\left(\frac{k+d}{\varepsilon^2}\right). \quad (14)$$

This makes apparent the pivotal role of the quantity $\min_i n_i/w_i$. In our algorithm, we design the update rule (cf. line 9 of Algorithm 1), so as to guaranteed that

$$\min_i \frac{n_i^t}{w_i^t} \gtrsim T_1 \geq \tilde{\Omega}\left(\frac{k+d}{\varepsilon^2}\right) \quad (15)$$

for all $1 \leq t \leq T$. In fact, this explains our choice of T_1 in Algorithm 1. Crucially, the aforementioned uniform concentration result allows us to reuse samples throughout the learning process instead of drawing fresh samples to estimate $L(h, w^t)$ in each round t (the latter approach clearly loses data efficiency). To conclude, to guarantee ε -uniform convergence for all rounds, it suffices to choose $T_1 = \tilde{\Omega}\left(\frac{k+d}{\varepsilon^2}\right)$.

Finally, recall that $n_i^t \gtrsim T_1 \bar{w}_i^t$ for each $i \in [k]$ and $t \leq T$, with $\bar{w}_i^t := \max_{1 \leq \tau \leq t} w_i^\tau$; taking $n_i^t \asymp T_1 \bar{w}_i^t$ (as opposed to $n_i^t \asymp T_1 w_i^t$) ensures that the sample size n_i^t is monotonically non-decreasing in t . With (15) in mind, the total number of samples collected within T rounds in Algorithm 1 obeys

$$\frac{1}{T_1} \sum_{i=1}^k n_i^T \asymp \sum_{i=1}^k \bar{w}_i^T =: \|\bar{w}^T\|_1. \quad (16)$$

This threshold $\|\bar{w}^T\|_1$ — or equivalently, the ℓ_1/ℓ_∞ norm of $\{w^t\}_{1 \leq t \leq T}$ — is a critical quantity that we wish to control; in particular, in the desirable scenario where $\|\bar{w}^T\|_1 \leq \tilde{O}(1)$, the total sample size obeys $\sum_{i=1}^k n_i^T \asymp T_1 \|\bar{w}^T\|_1 = \tilde{O}\left(\frac{k+d}{\varepsilon^2}\right)$.

4.2 Bounding the key quantity $\|\bar{w}^T\|_1$ by controlling the Hedge trajectory

Perhaps the most innovative (and most challenging) part of our analysis lies in controlling the ℓ_1/ℓ_∞ norm of $\{w_i^t\}_{1 \leq t \leq T}$, whose critical importance has been pointed out in Section 4.1.

Towards this end, the key lies in carefully tracking the dynamics of the Hedge algorithm. To elucidate the high-level idea, we first consider the following minimax optimization problem w.r.t. the set of loss vectors in the convex hull of a set \mathcal{Y} :

$$\min_{y \in \text{conv}(\mathcal{Y})} \max_{w \in \Delta(k)} w^\top y \quad (\text{or equivalently, } \max_{w \in \Delta(k)} \min_{y \in \text{conv}(\mathcal{Y})} w^\top y), \quad (17)$$

where the equivalence arises from von Neumann's minimax theorem (v. Neumann, 1928). Let us look at the following algorithm (cf. Algorithm 2) tailored to this minimax problem, assuming perfect knowledge about the loss vector.⁵

Algorithm 2: The Hedge algorithm for bilinear games.

- 1 **Input:** $\mathcal{Y} \subseteq [-1, 1]^k$, target accuracy level $\varepsilon \in (0, 1)$.
 - 2 **Initialization:** $T = \frac{100 \log(k)}{\varepsilon^2}$, $\eta = \frac{1}{10} \varepsilon$, and $W_i^1 = 1$ for all $1 \leq i \leq k$.
 - 3 **for** $t = 1, 2, \dots, T$ **do**
 - 4 compute $w_i^t \leftarrow \frac{W_i^t}{\sum_j W_j^t}$ for every $1 \leq i \leq k$.
 - 5 compute $y^t \leftarrow \arg \min_{y \in \mathcal{Y}} \langle w^t, y \rangle$.
 - 6 update $W_i^{t+1} \leftarrow W_i^t \exp(\eta y_i^t)$ for every $1 \leq i \leq k$.
-

This algorithm is often referred to as the Hedge algorithm, which is known to yield an ε -minimax solution within $O\left(\frac{\log(k)}{\varepsilon^2}\right)$ iterations. A challenging question relevant to our analysis is:

⁵Note that in Algorithm 1, we can only estimate the loss vector using the collected samples. Additional efforts are needed to reduce the variability (see line 14 in Algorithm 1 and Lemma 17).

Question: can we bound $\|\bar{w}^T\|_1 = \sum_{i=1}^k \max_{1 \leq t \leq T} w_i^t$ by poly-logarithmic terms?

As it turns out, we can answer this question affirmatively (see Lemma 3), and the key ideas will be elucidated in the remainder of this section.

To streamline presentation of our techniques, we assume without loss of generality that

$$\min_{y \in \text{conv}(\mathcal{Y})} \max_{w \in \Delta(k)} w^\top y = \max_{w \in \Delta(k)} \min_{y \in \text{conv}(\mathcal{Y})} w^\top y = 0.$$

Under this assumption, it is easily seen that [YXC: TODO]

$$\langle w^t, y^t \rangle = \min_{y \in \mathcal{Y}} \langle w^t, y \rangle \leq 0, \quad \forall t \in [T] \quad \text{and} \quad T^{-1} \sum_{t=1}^T \langle w^t, y^t \rangle \geq -O(\varepsilon). \quad (18)$$

Let us also assume for the moment that $-\langle w^t, y^t \rangle = O(\varepsilon)$ for any $t \in [T]$.⁶

4.2.1 Doubling w_j needs $\tilde{O}(1/\varepsilon^2)$ steps

[YXC: Need more edits.]

Instead of bounding $\|\bar{w}^T\|_1$ directly, our first attempt is to show that:

- there exist at most $\tilde{O}(1)$ coordinates $i \in [k]$ obeying $\max_{1 \leq t \leq T} w_i^t \geq 1/4$ (or some other universal constant).

In other words, we would like to show that the cardinality of the following set is small:

$$\mathcal{W} := \{i \in [k] \mid \max_{1 \leq t \leq T} w_i^t \geq 1/4\}. \quad (19)$$

To do so, note that for small stepsize η , one can find, for each $i \in \mathcal{W}$, a time interval $[s_i, e_i] \subseteq [0, T]$ obeying

$$1/16 \leq w_i^{s_i} \leq 1/8, \quad w_i^{e_i} \geq 1/4 \quad \text{and} \quad w_i^t \geq 1/8, \quad \forall t \in (s_i, e_i]. \quad (20)$$

In words, w_i^t at least doubles from $t = s_i$ to $t = e_i$. We claim for the moment that

$$e_i - s_i \geq \Omega(1/\varepsilon^2) \quad \forall i \in \mathcal{W}. \quad (21)$$

Additionally, observe that for any t , there exist at most 8 coordinates $i \in \mathcal{W}$ such that $s_i \leq t \leq e_i$ (since $w_i^t \geq 1/8$ for $t \in [s_i, e_i]$). This reveals that

$$8T \geq \sum_{i \in \mathcal{W}} (e_i - s_i) \geq |\mathcal{W}| \cdot \Omega(1/\varepsilon^2),$$

which combined with our choice of $T = O(\log(k/\delta)/\varepsilon^2)$ (cf. line 2 of Algorithm 1) yields

$$|\mathcal{W}| \leq O(T\varepsilon^2) = O(\log(k/\delta)).$$

Proof strategy for (21). In this proof, we shall exploit properties of a bilinear game where the opponent plays the best response. From (18) and the fact that y^t is the best response for w^t in each t , one sees that

$$\langle w^t, y^t \rangle \leq 0, \quad \langle w^t, y^\tau \rangle \geq \langle w^t, y^t \rangle, \quad \forall 1 \leq t \leq \tau \leq T. \quad (22)$$

Armed with (22), we claim that

$$\text{KL}(w^{t_1} \parallel w^{t_2}) \leq O(\eta^2(t_2 - t_1)), \quad 1 \leq t_1 \leq t_2 \leq T.$$

⁶While this assumption is not valid in most cases, one can divide $[T]$ into $\tilde{O}(1)$ disjoint subsets and then tackle each subset. As a trade-off, this strategy leads to some additional logarithmic factors (see Lemma 17).

To show this, our calculation proceeds as follows:

$$\begin{aligned}\text{KL}(w^{t_1} \| w^{t_2}) &= \sum_{i=1}^k w_i^{t_1} \log \left(\frac{w_i^{t_1}}{w_i^{t_2}} \right) = \sum_{i=1}^k w_i^{t_1} \log \left(\frac{\sum_{j=1}^k W_j^{t_2}}{\sum_{j=1}^k W_j^{t_1}} \right) + \sum_{i=1}^k w_i^{t_1} \log \left(\frac{W_i^{t_1}}{W_i^{t_2}} \right) \\ &= \log \left(\frac{\sum_{i=1}^k W_i^{t_2}}{\sum_{i=1}^k W_i^{t_1}} \right) - \eta \sum_{\tau=t_1}^{t_2-1} \sum_{i=1}^k w_i^{t_1} y_i^\tau \leq \log \left(\frac{\sum_{i=1}^k W_i^{t_2}}{\sum_{i=1}^k W_i^{t_1}} \right) - \eta(t_2 - t_1) \langle w^{t_1}, y^{t_1} \rangle,\end{aligned}$$

where the second identity holds since $w_i^t = \frac{W_i^t}{\sum_j W_j^t}$, the third identity is valid since $W_i^{t_2} = W_i^{t_1} \exp(\eta \sum_{\tau=t_1}^{t_2-1} y_i^\tau)$, and the last relation results from (22). In light of the properties

$$\eta = \Theta(\varepsilon), \quad -(w^{t_1})^\top y^{t_1} = O(\varepsilon), \quad \log \left(\frac{\sum_{i=1}^k W_i^{t_2}}{\sum_{i=1}^k W_i^{t_1}} \right) \leq O((t_2 - t_1)\eta^2), \quad (23)$$

we can further obtain $\text{KL}(w^{t_1} \| w^{t_2}) = O(\varepsilon^2(t_2 - t_1))$ and hence

$$t_2 - t_1 \geq \Omega(\varepsilon^{-2} \text{KL}(w^{t_1} \| w^{t_2})). \quad (24)$$

By taking $t_1 = s_i$ and $t_2 = e_i$, we can combine (24) and Pinsker's inequality to obtain

$$\begin{aligned}e_i - s_i &\geq \Omega(\varepsilon^{-2} \text{KL}(w^{s_i} \| w^{e_i})) \geq \Omega(\varepsilon^{-2} (\text{TV}(w^{s_i}, w^{e_i}))^2) \\ &\geq \Omega(\varepsilon^{-2} (w_i^{s_i} - w_i^{e_i})^2) = \Omega(1/\varepsilon^2).\end{aligned}$$

□

4.2.2 Coping with the segments

Naturally, one would hope to generalize the arguments in Section 4.2.1 to bound the size of the set:

$$\mathcal{W}(p) := \{i \in [k] \mid \max_{1 \leq t \leq T} w_i^t \in [2p, 4p]\} \quad (25)$$

for any $p \in [0, 1]$. Nevertheless, the arguments above fall short of delivering a desirable bound on $|\mathcal{W}(p)|$ when p is small. To be more specific, for each $i \in \mathcal{W}(p)$, let $[s_i, e_i]$ represent a time interval such that

$$p/2 \leq w_i^{s_i} \leq p, \quad w_i^{e_i} \geq 2p \quad \text{and} \quad w_i^t \geq p \quad \text{for any } s_i < t \leq e_i. \quad (26)$$

While we can derive $e_i - s_i \geq \Omega(p/\varepsilon^2)$ via the arguments in (24), this bound does not readily allow one to bound $|\mathcal{W}(p)|$ by $\tilde{O}(1/p)$, since the intervals $[s_i, e_i]$ for different i 's might have lots of overlaps.

To address this issue, we make the following key observation: if there exist some coordinates $i \in \mathcal{W}(p)$ sharing similar $[s_i, e_i]$, we can obtain an improved bound. For example, suppose that for each $i \in \overline{\mathcal{W}}(p) \subseteq \mathcal{W}(p)$, one has $s_i = s$ and $e_i = e$, then one can derive $e - s \geq \Omega(|\overline{\mathcal{W}}(p)|p/\varepsilon^2)$, which strengthens the original bound $\Omega(p/\varepsilon^2)$ if $|\overline{\mathcal{W}}(p)|$ is large. As such, it is helpful to merge those coordinates with similar $[s_i, e_i]$. To facilitate analysis, we introduce the notion of “segments.”

Definition 1 (Segment). *For any $p, x > 0$ and $i \in [k]$, we say that (t_1, t_2) is a (p, q, x) -segment if there exists a subset $\mathcal{I} \subseteq [k]$ such that*

- (i) $\sum_{i \in \mathcal{I}} w_i^{t_1} \in [p/2, p]$,
- (ii) $\sum_{i \in \mathcal{I}} w_i^{t_2} \geq p \exp(x)$,
- (iii) $\sum_{i \in \mathcal{I}} w_i^t \geq q$ for any $t_1 \leq t \leq t_2$.

We shall refer to t_1 as the starting point and t_2 as the ending point, and call \mathcal{I} the index set. Moreover, two segments (s_1, e_1) and (s_2, e_2) are said to be disjoint if $s_1 < e_1 \leq s_2 < e_2$ or $s_2 < e_2 \leq s_1 < e_1$.

For any $(p, p/2, x)$ -segment (t_1, t_2) , repeating the arguments in (24) allows one to derive [YXC: TODO]

$$t_2 - t_1 \geq \Omega(px^2/\varepsilon^2).$$

For each $i \in \mathcal{W}(p)$ (see its definition in (25)), there exists a $(\frac{p}{4}, \frac{p}{8}, \log(2))$ -segment (s_i, e_i) with index set $\mathcal{I} = \{i\}$. It then follows that

$$e_i - s_i \geq \Omega(p/\varepsilon^2) \quad \text{for each } i \in \mathcal{W}(p). \quad (27)$$

As a result, if the segments (s_i, e_i) are mutually disjoint, $|\mathcal{W}(p)|$ is at most $\tilde{O}(1/p)$. More generally, if we can divide the $|\mathcal{W}(p)|$ segments into disjoint blocks such that the segments belonging to the same block share the same starting and ending points, then we can also derive $|\mathcal{W}(p)| \leq \tilde{O}(1/p)$. Suppose that we have ℓ blocks with the i -th block (with starting-ending points as $(\tilde{s}^i, \tilde{e}^i)$) containing m_i coordinates, and suppose that $\tilde{s}^1 < \tilde{e}^1 \leq \tilde{s}^2 < \tilde{e}^2 \leq \dots \leq \tilde{s}^\ell < \tilde{e}^\ell$. Then from Definition 1, $(\tilde{s}^i, \tilde{e}^i)$ forms a $(m_i \cdot \frac{p}{4}, m_i \cdot \frac{p}{8}, \log(2))$ -segment with index set as the m_i coordinates in the i -th block, thereby indicating that

$$\tilde{e}^i - \tilde{s}^i \geq \Omega(m_i p / \varepsilon^2).$$

Summing over $1 \leq i \leq \ell$ leads to

$$|\mathcal{W}(p)| \leq \sum_{i=1}^{\ell} m_i \leq O\left(\sum_{i=1}^{\ell} \frac{(\tilde{e}^i - \tilde{s}^i)\varepsilon^2}{p}\right) \leq O\left(\frac{T\varepsilon^2}{p}\right) = \tilde{O}\left(\frac{1}{p}\right), \quad (28)$$

which in turn implies that $\sum_{i \in \mathcal{W}(p)} \max_{1 \leq t \leq T} w_i^t \leq \tilde{O}(1)$. With standard doubling arguments, it follows that $\sum_{i \in [k]} \max_{1 \leq t \leq T} w_i^t = \tilde{O}(1)$.

In light of the above observation, we introduce the following concept of *regular* configurations.⁷

Definition 2 (Configuration). A configuration Conf is a set of intervals $\text{Conf} = \{[a_i, b_i]\}_{i=1}^m$ obeying $b_i > a_i$ for each $i \in [m]$ (note that repeated elements are allowed). A configuration Conf is said to be *regular* if, for any $i, j \in [m]$, one of the following three properties holds:

- (a) $a_i < b_i \leq a_j < b_j$;
- (b) $a_j < b_j \leq a_i < b_i$;
- (c) $a_i = a_j, b_i = b_j$.

In words, (28) asserts that if $\{[s_i, e_i]\}_{i \in \mathcal{W}(p)}$ forms a *regular* configuration, then we have $|\mathcal{W}(p)| = \tilde{O}(1/p)$. However, a general configuration of the segments might be *irregular* because it is possible that two segments are not disjoint (see Figure 2). To address this issue, we find it helpful to construct a regular configuration with sub-segments⁸ of the original segments.

Our first step is to align one side of the segments. We then divide the whole learning process into disjoint blocks such that the segments in each block have a common inner point (see Figure 3). In the meantime, a segment is discarded if it intersects with more than one blocks. We show that there exists a regular configuration of the blocks such that at most $(1 - \frac{1}{3(\log_2(T)+1)})|\mathcal{W}(p)|$ segments are discarded. In other words, at least $\frac{1}{3(\log_2(T)+1)} \cdot |\mathcal{W}(p)|$ segments are contained by these blocks. Since the blocks are disjoint, it suffices to operate on one block. Without loss of generality, assume that there is only one block, which also means that all segments have a common inner point t_{mid} . Then we divide this block into two parts according to t_{mid} (see Figure 4). Given that

$$\log(w_i^{e_i}/w_i^{t_{\text{mid}}}) + \log(w_i^{t_{\text{mid}}}/w_i^{s_i}) \geq \log(2),$$

⁷See an example in Figure 1.

⁸A sub-segment is equivalent to a sub-interval of a segment. We refer to Figure 6 for more details.

one has: either (s_i, t_{mid}) is a $(\frac{p}{4}, \frac{p}{8}, \frac{\log(2)}{2})$ -segment with index i , or (t_{mid}, e_i) is a $(w_i^{t_{\text{mid}}}, \frac{p}{8}, \frac{\log(2)}{2})$ -segment with index set $\{i\}$. Since $w_i^{t_{\text{mid}}} \in [p/4, p/2]$, we can roughly view (t_{mid}, e_i) as a $(\frac{p}{4}, \frac{p}{8}, \frac{\log(2)}{2})$ -segment in the latter case. Therefore, one can find at least $\frac{|\mathcal{W}(p)|}{3(\log_2(T)+1)} \frac{1}{2}$ different $(\frac{p}{4}, \frac{p}{8}, \frac{\log(2)}{2})$ segments with the same starting (ending) points.

Without loss of generality, we assume that these $(\frac{p}{4}, \frac{p}{8}, \frac{\log(2)}{2})$ -segments share a common starting point. We denote the common starting point as e_0 , and re-order the coordinates so that e_i is non-decreasing in i (see Figure 3). Then we construct $O(\log_2(k))$ regular configurations using a recursion (see Lemma 16). This allows us to show that at least one of these configurations contains $\tilde{O}(|\mathcal{W}(p)|) \left(\frac{p}{4}, \frac{p}{8}, \frac{\log(2)}{2(\log_2(k)+2)}\right)$ -segments (see Figures 5-6).

5 Analysis for VC classes (proof of Theorem 1)

The main steps for establishing Theorem 1 lie in proving three key lemmas, as stated below.

The first lemma is concerned with the hypothesis $h^t = \arg \min_{h \in \mathcal{H}} \hat{L}^t(h, w^t)$ (cf. line 11 of Algorithm 1); in words, h^t is the minimizer of the empirical loss function $\hat{L}^t(\cdot, w^t)$, computed using samples obtained up to the t -th round. The following lemma tells us that: even though h^t is an empirical minimizer, it almost optimizes the weighted population loss $L(\cdot, w^t)$. In other words, this lemma justifies that the adaptive sampling scheme proposed in Algorithm 1 ensures faithfulness of the empirical loss and its minimizer.

Lemma 1. *With probability at least $1 - \delta/4$,*

$$L(h^t, w^t) \leq \min_{h \in \mathcal{H}} L(h, w^t) + \varepsilon_1 \quad (29)$$

holds for all $1 \leq t \leq T$, where h^t (resp. w^t) is the hypothesis (resp. weight vector) computed in round t of Algorithm 1.

Proof. See Section C.1. □

Next, assuming that (29) holds, we can resort to standard analysis for the Hedge algorithm to demonstrate the quality of the final output h^{final} .

Lemma 2. *Suppose that lines 6-11 in Algorithm 1 are replaced with some oracle that returns a hypothesis h^t satisfying $L(h^t, w^t) \leq \min_{h \in \mathcal{H}} L(h, w^t) + \varepsilon_1$ in the t -th round for each $1 \leq t \leq T$. With probability exceeding $1 - \delta/4$, the hypothesis h^{final} output by Algorithm 1 is ε -optimal in the sense that*

$$\max_{1 \leq i \leq k} L(h^{\text{final}}, e_i) \leq \min_{h \in \mathcal{H}} \max_{1 \leq i \leq k} L(h, e_i) + \varepsilon. \quad (30)$$

Proof. See Section C.2. □

Taking Lemma 1 and Lemma 2 together, one can readily see that Algorithm 1 returns an ε -optimal randomized hypothesis h^{final} with probability at least $1 - \delta/2$. The next step then lies in bounding the total number of samples that has been collected Algorithm 1. Towards this end, recall that $\bar{w}_i^T = \max_{1 \leq t \leq T} w_i^t$ for each $i \in [k]$. Recognizing that $\hat{w}_i^t \leq \bar{w}_i^t$ for each $t \in [T]$ and $i \in [k]$, we can bound the total sample size by

$$\begin{aligned} \text{(sample size)} \quad T_1 \sum_{i=1}^k \hat{w}_i^T + k + T \left(k \sum_{i=1}^k \bar{w}_i^T + k \right) &\leq (T_1 \|\bar{w}^T\|_1 + kT \|\bar{w}^T\|_1) + k(T+1) \\ &\leq \frac{d \log\left(\frac{d}{\varepsilon}\right) + k \log\left(\frac{k}{\delta \varepsilon}\right)}{\varepsilon^2} \cdot \|\bar{w}^T\|_1. \end{aligned} \quad (31)$$

Consequently, everything then comes down to bounding $\|\bar{w}^T\|_1$, for which we resort to the following lemma.

Lemma 3. Assume Line 6-11 in Algorithm 1 is replaced by some oracle which returns a hypothesis h^t satisfies that $L(h^t, w^t) \leq \min_{h \in \mathcal{H}} L(h, w^t) + \varepsilon_1$ in the t -th round for each $1 \leq t \leq T$. With probability at least $1 - \delta/4$, one has $\|\bar{w}^T\|_1$ is bounded by

$$\|\bar{w}^T\|_1 \leq O\left(\log^5(k) \log\left(\frac{1}{\varepsilon}\right) \log^2\left(\frac{k}{\delta\varepsilon}\right)\right).$$

It is noteworthy that the proof of Lemma 3 is the most technically challenging part of the analysis; we postpone this proof to Appendix C.3.

Combining Lemma 3 with (31) immediately reveals that, with probability at least $1 - \delta$, the sample complexity of Algorithm 1 is bounded by

$$O\left(\frac{d \log\left(\frac{d}{\varepsilon}\right) + k \log\left(\frac{k}{\delta\varepsilon}\right)}{\varepsilon^2} \cdot \left(\log^5(k) \log\left(\frac{1}{\varepsilon}\right) \log^2\left(\frac{k}{\delta\varepsilon}\right)\right)\right),$$

as claimed in Theorem 1. It remains to prove the above key lemmas, which we postpone to Section C.

6 Necessity of randomization

Given that the best-known sample complexities prior to our work were derived for algorithms that either output randomized hypotheses or invoke majority votes, Awasthi et al. (2023) raised the question about how the sample complexity is impacted if only deterministic (or “proper”) hypotheses are permitted as the output of the learning algorithms. As it turns out, the restriction to deterministic hypotheses substantially worsens the sample efficiency, as revealed by the following theorem.

Theorem 2. Assume that $d \geq 2 \log(8k)$. Consider any $\varepsilon \in (0, 1/100)$, and let $N_0 = \frac{2^d - 1}{k}$. One can find

- a hypothesis class \mathcal{H} containing at most $kN_0 + 1$ hypothesis,
- a collection of k distributions $\mathcal{D} := \{\mathcal{D}_i\}_{i=1}^k$,
- a loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \rightarrow [-1, 1]$,

such that it takes at least $\frac{dk}{240000\varepsilon^2}$ samples to find $h \in \mathcal{H}$ obeying

$$\max_{1 \leq i \leq k} \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\ell(h, (x, y))] \leq \min_{h' \in \mathcal{H}} \max_{i \in [k]} \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\ell(h', (x, y))] + \varepsilon \quad (32)$$

with probability exceeding $3/4$.

Let us briefly describe the high-level strategy for our construction of the hard instance: for each $i \in [k]$, we build a hypothesis set \mathcal{H}_i that performs poorly solely on \mathcal{D}_i . To discriminate the optimal hypothesis — denoted by h^* — from some \mathcal{H}_i , the learner has to call $\Omega(\log(|\mathcal{H}_i|)/\varepsilon^2) = \tilde{\Omega}(d/\varepsilon^2)$ times to $\text{Query}(\mathcal{D}_i)$ in expectation. The result follows by taking sum over $i \in [k]$.

Proof of Theorem 2. Note that $N_0 = \frac{2^d - 1}{k}$. Set $N = kN_0 + 1 = 2^d$. Set $\mathcal{X} = \{-1, 0, 1\}^{kN}$. We set $\mathcal{Y} = \{1\}$ to be a set with only one element. Without loss of generality, we write $\ell(h, (x, y)) = \ell(h, x)$.

We now describe our construction. There are N hypotheses in \mathcal{H} , where each hypothesis corresponds to k dimensions of the ground set \mathcal{X} . Without loss of generality, for $h \in \mathcal{H}$, we let $\mathcal{I}_h = \{j_{h,i}\}_{i=1}^k$ be the k dimensions related to h . Note that $\mathcal{I}_h \cap \mathcal{I}_{h'} = \emptyset$ for $h \neq h'$. We construct \mathcal{H} as $\mathcal{H} = (\cup_{i=1}^k \mathcal{H}_i) \cup \{h^*\}$. Then we define $h(x)$ and $\ell(h, x)$ as $h(x) = \ell(h, x) = x_{i'}$ where $i' = \arg \min_{i \in \mathcal{I}_h, x_i \neq 0}$. Now we design the k distributions $\{\mathcal{D}_i\}_{i=1}^k$. Fix $i \in [k]$, we let $\mathbb{P}_{\mathcal{D}_i}[x_\ell] = \prod_{\ell=1}^{kN} \mathbb{P}_{\mathcal{D}_{i,\ell}}[x_\ell]$, where $\mathbb{P}_{\mathcal{D}_{i,\ell}}[x_\ell] = \mathbb{I}[x_\ell = 0]$ for $\ell \notin \{j_{h,i} | h \in \mathcal{H}\}$, $\mathbb{P}_{\mathcal{D}_{i,\ell}}[x_\ell] = \frac{1}{2} \mathbb{I}[x_\ell = 1] + \frac{1}{2} \mathbb{I}[x_\ell = -1]$ for $\ell \in \{j_{h,i} | h \notin \mathcal{H}_i\}$, and $\mathbb{P}_{\mathcal{D}_{i,\ell}}[x_\ell] = (\frac{1}{2} + 4\varepsilon) \mathbb{I}[x_\ell = 1] + (\frac{1}{2} - 4\varepsilon) \mathbb{I}[x_\ell = -1]$ for $\ell \in \{j_{h,i} | h \in \mathcal{H}_i\}$. In words, for each $i \in [k]$, we draw $x \sim \mathcal{D}_i$ by independently generate each coordinate of x , and there are $|\mathcal{H}| = N$ non-zero coordinates. Furthermore, for $x \sim \mathcal{D}_i$ and $h \in \mathcal{H}$, we have $\ell(h, x) = x_{j_{h,i}}$. Through this construction, we have the following properties:

- (i). $\ell(h, x) \in [-1, 1]$ for any $h \in \mathcal{H}$ and $x \in \mathcal{X}$;
- (ii). $\mathbb{E}_{x \sim \mathcal{D}_i}[\ell(h, x)] = \mathbb{E}_{x \sim \mathcal{D}_{i, j_{h,i}}} [x_{j_{h,i}}] = 8\varepsilon \cdot \mathbb{I}[h \in \mathcal{H}_i]$ for any $i \in [k]$ and $h \in \mathcal{H}$;
- (iii). the only ε -optimal hypothesis is h^* because for any $h \neq h^*$, there exists some i such that $h \in \mathcal{H}_i$;
- (iv). $h(x) \in \{-1, 1\}$ for $x \in (\cup_{i=1}^k \text{supp}(\mathcal{D}_i))^9$, and $|\mathcal{H}| = N = kN_0 + 1 = 2^d$, which imply $\text{VC}(\mathcal{H}) \leq \log_2(N) \leq d$ over $(\cup_{i=1}^k \text{supp}(\mathcal{D}_i))$;
- (v). $\ell(h, x)$ could be regarded as a function of $h(x)$ because $\ell(h, x) = h(x)$.

For each call to $\text{Query}(\mathcal{D}_i)$, we can get independent observations $\{x_{j_{h,i}}\}_{h \in \mathcal{H}}$ where $x_{j_{h,i}} \sim D_{i, j_{h,i}}$ for each $h \in \mathcal{H}$. Now we let the number of calls to $\text{Query}(\mathcal{D}_i)$ be M_i for $i \in [k]$. Our target is to show to distinguish h^* from \mathcal{H}_i , M_i has to be at least $\Omega(d/\varepsilon^2)$.

Suppose now that there is an algorithm \mathcal{G} with numbers of samples $\{M_i\}_{i=1}^k$ such that the output is h^* with probability at least $\frac{3}{4}$. Let $\mathbb{P}_{\mathcal{G}}[\cdot]$ and $\mathbb{E}_{\mathcal{G}}[\cdot]$ denote respectively the probability and expectation under running the algorithm \mathcal{G} . Let h_{out} be the output hypothesis. It then follows that

$$\mathbb{P}_{\mathcal{G}}[h_{\text{out}} = h^*] \geq \frac{3}{4}.$$

Let $\Pi_{\mathcal{H}}$ be the set of permutations over \mathcal{H} . Let $\text{U}(\Pi_{\mathcal{H}})$ be the uniform distribution over $\Pi(\mathcal{H})$. With a slight abuse of notations, for $x \in \{-1, 0, 1\}^{kN}$ and $\sigma \in \Pi_{\mathcal{H}}$, we define $\sigma(x)$ to be the vector y such that $y_{j_{h,i}} = x_{j_{\sigma(h),i}}$ for all $h \in \mathcal{H}$ and $i \in [k]$. Let \mathcal{G}' be the algorithm with \mathcal{H} replaced by $\sigma(\mathcal{H})$ in the input where $\sigma \sim \text{U}(\Pi_{\mathcal{H}})$. Recognizing \mathcal{G} returns the optimal hypothesis with probability at least $3/4$ for all problem instances, we can see that

$$\mathbb{P}_{\mathcal{G}'}[h_{\text{out}} = h^*] \geq \frac{3}{4}.$$

We note that \mathcal{G}' is a symmetric algorithm with respect to the hypothesis set. Formally, we have the lemma below to bound the probability of returning a sub-optimal hypothesis.

Lemma 4. Fix $m \geq 0$ and $\tilde{i} \in [k]$. Suppose $\mathbb{P}_{\mathcal{G}'}[h_{\text{out}} = h^*, M_{\tilde{i}} \leq m] \geq \frac{1}{2}$. It then holds that for any $h \in \mathcal{H}_{\tilde{i}}$

$$\mathbb{P}_{\mathcal{G}'}[h_{\text{out}} = h, M_{\tilde{i}} \leq m] \geq \frac{1}{2} \mathbb{P}_{\mathcal{G}'}[h_{\text{out}} = h^*, M_{\tilde{i}} \leq m] \exp(-80\sqrt{m\varepsilon} - 40m\varepsilon^2).$$

Moreover, it holds that $m \geq \frac{\log(N_0/4)}{30000\varepsilon^2}$.

By Lemma 4, we have

$$\mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h^*, M_{\tilde{i}} \leq \frac{\log(N_0/4)}{30000\varepsilon^2} \right] \leq \frac{1}{2}. \quad (33)$$

Observing that $\mathbb{P}_{\mathcal{G}'}[h_{\text{out}} = h^*] \geq \frac{3}{4}$, we learn that

$$\mathbb{P}_{\mathcal{G}'}[h_{\text{out}} = h^*, M_{\tilde{i}} \geq \frac{\log(N_0/4)}{30000\varepsilon^2}] \geq \frac{1}{4}, \quad (34)$$

which implies that $\mathbb{E}_{\mathcal{G}'}[M_{\tilde{i}}] \geq \frac{\log(N_0/4)}{120000\varepsilon^2} \geq \frac{d - \log_2(8k)}{120000\varepsilon^2} \geq \frac{d}{240000\varepsilon^2}$. Summing over $i \in [k]$ gives

$$\mathbb{E}_{\mathcal{G}'} \left[\sum_{i=1}^k M_{\tilde{i}} \right] \geq \frac{dk}{240000\varepsilon^2}. \quad (35)$$

The proof is thus completed. \square

Algorithm 3: Hedge for Multi-distribution Learning on Rademacher Class (MDL – Hedge – Rademacher)

1 **input:** k data distributions $\{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_k\}$, hypothesis class \mathcal{H} , target accuracy level ε , target success rate $1 - \delta$, the constant $\{C_n\}_{n \geq 1}$ in Assumption 1.

2 **hyper-parameters:** stepsize $\eta = \frac{1}{100}\varepsilon$, number of rounds $T = \frac{20000 \log(\frac{k}{\delta})}{\varepsilon^2}$, auxiliary accuracy level $\varepsilon_1 = \frac{1}{100}\varepsilon$, auxiliary sub-sample-size $T_1 = \min \left\{ t \geq \frac{400(k \log(\frac{k}{\varepsilon_1}) + \log(\frac{1}{\delta}))}{\varepsilon_1^2} \mid C_t \leq \frac{\varepsilon_1}{1200} \right\}$.

3 **initialization:** for all $i \in [k]$, set $W_i^1 = 1$, $\hat{w}_i^0 = 0$ and $n_i^0 = 0$; $\mathcal{S} = \emptyset$.

4 draw $\lceil 12 \log(2k) \rceil$ samples from \mathcal{D}_i for each i , and add these samples to \mathcal{S} .

5 **for** $t = 1, 2, \dots, T$ **do**

6 set $w^t = [w_i^t]_{1 \leq i \leq k}$ and $\hat{w}^t = [\hat{w}_i^t]_{1 \leq i \leq k}$, where $w_i^t \leftarrow \frac{W_i^t}{\sum_j W_j^t}$ and $\hat{w}_i^t \leftarrow \hat{w}_i^{t-1}$ for all $i \in [k]$.

7 /* recompute \hat{w}^t & draw new samples for \mathcal{S}_w only if w^t changes sufficiently. */

8 **if** there exists $j \in [k]$ such that $w_j^t \geq 2\hat{w}_j^{t-1}$ **then**

9 $\hat{w}_i^t \leftarrow \max\{w_i^t, \hat{w}_i^{t-1}\}$ for all $i \in [k]$;

10 **for** $i = 1, \dots, k$ **do**

11 $n_i^t \leftarrow \lceil T_1 \hat{w}_i^t \rceil$;

12 draw $n_i^t - n_i^{t-1}$ independent samples from \mathcal{D}_i , and add these samples to \mathcal{S} .

13 /* estimate the near-optimal hypothesis for weighted data distributions. */

14 compute $h^t \leftarrow \arg \min_{h \in \mathcal{H}} \hat{L}(h, w^t)$, where

$$\hat{L}^t(h, w^t) := \sum_{i=1}^k \frac{w_i^t}{\check{n}_i^t} \cdot \sum_{j=1}^{\check{n}_i^t} \ell(h, (x_{i,j}, y_{i,j})) \quad (36)$$

15 with $\check{n}_i^t = \min\{\lceil T_1 w_i^t + 12 \log(2k) \rceil, T_1\}$, and $(x_{i,j}, y_{i,j})$ being the j -th datapoint from \mathcal{D}_i in \mathcal{S} .

16 /* estimate the loss vector and execute weighted updates. */

17 $\bar{w}_i^t \leftarrow \max_{1 \leq \tau \leq t} w_i^\tau$ for all $i \in [k]$.

18 **for** $i = 1, \dots, k$ **do**

19 draw $\lceil k \bar{w}_i^t \rceil$ independent samples — denoted by $\{(x_{i,j}^t, y_{i,j}^t)\}_{j=1}^{\lceil k \bar{w}_i^t \rceil}$ — from \mathcal{D}_i , and set

$$\hat{r}_i^t = \frac{1}{\lceil k \bar{w}_i^t \rceil} \sum_{j=1}^{\lceil k \bar{w}_i^t \rceil} \ell(h^t, (x_{i,j}^t, y_{i,j}^t));$$

20 update the weight as $W_i^{t+1} = W_i^t \exp(\eta \hat{r}_i^t)$. // Hedge updates.

21 **output:** a randomized hypothesis h^{final} as a uniform distribution over $\{h^t\}_{t=1}^T$.

7 Extension: learning Rademacher classes

In this section, we study how to adapt our algorithm and theory to accommodate MDL for Rademacher classes.

7.1 Preliminaries: Rademacher complexity

Let us first introduce the formal definition of the Rademacher complexity; more detailed introduction can be found in [Shalev-Shwartz and Ben-David \(2014\)](#).

Definition 3 (Rademacher complexity). *Given a distribution \mathcal{D} supported on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ and a positive integer n , the (average) Rademacher complexity is defined as*

$$\text{Rad}_n(\mathcal{D}) := \mathbb{E}_{\{z_i\}_{i=1}^n} \left[\mathbb{E}_{\{\sigma_i\}_{i=1}^n} \left[\max_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(h, z_i) \right] \right], \quad (37)$$

where $\{z_i\}_{i=1}^n$ are drawn independently from \mathcal{D} , and $\{\sigma_i\}_{i=1}^n$ are i.i.d. Rademacher random variables obeying $\mathbb{P}\{\sigma_i = 1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$ for each $1 \leq i \leq n$.

Next, we would like to make an assumption on the Rademacher complexity of mixtures of distributions. Denoting by $\mathcal{D}(w)$ the mixed distribution

$$\mathcal{D}(w) := \sum_{i=1}^k w_i \mathcal{D}_i \quad (38)$$

for any probability vector $w \in \Delta(k)$, we can state our assumption as follows.

Assumption 1. *For each $n \geq 1$, there exists a universal constant $C_n > 0$ such that*

$$\text{Rad}_n(\mathcal{D}(w)) \leq C_n \quad (39)$$

holds for all $w \in \Delta^k$.

Remark 1. *One might raise a natural question about Assumption 1: can we only assume $\text{Rad}_n(\mathcal{D}_i) \leq C_n$ for $i \in [k]$ without incurring a worse sample complexity? The answer is, however, negative. In fact, the Rademacher complexity $\text{Rad}_n(\mathcal{D}(w))$ is not convex in w , and hence we fail to use $\max_i \text{Rad}_n(\mathcal{D}_i)$ to bound $\max_{w \in \Delta(k)} \text{Rad}_n(\mathcal{D}(w))$. The interested reader is referred to Appendix F.3 for more details.*

It is well known that $\text{VC-dim}(\mathcal{H}) \leq d$ implies Assumption 1 holds with $C_n = \sqrt{\frac{2d \log(en/d)}{n}}$ ([Mohri et al., 2018](#)).

To facilitate our analysis in this section, we find it helpful to introduce the notion of the weighted Rademacher complexity as follows.

Definition 4 (Weighted Rademacher complexity). *Given a collection of distributions $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^k$ and a set of positive integers $\{n_i\}_{i=1}^k$, the weighted (average) Rademacher complexity is defined as*

$$\widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k}(\mathcal{D}) := \mathbb{E}_{\{z_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\mathbb{E}_{\{\sigma_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\frac{1}{\sum_{i=1}^k n_i} \max_{h \in \mathcal{H}} \sum_{i=1}^k \sum_{j=1}^{n_i} \sigma_i^j \ell(h, z_i^j) \right] \right], \quad (40)$$

where $\{\{z_i^j\}_{j=1}^{n_i}\}_{i=1}^k$ are independently generated with each z_i^j drawn from \mathcal{D}_i , and $\{\{\sigma_i^j\}_{j=1}^{n_i}\}_{i=1}^k$ are independent Rademacher random variables obeying $\mathbb{P}\{\sigma_i^j = 1\} = \mathbb{P}\{\sigma_i^j = -1\} = 1/2$. Throughout the rest of this paper, we shall often abbreviate $\widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} = \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k}(\mathcal{D})$.

⁹We use $\text{supp}(\mathcal{D})$ to denote the support of the distribution \mathcal{D}

The weighted Rademacher complexity defined above satisfies an important property below.

Lemma 5. *For any two groups of positive integers $\{n_i\}_{i=1}^k$ and $\{m_i\}_{i=1}^k$, it holds that*

$$\begin{aligned} \left(\sum_{i=1}^k n_i \right) \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} &\leq \left(\sum_{i=1}^k (m_i + n_i) \right) \widetilde{\text{Rad}}_{\{m_i + n_i\}_{i=1}^k} \\ &\leq \left(\sum_{i=1}^k n_i \right) \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} + \left(\sum_{i=1}^k m_i \right) \widetilde{\text{Rad}}_{\{m_i\}_{i=1}^k}. \end{aligned} \quad (41)$$

In addition, the following lemma allows us to bound the weighted Rademacher complexity under Assumption 1.

Lemma 6. *Consider any $\{n_i\}_{i=1}^k$ obeying $n_i \geq 12 \log(2k)$ for each $i \in [k]$. By taking $w \in \Delta^k$ with $w_i = \frac{n_i}{\sum_{l=1}^k n_l}$, one has*

$$\widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} \leq 72 \text{Rad}_{\sum_{i=1}^k n_i}(\mathcal{D}(w)).$$

7.2 Algorithm and sample complexity

Let us present our algorithm in Algorithm 3, which seeks to learn a Rademacher class in the presence of multiple distributions. Note that Algorithm 3 is also a Hedge-like algorithm to learn a convex (concave) game. Its major difference from Algorithm 1 lies in the subroutine to learn h^t (see lines 6-12 in Algorithm 3). More precisely, to compute the estimator $\hat{L}^t(h, w^t)$ for $L(h, w^t)$, instead of using the first n_i^t samples from \mathcal{D}_i for each $i \in [k]$, we choose to use the first

$$\check{n}_i^t = \min \{ \lceil T_1 w_i^t + 12 \log(2k) \rceil, T_1 \}$$

samples from \mathcal{D}_i for each i . Formally, we have the following theoretical guarantees.

Theorem 3. *Suppose Assumption 1 holds. With probability at least $1 - \delta$, the output h^{final} returned by Algorithm 3 satisfies*

$$\max_i L(h^{\text{final}}, e_i) = \max_i \frac{1}{T} \sum_{t=1}^T L(h^t, e_i) \leq \min_{h \in \mathcal{H}} \max_i L(h, e_i) + \varepsilon.$$

Meanwhile, the sample complexity of Algorithm 3 is bounded by

$$O \left(T_\epsilon \cdot \log^5(k) \log \left(\frac{1}{\varepsilon} \right) \log^2 \left(\frac{k}{\delta \varepsilon} \right) \right),$$

where T_ϵ is defined as

$$T_\epsilon := \min \left\{ t \geq \frac{400 \left(k \log \left(\frac{k}{\varepsilon_1} \right) + \log \left(\frac{1}{\delta} \right) \right)}{\varepsilon_1^2} \mid C_t \leq \frac{\varepsilon_1}{1200} \right\}.$$

In the case where $\text{VC-dim}(\mathcal{H}) \leq d$, we have $C_n \leq \sqrt{\frac{2d \log(en/d)}{n}}$, which implies that $T_\epsilon = \tilde{O} \left(\frac{d+k}{\varepsilon^2} \right)$ and a sample complexity bound of $\tilde{O} \left(\frac{d+k}{\varepsilon^2} \right)$.

Proof of Theorem 3. In view of Lemma 2 and Lemma ??, it suffices to show that running Algorithm 3 results in $L(h^t, w^t) \leq \min_{h \in \mathcal{H}} L(h, w^t) + \varepsilon_1$ for any $1 \leq t \leq T$, a property that holds with probability at least $1 - \delta/4$. Formally, we have the lemma below.

Lemma 7. *Suppose Assumption 1 holds. With probability at least $1 - \delta/4$, the iterates of Algorithm 3 satisfy*

$$L(h^t, w^t) \leq \min_{h \in \mathcal{H}} L(h, w^t) + \varepsilon_1 \quad (42)$$

for any $1 \leq t \leq T$.

The proof of Lemma 7 is postponed to Appendix F. □

8 Extension: oracle-efficient multi-group learning

Given a hypothesis set \mathcal{H} where each $h \in \mathcal{H} : \mathcal{X} \rightarrow \mathcal{Y}$, a set of groups \mathcal{G} such that each $g \in \mathcal{G}$ is a subset of \mathcal{X} , and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$ and a distribution \mathcal{D} with support $\mathcal{X} \times \mathcal{Y}$, define $L_{\mathcal{D}}(h \mid g) := \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(h(x), y) \mid x \in g]$. Define $P_g = \mathbb{P}_{(x,y) \sim \mathcal{D}}[x \in g]$. Let $\gamma = \min_{g \in \mathcal{G}} P_g$. The goal of the learning algorithm is to find a (possibly randomized) hypothesis h to minimize $L_{\mathcal{D}}(h \mid g)$ for all $g \in \mathcal{G}$ up to some threshold $\varepsilon \in (0, 1]$. That is,

$$L_{\mathcal{D}}(h \mid g) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h' \mid g) + \varepsilon, \forall g \in \mathcal{G}.$$

We continue with the compatibility assumption as below, which ensures the existence of such a near-optimal hypothesis h .

Assumption 2. *There exists $h^* \in \mathcal{H}$, such that*

$$L_{\mathcal{D}}(h^* \mid g) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h \mid g) + \frac{\varepsilon}{8}, \quad \forall g \in \mathcal{G}.$$

The learning algorithm As presented in Algorithm 4, we first sample $N = O\left(\frac{\log(|\mathcal{G}|/\delta) + d \log(d/\varepsilon)}{\gamma \varepsilon^2}\right)$ datapoints from \mathcal{D} , and then estimate the optimal value f_g for each group $g \in \mathcal{G}$. By taking $L_{\mathcal{D}}(h \mid g) - f_g$ to be the loss function, we can then invoke a Hedge algorithm over \mathcal{G} to solve the following problem:

$$\min_{h \in \Delta(\mathcal{H})} \max_{g \in \mathcal{G}} (L_{\mathcal{D}}(h \mid g) - f_g). \quad (43)$$

According to Assumption 2, it suffices to find some (possibly randomized) h such that

$$\max_{g \in \mathcal{G}} (L_{\mathcal{D}}(h \mid g) - f_g) \leq \max_{g \in \mathcal{G}} (L_{\mathcal{D}}(h^* \mid g) - f_g) + O(\varepsilon)$$

Formally, we have the following theorem.

Theorem 4. *Assume $\varepsilon \in (0, 1/10]$. Suppose Assumption 2 holds. By running Algorithm 4, with probability at least $1 - \delta$, the output h^{final} satisfies that*

$$L_{\mathcal{D}}(h^{\text{final}} \mid g) = \frac{1}{T} \sum_{t=1}^T L_{\mathcal{D}}(h^t \mid g) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h \mid g) + \varepsilon \quad (44)$$

for any $g \in \mathcal{G}$. Meanwhile, the sample complexity of Algorithm 4 is bounded by

$$O\left(\frac{d \log(d/\varepsilon) + \log(|\mathcal{G}|/\delta)}{\gamma \varepsilon^2}\right).$$

Moreover, Algorithm 4 only access \mathcal{H} with an ERM oracle.

Under Assumption 2, Theorem 4 recovers the result in Tosh and Hsu (2022) without direct access to the hypothesis set \mathcal{H} . In high-level idea, the proof of Theorem 4 is based on the uniform convergence argument, which is a natural extension of our main results.

9 Discussion

In this paper, we have settled the problem of achieving optimal sample complexity in multi-distribution learning, assuming the availability of adaptive (or on-demand) sampling. We have put forward a novel oracle-efficient algorithm that provably attains a sample complexity of $\tilde{O}\left(\frac{d+k}{\varepsilon^2}\right)$ for VC classes, which matches the

best-known lower bound up to some logarithmic factor. From the technical perspective, the key novelty of our analysis lies in carefully bounding the trajectory of the Hedge algorithm on a convex (concave) optimization problem. We have further unveiled the necessity of randomization, revealing that a considerably larger sample size is necessary if the learning algorithm is constrained to return deterministic hypotheses. Notably, our work manages to solve three open problems presented in COLT 2023 (namely, [Awasthi et al. \(2023, Problems 1, 3 and 4\)](#)).

Our work not only addresses existing challenges but also opens up several directions for future exploration. To begin with, while our sample complexity results are optimal up to logarithmic factors, further studies are needed in order to sharpen the logarithmic dependency. Additionally, the current paper assumes a flexible sampling protocol that allows the learner to take samples arbitrarily from any of the k distributions; how will the sample complexity be impacted under additional constraints imposed on the sampling process? Furthermore, can we extend our current analysis (which bounds the dynamics of the Hedge algorithm) to control the trajectory of more general first-order/second-order algorithms, in the context of robust online learning? Another venue for exploration is the extension of our multi-distribution learning framework to tackle other related tasks like multi-calibration ([Hébert-Johnson et al., 2018](#); [Haghtalab et al., 2023](#)). We believe that our algorithmic and analysis framework can shed light on making progress in all of these directions.

Acknowledgements

We thank Eric Zhao for answering numerous questions about the open problems. YC is supported in part by the Alfred P. Sloan Research Fellowship, the NSF grants CCF-1907661, DMS-2014279, IIS-2218713 and IIS-2218773. JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, NSF CIF 2212262, ONR Young Investigator Award, and NSF CAREER Award 2144994.

A Additional figures

In this section, we provide several examples with figures, in order to help the readers understand our strategy in obtaining a regular configuration in Section 4.2.2 and Lemma 16. Let us provide a brief introduction to these figures.

In Figure 1, we present an example of a regular configuration. In this example, letting $[\tilde{s}_i, \tilde{e}_i]$ be the interval for the i -th block, we have $\tilde{e}_i - \tilde{s}_i \geq \Omega(m_i p / \varepsilon^2)$, where m_i is the number of coordinates in the i -th block. By observing that

$$\sum_i \frac{m_i p}{\varepsilon^2} = O(\sum_i \tilde{e}_i - \tilde{s}_i) = O(T),$$

we can derive

$$|\mathcal{W}(p)| = \sum_i m_i = O(T \varepsilon^2 / p) = \tilde{O}(1/p).$$

In addition, we give an example of irregular configuration in Figure 2. Due to the non-disjoint segments, one cannot perform the arguments above to bound $|\mathcal{W}(p)|$.

In Figure 3, we provide an example of the partition of blocks, and in Figure 4, we illustrate how to align one side of the segments using a common inner point.

In Figure 5 and Figure 6, we illustrate how to construct the regular configurations using a group of segments with the same starting points in the case where $k = 8$. In this toy example, we have in total 5 configurations with 5 different colors. For each $\left(\frac{p}{4}, \frac{p}{8}, \frac{\log(2)}{2}\right)$ -segment in Figure 5, it forms an $\left(\frac{p}{4}, \frac{p}{8}, \frac{\log(2)}{2(\log_2(k)+2)}\right)$ -segment to at least one of these configurations. According to the pigeonhole principle, there must be at least one regular configuration with a number $\frac{|\mathcal{W}(p)|}{6(\log_2(T)+1)(\log_2(k)+2)}$ of $\left(\frac{p}{4}, \frac{p}{8}, \frac{\log(2)}{2(\log_2(k)+2)}\right)$ -segments (see Lemma 16).

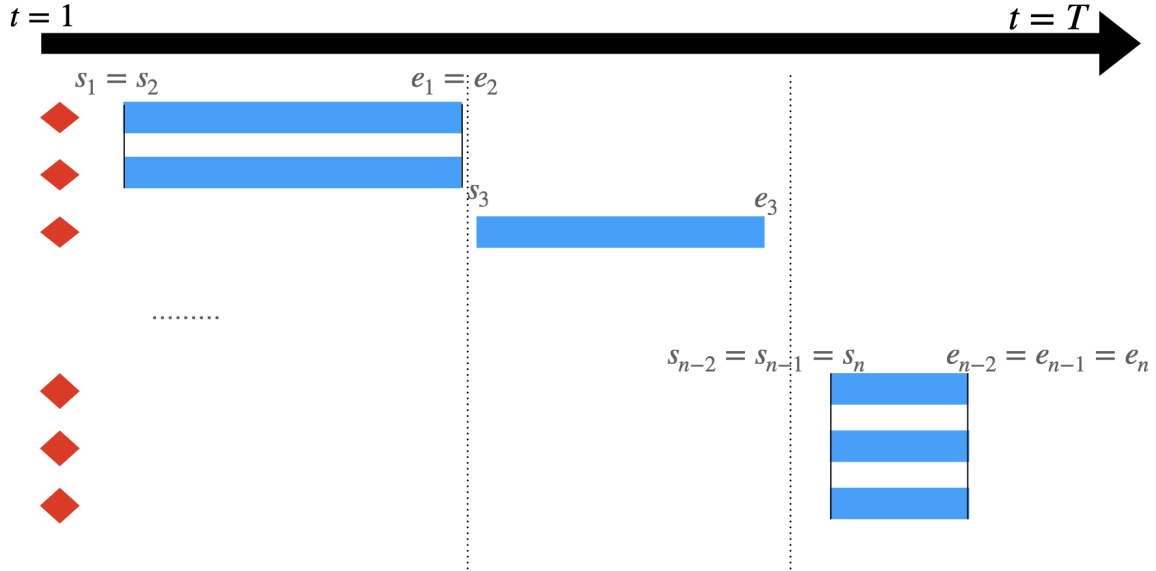


Figure 1: Regular configuration.

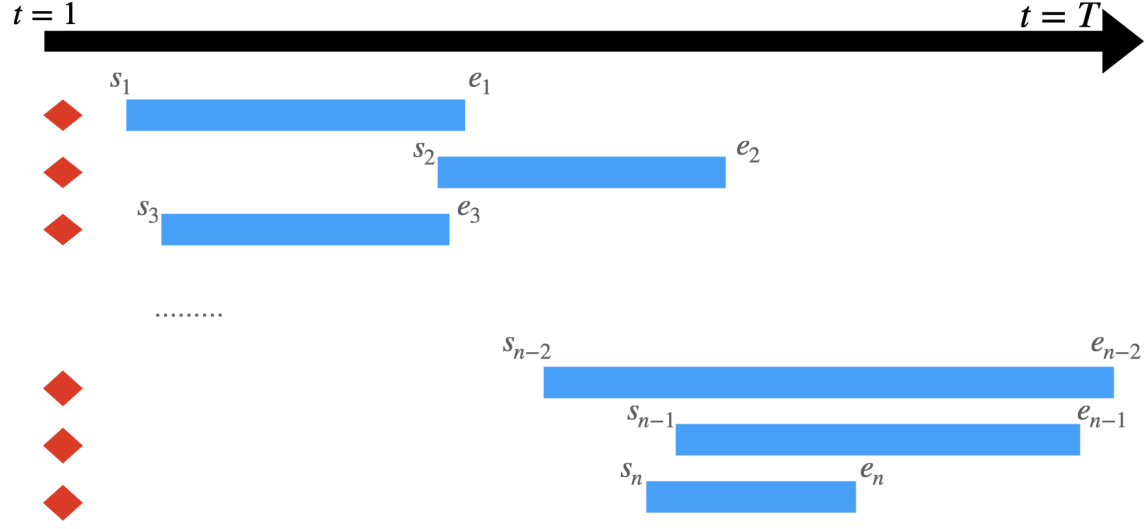


Figure 2: General irregular configuration.

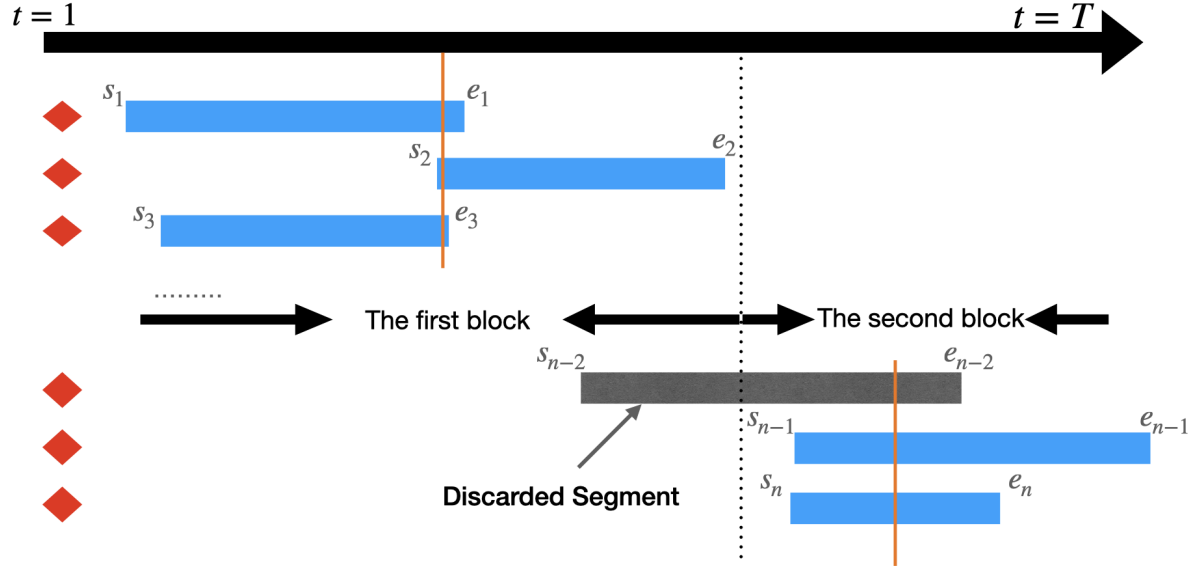


Figure 3: Partition of blocks.

B Auxiliary lemmas

In this section, we introduce several technical lemmas that are used multiple times in our analysis.

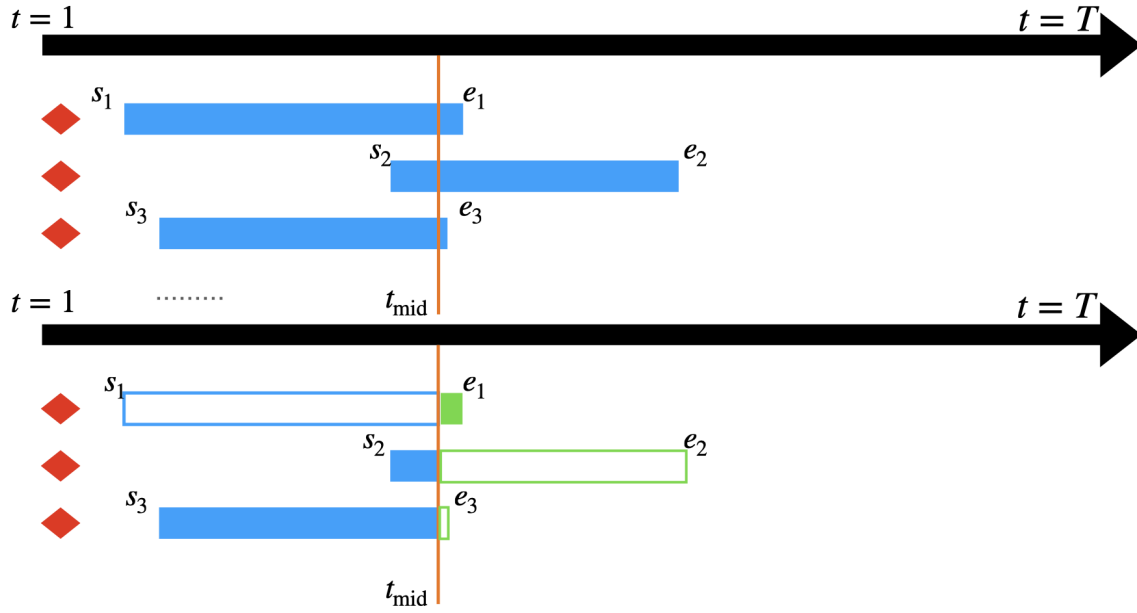


Figure 4: Aligning one side of the configuration. The unfilled part of the segments means that the variation over $w_i^{s_i} \rightarrow w_i^{t_{\text{mid}}}$ (i.e., $\log(w_i^{t_{\text{mid}}}/w_i^{s_i})$) is not significant enough.

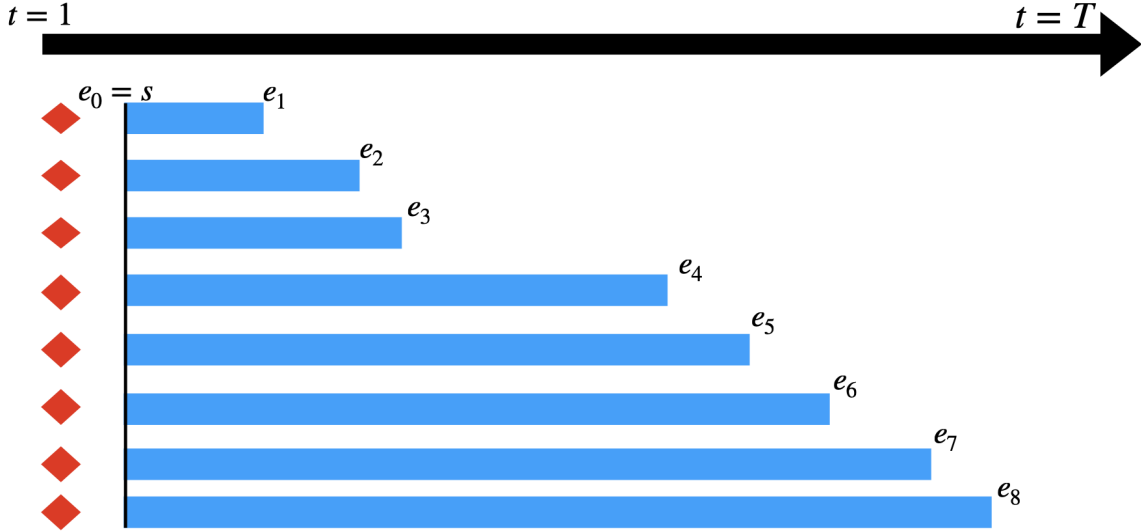


Figure 5: A group of segments with common starting points.

We begin by introducing three handy concentrations inequalities. The first result is the well-renowned Freedman inequality ([Freedman, 1975](#)), which assists in deriving variance-aware concentration inequalities for

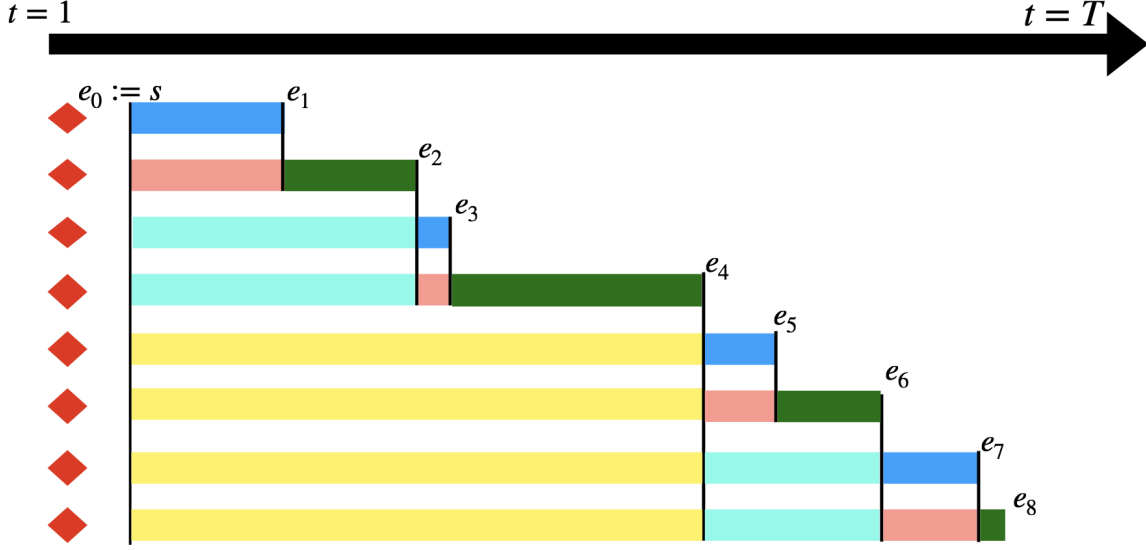


Figure 6: Construction of the regular configurations. Each segment is cut into at most $\log_2(k) + 1$ sub-segments with different colors.

martingales.

Lemma 8 (Freedman's inequality ([Freedman, 1975](#))). *Let $(M_n)_{n \geq 0}$ be a martingale obeying $M_0 = 0$. Define $V_n := \sum_{k=1}^n \mathbb{E}[(M_k - M_{k-1})^2 | \mathcal{F}_{k-1}]$ for each $n \geq 0$, where \mathcal{F}_k denotes the σ -algebra generated by (M_1, M_2, \dots, M_k) . Suppose that $M_k - M_{k-1} \leq 1$ for all $k \geq 1$. Then for any $x > 0$ and $y > 0$, one has*

$$\mathbb{P}(M_n \geq nx, V_n \leq ny) \leq \exp\left(-\frac{nx^2}{2(y + \frac{1}{3}x)}\right). \quad (45)$$

The second concentration result bounds the difference between the sum of a sequence of random variables and the sum of their respective conditional means (w.r.t. the associated σ -algebra).

Lemma 9 (Lemma 10 in [Zhang et al. \(2022\)](#)). *Let X_1, X_2, \dots be a sequence of random variables taking value in the interval $[0, l]$. For any $k \geq 1$, let \mathcal{F}_k be the σ -algebra generated by (X_1, X_2, \dots, X_k) , and define $Y_k := \mathbb{E}[X_k | \mathcal{F}_{k-1}]$. Then for any $\delta > 0$, we have*

$$\begin{aligned} \mathbb{P}\left\{\exists n \in \mathbb{N}, \sum_{k=1}^n X_k \geq 3 \sum_{k=1}^n Y_k + l \log \frac{1}{\delta}\right\} &\leq \delta, \\ \mathbb{P}\left\{\exists n \in \mathbb{N}, \sum_{k=1}^n Y_k \geq 3 \sum_{k=1}^n X_k + l \log \frac{1}{\delta}\right\} &\leq \delta. \end{aligned}$$

The third concentration result is the Mcdiarmid inequality, a celebrated inequality widely used to control the fluctuation of multivariate functions when the input variables are independently generated.

Lemma 10 (Mcdiarmid's inequality). *Let X_1, X_2, \dots, X_n be a sequence of independent random variables, with X_i supported on \mathcal{X}_i . Let $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_n \rightarrow \mathbb{R}$ be a function such that: for any $i \in [n]$ and any $\{x_1, \dots, x_n\} \in \mathcal{X}_1 \times \dots \times \mathcal{X}_n$,*

$$\sup_{x'_i \in \mathcal{X}_i} |f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x'_i, \dots, x_n)| \leq c$$

holds for some quantity $c > 0$. It then holds that

$$\mathbb{P}\left\{|f(X_1, X_2, \dots, X_n) - \mathbb{E}[f(X_1, X_2, \dots, X_n)]| \geq \varepsilon\right\} \leq 2 \exp\left(-\frac{2\varepsilon^2}{nc^2}\right).$$

Additionally, the following lemma presents a sort of the data processing inequality w.r.t. the Kullback-Leibler (KL) divergence, which is a classical result from information theory.

Lemma 11. *Let \mathcal{X} and \mathcal{Y} be two sets, and consider any function $f : \mathcal{X} \rightarrow \mathcal{Y}$. For any two random variables X_1 and X_2 supported on \mathcal{X} , it holds that*

$$\text{KL}(\mu(X_1) \parallel \mu(X_2)) \geq \text{KL}(\mu(f(X_1)) \parallel \mu(f(X_2))), \quad (46)$$

where we use $\mu(Z)$ to denote the distribution of a random variable Z .

Lastly, let us make note of an elementary bound regarding the KL divergence between two Bernoulli distributions.

Lemma 12. *Consider any $q > 0$ and $x \in [0, \log(2)]$. Also, consider any $y, y' \in (0, 1)$ obeying $y \geq q$ and $y' \geq \exp(x)y$. It then holds that*

$$\text{KL}(\text{Ber}(y) \parallel \text{Ber}(y')) \geq \frac{qx^2}{4},$$

where $\text{Ber}(z)$ denotes the Bernoulli distribution with mean z .

Proof. To begin with, the function defined below satisfies

$$f(a, b) := \text{KL}(\text{Ber}(a) \parallel \text{Ber}(b)) = a \log\left(\frac{a}{b}\right) + (1-a) \log\left(\frac{1-a}{1-b}\right).$$

For any $0 < a \leq b \leq 1$, it is readily seen that

$$\frac{\partial f(a, b)}{\partial b} = -\frac{a}{b} + \frac{1-a}{1-b} = \frac{b-a}{b(1-b)} \geq 0.$$

It follows from our assumptions $y \geq q$ and $y' \geq \exp(x)y$ that

$$\begin{aligned} \text{KL}(\text{Ber}(y) \parallel \text{Ber}(y')) &= f(y, y') = f(y, y) + \int_y^{y'} \frac{\partial f(y, z)}{\partial z} dz = \int_y^{y'} \frac{z-y}{z(1-z)} dz \\ &\geq \frac{1}{y'} \int_y^{y'} (z-y) dz \geq \frac{(y'-y)^2}{2y'} \\ &\geq \frac{(y'-y)(1-\exp(-x))}{2} \\ &\geq \frac{y(\exp(x)-1)^2}{4} \geq \frac{qx^2}{4}, \end{aligned}$$

where the penultimate inequality uses $x \in [0, \log(2)]$, and the last inequality holds since $y \geq q$. \square

C Proofs of auxiliary lemmas for VC classes

C.1 Proof of Lemma 1

For ease of presentation, suppose there exists a dataset $\tilde{\mathcal{D}}$ containing T_1 independent samples drawn from each distribution \mathcal{D}_i ($1 \leq i \leq k$), so that in total it contains kT_1 samples. We find it helpful to introduce the following notation.

- For each $i \in [k]$ and $j \in [n_i]$, denote by $(x_{i,j}, y_{i,j})$ the j -th sample in $\tilde{\mathcal{S}}$ that is drawn from \mathcal{D}_i .
- For each set of integers $n = \{n_i\}_{i=1}^k \in \mathbb{N}^k$, we define $\tilde{\mathcal{S}}(n)$ to be the dataset containing $\{(x_{i,j}, y_{i,j})\}_{1 \leq j \leq n_i}$ for all $i \in [k]$; namely, it comprises, for each $i \in [k]$, the first n_i samples in $\tilde{\mathcal{S}}$ that are drawn from \mathcal{D}_i .
- We shall also let $\tilde{\mathcal{S}}^+(n) = \{ \{(x_{i,j}^+, y_{i,j}^+)\}_{j=1}^{n_i} \}_{i=1}^k$ be an *independent copy* of $\tilde{\mathcal{S}}(n)$, where for each $i \in [k]$, $\{(x_{i,j}^+, y_{i,j}^+)\}$ are independent samples drawn from \mathcal{D}_i .

Equipped with the above notation, we are ready to present our proof.

Step 1: concentration bounds for any fixed $n = \{n_i\}_{i=1}^k$ and $w \in \Delta(k)$. Consider first any fixed $n = \{n_i\}_{i=1}^k$ obeying $0 \leq n_i \leq T_1$ for all $i \in [k]$, and any fixed $w \in \Delta(k)$. For any quantity $\lambda \in [0, \min_{i \in [k]} \frac{n_i}{w_i}]$, if we take

$$E(\lambda, n, w) := \mathbb{E}_{\tilde{\mathcal{S}}(n)} \left[\max_{h \in \mathcal{H}} \exp \left(\lambda \left\{ \sum_{i=1}^k w_i \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(h, (x_{i,j}, y_{i,j})) - L(h, w) \right\} \right) \right] \quad (47)$$

with the expectation taken over the randomness of $\tilde{\mathcal{S}}(n)$, then we can apply a standard “symmetrization” trick to bound $E(\lambda, n, w)$ as follows:

$$\begin{aligned} E(\lambda, n, w) &:= \mathbb{E}_{\tilde{\mathcal{S}}(n)} \left[\max_{h \in \mathcal{H}} \exp \left(\lambda \left\{ \sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} \ell(h, (x_{i,j}, y_{i,j})) - L(h, w) \right\} \right) \right] \\ &= \mathbb{E}_{\tilde{\mathcal{S}}(n)} \left[\max_{h \in \mathcal{H}} \exp \left(\lambda \left\{ \sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} \ell(h, (x_{i,j}, y_{i,j})) - \mathbb{E}_{\tilde{\mathcal{S}}^+(n)} \left[\sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} \ell(h, (x_{i,j}^+, y_{i,j}^+)) \right] \right\} \right) \right] \\ &\leq \mathbb{E}_{\tilde{\mathcal{S}}(n)} \left[\max_{h \in \mathcal{H}} \mathbb{E}_{\tilde{\mathcal{S}}^+(n)} \left[\exp \left(\lambda \left\{ \sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} (\ell(h, (x_{i,j}, y_{i,j})) - \ell(h, (x_{i,j}^+, y_{i,j}^+))) \right\} \right) \right] \right] \\ &\leq \mathbb{E}_{\tilde{\mathcal{S}}(n), \tilde{\mathcal{S}}^+(n)} \left[\max_{h \in \mathcal{H}} \exp \left(\lambda \left\{ \sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} (\ell(h, (x_{i,j}, y_{i,j})) - \ell(h, (x_{i,j}^+, y_{i,j}^+))) \right\} \right) \right], \end{aligned} \quad (48)$$

where the last two inequalities follow from Jensen’s inequality.

Next, let $\sigma(n) := \{\{\sigma_{i,j}\}_{j=1}^{n_i}\}_{i=1}^k$ be a collection of i.i.d. Rademacher random variables obeying $\mathbb{P}(\sigma_{i,j} = 1) = \mathbb{P}(\sigma_{i,j} = -1) = 1/2$. Denoting $\mathcal{C} = \{(x_{i,j}, y_{i,j})\} \cup \{(x_{i,j}^+, y_{i,j}^+)\}$, we obtain

$$\begin{aligned} &\mathbb{E}_{\tilde{\mathcal{S}}(n), \tilde{\mathcal{S}}^+(n)} \left[\max_{h \in \mathcal{H}} \exp \left(\lambda \left\{ \sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} (\ell(h, (x_{i,j}, y_{i,j})) - \ell(h, (x_{i,j}^+, y_{i,j}^+))) \right\} \right) \right] \\ &= \mathbb{E}_{\tilde{\mathcal{S}}(n), \tilde{\mathcal{S}}^+(n)} \left[\mathbb{E}_{\sigma(n)} \left[\max_{h \in \mathcal{H}} \exp \left(\lambda \left\{ \sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} \sigma_{i,j} (\ell(h, (x_{i,j}, y_{i,j})) - \ell(h, (x_{i,j}^+, y_{i,j}^+))) \right\} \right) \mid \mathcal{C} \right] \right]. \end{aligned} \quad (49)$$

Note that for any dataset \mathcal{C} with cardinality $|\mathcal{C}|$, the Sauer–Shelah lemma (Wainwright, 2019, Proposition 4.18) together with our assumption that $\text{VC-dim}(\mathcal{H}) \leq d$ tells us that the cardinality of the following set obeys

$$|\mathcal{H}(\mathcal{C})| \leq (|\mathcal{C}| + 1)^d \leq (|\tilde{\mathcal{S}}| + |\tilde{\mathcal{S}}^+| + 1)^d \leq (2kT_1 + 1)^d, \quad (50)$$

where $\mathcal{H}(\mathcal{C})$ denotes the set obtained by applying all $h \in \mathcal{H}$ to the data points in \mathcal{C} , namely,

$$\mathcal{H}(\mathcal{C}) := \left\{ (h(x_{1,1}), h(x_{1,1}^+), h(x_{1,2}), h(x_{1,2}^+), \dots) \mid h \in \mathcal{H} \right\}. \quad (51)$$

We shall also define $\mathcal{H}_{\min, \mathcal{C}} \subseteq \mathcal{H}$ to be the *minimum-cardinality subset* of \mathcal{H} that results in the same outcome as \mathcal{H} when applied to \mathcal{C} , namely,

$$\mathcal{H}_{\min, \mathcal{C}}(\mathcal{C}) = \mathcal{H}(\mathcal{C}) \quad \text{and} \quad |\mathcal{H}_{\min, \mathcal{C}}| = |\mathcal{H}(\mathcal{C})|.$$

With these in place, we can demonstrate that

$$\begin{aligned}
& \mathbb{E}_{\sigma(n)} \left[\max_{h \in \mathcal{H}} \exp \left(\lambda \left\{ \sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} \sigma_{i,j} \left(\ell(h, (x_{i,j}, y_{i,j})) - \ell(h, (x_{i,j}^+, y_{i,j}^+)) \right) \right\} \right) \middle| \mathcal{C} \right] \\
&= \mathbb{E}_{\sigma(n)} \left[\max_{h \in \mathcal{H}_{\min, \mathcal{C}}} \exp \left(\lambda \left\{ \sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} \sigma_{i,j} \left(\ell(h, (x_{i,j}, y_{i,j})) - \ell(h, (x_{i,j}^+, y_{i,j}^+)) \right) \right\} \right) \middle| \mathcal{C} \right] \\
&\leq \mathbb{E}_{\sigma(n)} \left[\sum_{h \in \mathcal{H}_{\min, \mathcal{C}}} \exp \left(\lambda \left\{ \sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} \sigma_{i,j} \left(\ell(h, (x_{i,j}, y_{i,j})) - \ell(h, (x_{i,j}^+, y_{i,j}^+)) \right) \right\} \right) \middle| \mathcal{C} \right] \\
&\leq |\mathcal{H}_{\min, \mathcal{C}}| \max_{h \in \mathcal{H}_{\min, \mathcal{C}}} \mathbb{E}_{\sigma(n)} \left[\exp \left(\lambda \left\{ \sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} \sigma_{i,j} \left(\ell(h, (x_{i,j}, y_{i,j})) - \ell(h, (x_{i,j}^+, y_{i,j}^+)) \right) \right\} \right) \middle| \mathcal{C} \right] \\
&\leq (2kT_1 + 1)^d \max_{h \in \mathcal{H}} \prod_{i=1}^k \prod_{j=1}^{n_i} \mathbb{E}_{\sigma_{i,j}} \left[\exp \left(\lambda \left\{ \frac{w_i}{n_i} \sigma_{i,j} \left(\ell(h, (x_{i,j}, y_{i,j})) - \ell(h, (x_{i,j}^+, y_{i,j}^+)) \right) \right\} \right) \middle| \mathcal{C} \right] \\
&\leq (2kT_1 + 1)^d \exp \left(2\lambda^2 \sum_{i=1}^k \frac{(w_i)^2}{n_i} \right). \tag{52}
\end{aligned}$$

Here, the last inequality makes use of fact $|\ell(h, (x_{i,j}, y_{i,j})) - \ell(h, (x_{i,j}^+, y_{i,j}^+))| \leq 2$ as well as the following elementary inequality

$$\mathbb{E}_{\sigma_{i,j}} [\exp(\sigma_{i,j} x)] = \frac{1}{2} (\exp(x) + \exp(-x)) \leq \exp(0.5x^2).$$

Taking (48), (49) and (52) together reveals that

$$E(\lambda) \leq (2kT_1 + 1)^d \exp \left(2\lambda^2 \sum_{i=1}^k \frac{(w_i)^2}{n_i} \right). \tag{53}$$

Repeating the same arguments also yields an upper bound on the following quantity:

$$\begin{aligned}
\bar{E}(\lambda) &:= \mathbb{E}_{\bar{\mathcal{S}}(n)} \left[\max_{h \in \mathcal{H}} \exp \left(\lambda \left\{ L(h, w) - \sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} \ell(h, (x_{i,j}, y_{i,j})) \right\} \right) \right] \\
&\leq (2kT_1 + 1)^d \exp \left(2\lambda^2 \sum_{i=1}^k \frac{(w_i)^2}{n_i} \right)
\end{aligned}$$

for any $\lambda \in [0, \min_{i \in [k]} \frac{n_i}{w_i}]$. Taking the above two inequalities and applying the Markov inequality reveal that, for any $0 < \varepsilon' \leq 1$,

$$\begin{aligned}
& \mathbb{P} \left(\max_{h \in \mathcal{H}} \left| \sum_{i=1}^k w_i \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(h, (x_{i,j}, y_{i,j})) - L(h, w) \right| \geq \varepsilon' \right) \\
&\leq \min_{0 \leq \lambda \leq \min_i \frac{n_i}{w_i}} \frac{E(\lambda) + \bar{E}(\lambda)}{\exp(\lambda \varepsilon')} \\
&\leq \min_{0 \leq \lambda \leq \min_i \frac{n_i}{w_i}} 2 \cdot (2kT_1 + 1)^d \exp \left(2\lambda^2 \sum_{i=1}^k \frac{(w_i)^2}{n_i} - \lambda \varepsilon' \right). \tag{54}
\end{aligned}$$

Step 2: uniform concentration bounds over epsilon-nets w.r.t. n and w . Next, we move on to extend the above result to uniform concentration bounds over all possible n and w . Towards this, let us first introduce a couple of notation.

- Let us use $\Delta_{\varepsilon_2}(k) \subseteq \Delta(k)$ to denote an ε_2 -net of $\Delta(k)$ — namely, for any $x \in \Delta(k)$, there exists a vector $x_0 \in \Delta_{\varepsilon_2}(k)$ obeying $\|x - x_0\|_\infty \leq \varepsilon_2$. We shall choose $\Delta_{\varepsilon_2}(k)$ properly so that

$$|\Delta_{\varepsilon_2}(k)| \leq (1/\varepsilon_2)^k.$$

- Define the following set

$$\mathcal{B} = \left\{ n = \{n_i\}_{i=1}^k, w = \{w_i\}_{i=1}^k \mid \frac{n_i}{w_i} \geq \frac{T_1}{2}, 0 \leq n_i \leq T_1, \forall i \in [k], w \in \Delta_{\varepsilon_1/(8k)}(k) \right\},$$

which clearly satisfies

$$|\mathcal{B}| \leq T_1^k \cdot \left(\frac{8k}{\varepsilon_1} \right)^k.$$

Applying the union bound yields that, for any $0 < \varepsilon' \leq 1$,

$$\begin{aligned} & \mathbb{P} \left(\exists (n, w) \in \mathcal{B}, \max_{h \in \mathcal{H}} \left| \sum_{i=1}^k w_i \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(h, (x_{i,j}, y_{i,j})) - L(h, w) \right| \geq \varepsilon' \right) \\ & \leq \sum_{(n, w) \in \mathcal{B}} \min_{0 \leq \lambda \leq \min_i \frac{n_i}{w_i}} 2 \cdot (2kT_1 + 1)^d \exp \left(2\lambda^2 \sum_{i=1}^k \frac{(w_i)^2}{n_i} - \lambda \varepsilon' \right) \\ & \leq \sum_{(n, w) \in \mathcal{B}} \min_{0 \leq \lambda \leq \frac{T_1}{2}} 2 \cdot (2kT_1 + 1)^d \exp \left(2\lambda^2 \cdot \frac{2}{T_1} - \lambda \varepsilon' \right) \\ & \leq \sum_{(n, w) \in \mathcal{B}} 2 \cdot (2kT_1 + 1)^d \exp \left(-\frac{T_1(\varepsilon')^2}{16} \right) \\ & \leq |\mathcal{B}| \cdot 2 \cdot (2kT_1 + 1)^d \exp \left(-\frac{T_1(\varepsilon')^2}{16} \right) \\ & \leq 2 \cdot (8kT_1/\varepsilon_1)^k (2kT_1 + 1)^d \cdot \exp \left(-\frac{T_1(\varepsilon')^2}{16} \right), \end{aligned}$$

where the second inequality holds since $\sum_{i=1}^k \frac{w_i^2}{n_i} \leq \frac{2}{T_1} \sum_{i=1}^k w_i = \frac{2}{T_1}$ (according to the definition of \mathcal{B}).

Step 3: concentration bounds w.r.t. n^t and w^t . Let \mathcal{S}^t denote the value of \mathcal{S} after line 10 of Algorithm 1 in the t -th round. Recall that $n^t = [n_i^t]_{1 \leq i \leq k}$ denotes the number of samples for all k distributions in \mathcal{S}^t , and let $w^t = [w_i^t]_{1 \leq i \leq k}$ represent the weight iterates in the t -th round. It is easily seen from lines 6 and 9 of Algorithm 1 that $n_i^t \leq T_1$ and $n_i^t/w_i^t \geq n_i^t/(2\hat{w}_i^t) \geq T_1/2$. For analysis purposes, it suffices to take $\mathcal{S}^t = \tilde{\mathcal{S}}(n^t)$.

In view of the update rule in Algorithm 1, one can always find $(n^t, \tilde{w}^t) \in \mathcal{B}$ satisfying $\|\tilde{w}^t - w^t\|_1 \leq k\|\tilde{w}^t - w^t\|_\infty \leq \varepsilon_1/8$. As a result, for any $0 < \varepsilon' \leq 1$, we can deduce that

$$\begin{aligned} & \mathbb{P} \left(\exists t \in [T], \max_{h \in \mathcal{H}} \left| \sum_{i=1}^k w_i^t \frac{1}{n_i^t} \sum_{j=1}^{n_i^t} \ell(h, (x_{i,j}, y_{i,j})) - L(h, w^t) \right| \geq \varepsilon' + \frac{\varepsilon_1}{4} \right) \\ & \leq \mathbb{P} \left(\exists t \in [T], \max_{h \in \mathcal{H}} \left| \sum_{i=1}^k \tilde{w}_i^t \frac{1}{n_i^t} \sum_{j=1}^{n_i^t} \ell(h, (x_{i,j}, y_{i,j})) - L(h, \tilde{w}^t) \right| \geq \varepsilon' \right) \\ & \leq 2 \cdot (8kT_1/\varepsilon_1)^k (2kT_1 + 1)^d \cdot \exp \left(-\frac{T_1(\varepsilon')^2}{16} \right), \end{aligned} \tag{55}$$

where the second inequality arises from the fact that $\frac{1}{n_i} \sum_{j=1}^{n_i} \ell(h, (x_{i,j}, y_{i,j})) \in [-1, 1]$ and $L(h, \tilde{w}^t) \in [-1, 1]$. Taking $\varepsilon' = \varepsilon_1/4$ and substituting $T_1 = \frac{4000(k \log(k/\varepsilon_1) + d \log(kd/\varepsilon_1) + \log(1/\delta))}{\varepsilon_1^2}$ into (55), we can obtain

$$\mathbb{P} \left(\exists t \in [T], \max_{h \in \mathcal{H}} \left| \sum_{i=1}^k w_i^t \cdot \frac{1}{n_i^t} \sum_{j=1}^{n_i^t} \ell(h, (x_{i,j}, y_{i,j})) - L(h, w^t) \right| \geq \frac{\varepsilon_1}{2} \right)$$

$$\begin{aligned}
&\leq 2 \cdot (8kT_1/\varepsilon_1)^k (2kT_1 + 1)^d \cdot \exp\left(-\frac{T_1 \varepsilon_1^2}{16}\right) \\
&\leq 2 \cdot (8kT_1/\varepsilon_1)^k (2kT_1 + 1)^d \cdot \exp\left(-10(k \log(k/\varepsilon_1) + d \log(kd/\varepsilon_1) + \log(1/\delta))\right) \\
&\leq 2 \cdot (8kT_1/\varepsilon_1)^k (2kT_1 + 1)^d \cdot (k/\varepsilon_1)^{-10k} \cdot (kd/\varepsilon_1)^{-10d} \cdot \delta \\
&\leq \delta/4.
\end{aligned} \tag{56}$$

Step 4: putting all this together. Recalling that

$$\hat{L}^t(h, w^t) = \sum_{i=1}^k w_i^t \cdot \frac{1}{n_i^t} \sum_{j=1}^{n_i^t} \ell(h, (x_{i,j}, y_{i,j})),$$

one can see from (56) that, with probability exceeding $1 - \delta/4$,

$$\left| \hat{L}^t(h, w^t) - L(h, w^t) \right| \leq \frac{\varepsilon_1}{2} \tag{57}$$

holds simultaneously for all $t \in [T]$ and all $h \in \mathcal{H}$. Additionally, observing that

$$h^t = \arg \min_{h \in \mathcal{H}} \hat{L}^t(h, w^t), \tag{58}$$

we can immediately deduce that

$$L(h^t, w^t) \leq \hat{L}(h^t, w^t) + \frac{\varepsilon_1}{2} = \min_{h \in \mathcal{H}} \hat{L}(h, w^t) + \frac{\varepsilon_1}{2} \leq \min_{h \in \mathcal{H}} L(h, w^t) + \varepsilon_1. \tag{59}$$

This concludes the proof of Lemma 1.

C.2 Proof of Lemma 2

Before proceeding, let us introduce some additional notation. Let $\delta' := \frac{\delta}{4(T+k+1)}$, and define

$$\text{OPT} := \min_{h \in \mathcal{H}} \max_{1 \leq i \leq k} L(h, e_i)$$

to be the optimal objective value. Additionally, set

$$v^t := L(h^t, w^t) - \text{OPT}. \tag{60}$$

It follows from the assumption of this lemma (i.e., $L(h^t, w^t) \leq \min_{h \in \mathcal{H}} L(h, w^t) + \varepsilon_1$) that

$$v^t \leq \min_{h \in \mathcal{H}} L(h, w^t) - \text{OPT} + \varepsilon_1 = \min_{h \in \mathcal{H}} L(h, w^t) - \min_{h \in \mathcal{H}} \max_i L(h, e_i) + \varepsilon_1 \leq \varepsilon_1, \quad \forall 1 \leq t \leq T. \tag{61}$$

We now begin to present the proof. In view of the Azuma-Hoeffding inequality and the union bound, we see that with probability at least $1 - (k+1)\delta'$,

$$\left| \sum_{t=1}^T \langle w^t, \hat{r}^t \rangle - \sum_{t=1}^T L(h^t, w^t) \right| \leq 2\sqrt{T \log(1/\delta')}, \tag{62a}$$

$$\left| \sum_{t=1}^T \hat{r}_i^t - \sum_{t=1}^T L(h^t, e_i) \right| \leq 2\sqrt{T \log(1/\delta')}. \tag{62b}$$

These motivate us to look at $\sum_{t=1}^T \langle w^t, \hat{r}^t \rangle$ (resp. $\sum_{t=1}^T \hat{r}_i^t$) as a surrogate for $\sum_{t=1}^T L(h^t, w^t)$ (resp. $\sum_{t=1}^T L(h^t, e_i)$).

We then resort to standard analysis for the Hedge algorithm. Specifically, direct computation gives

$$\log \left(\frac{\sum_{i=1}^k W_i^{t+1}}{\sum_{i=1}^k W_i^t} \right) \stackrel{(i)}{=} \log \left(\sum_{i=1}^k w_i^t \exp(\eta \hat{r}_i^t) \right) \stackrel{(ii)}{\leq} \log \left(\sum_{i=1}^k w_i^t (1 + \eta \hat{r}_i^t + \eta^2 (\hat{r}_i^t)^2) \right)$$

$$\leq \log \left(1 + \eta \sum_{i=1}^k w_i^t \hat{r}_i^t + \eta^2 \sum_{i=1}^k w_i^t (\hat{r}_i^t)^2 \right) \leq \eta \sum_{i=1}^k w_i^t \hat{r}_i^t + \eta^2. \quad (63)$$

Here, (i) is valid since $w_i^t = \frac{W_i^t}{\sum_j W_j^t}$ and $W_i^{t+1} = W_i^t \exp(\eta \hat{r}_i^t)$ (cf. lines 5 and 15 of Algorithm 1); (ii) arises from the elementary inequality $e^x \leq 1 + x + x^2$ for $x \in [0, 1]$ as well as the facts that $\eta \leq 1$ and $|\hat{r}_i^t| \leq 1$. Summing the inequality (63) over all t and rearranging terms, we are left with

$$\begin{aligned} \eta \sum_{t=1}^T \langle w^t, \hat{r}^t \rangle &\geq \sum_{t=1}^T \left\{ \log \left(\frac{\sum_{i=1}^k W_i^{t+1}}{\sum_{i=1}^k W_i^t} \right) - \eta^2 \right\} \\ &= \log \left(\sum_{i=1}^k W_i^{T+1} \right) - \log \left(\sum_{i=1}^k W_i^1 \right) - T\eta^2 \\ &\geq \max_{1 \leq i \leq k} \log(W_i^{T+1}) - \log(k) - T\eta^2 \\ &\geq \eta \max_{1 \leq i \leq k} \sum_{t=1}^T \hat{r}_i^t - \log(k) - T\eta^2, \end{aligned} \quad (64)$$

where the penultimate line makes use of $W_i^1 = 1$ for all $i \in [k]$, and the last line holds since $\log(W_i^{T+1}) = \log(W_i^T \exp(\eta \hat{r}_i^T)) \geq \eta \hat{r}_i^T$. Dividing both sides by η yields

$$\sum_{t=1}^T \langle w^t, \hat{r}^t \rangle \geq \max_i \sum_{t=1}^T \hat{r}_i^t - \left(\frac{\log(k)}{\eta} + \eta T \right). \quad (65)$$

Combine the above inequality with (62) to show that, with probability at least $1 - (k+1)\delta'$,

$$\sum_{t=1}^T L(h^t, w^t) \geq \max_{1 \leq i \leq k} \sum_{t=1}^T L(h^t, e_i) - \left(\frac{\log(k)}{\eta} + \eta T + 4\sqrt{T \log(1/\delta')} \right). \quad (66)$$

Recalling that $\varepsilon_1 = \eta = \frac{1}{100}\varepsilon$ and $T = \frac{20000 \log(\frac{k}{\delta'\varepsilon})}{\varepsilon^2}$, we can derive

$$\begin{aligned} \max_{1 \leq i \leq k} \sum_{t=1}^T L(h^t, e_i) &\leq \text{TOPT} + \sum_{t=1}^T v^t + \left(\frac{\log(k)}{\eta} + \eta T + 4\sqrt{T \log(1/\delta')} \right) \\ &\leq \text{TOPT} + T\varepsilon_1 + \left(\frac{\log(k)}{\eta} + \eta T + 4\sqrt{T \log(1/\delta')} \right) \\ &\leq \text{TOPT} + T\varepsilon, \end{aligned} \quad (67)$$

where the penultimate line results from (61). Given that h^{final} is taken to be uniformly distributed over $\{h^t\}_{1 \leq t \leq T}$, we arrive at

$$\max_{1 \leq i \leq k} L(h^{\text{final}}, e_i) = \max_{1 \leq i \leq k} \frac{1}{T} \sum_{t=1}^T L(h^t, e_i) \leq \text{OPT} + \varepsilon \quad (68)$$

with probability at least $1 - (k+1)\delta'$. This concludes the proof by recalling that $\delta' = \frac{\delta}{4(T+k+1)}$.

Remark 2. Note that the proof of this lemma works as long as $\hat{r}_i^t \in [0, 1]$ is an unbiased estimate of $L(h^t, e_i)$ for each $i \in [k]$, regardless of how many samples are used to construct \hat{r}_i^t .

C.3 Proof of Lemma 3

Set $\delta' = \delta/(32T^4k^2)$, and define

$$\mathcal{W}_j := \{i \in [k] \mid \max_{1 \leq t \leq T} w_i^t \in (2^{-j}, 2^{-(j-1)}]\}, \quad 1 \leq j \leq \lfloor \log_2(k) \rfloor + 1 \quad (69a)$$

$$\overline{\mathcal{W}} := [k]/\cup_j \mathcal{W}_j. \quad (69b)$$

In other words, we divide the k distributions into a logarithmic number of groups $\{\mathcal{W}_j\}$, where each \mathcal{W}_j consists of those distributions whose corresponding $\max_t w_i^t$ are on the same order. The main step in establishing Lemma 3 lies in bounding the size of each \mathcal{W}_j , as summarized below.

Lemma 13. *Suppose that the assumptions of Lemma 3 hold. Then with probability exceeding $1 - 8T^4 k \delta'$,*

$$|\mathcal{W}_j| \leq 8 \cdot 10^7 \cdot \left((\log_2(k) + 1)^4 (\log(k) + \log(1/\delta'))^2 (\log_2(T) + 1) \right) \cdot 2^j \quad (70)$$

holds all $1 \leq j \leq \log_2(k) - 2$.

In words, Lemma 13 asserts that the cardinality of each \mathcal{W}_j is upper bounded by

$$|\mathcal{W}_j| \leq \tilde{O}(2^j).$$

Importantly, this lemma tells us that, with probability at least $1 - 8T^4 k^2 \delta' = 1 - \delta/4$, one has

$$\begin{aligned} \|\overline{w}^T\|_1 &= \sum_{i=1}^k \max_{1 \leq t \leq T} w_i^t \leq k \cdot 2^{-(\lfloor \log_2(k) \rfloor - 2 - 1)} + \sum_{j=1}^{\lfloor \log_2(k) \rfloor - 2} |\mathcal{W}_j| 2^{-(j-1)} \\ &\leq k \cdot \frac{16}{k} + \sum_{j=1}^{\lfloor \log_2(k) \rfloor - 2} |\mathcal{W}_j| 2^{-(j-1)} \\ &\leq 2 \cdot 10^8 \cdot \left((\log_2(k) + 1)^5 (\log(Tk) + \log(1/\delta))^2 (\log_2(T) + 1) \right), \end{aligned}$$

where the first inequality is valid since $\max_{1 \leq t \leq T} w_i^t \leq 2^{-(j-1)}$ holds for any $i \in \mathcal{W}_j$. This immediately concludes the proof of Lemma 3, as long as Lemma 13 can be established. Proving Lemma 13 is the most challenging part of our analysis, and we dedicate the next section (Section D) to the proof of Lemma 13.

D Controlling the Hedge trajectory (proof of Lemma 13)

This section is devoted to proving Lemma 13. The proof relies heavily on the concepts of “segments” and “configurations” introduced in Section 4.2. For convenience, we restate these definitions below.

Definition 5 (Segment (restated)). *For any $p, x > 0$ and $i \in [k]$, we say that (t_1, t_2) is a (p, q, x) -segment if there exists a subset $\mathcal{I} \subseteq [k]$ such that*

- (i) $\sum_{i \in \mathcal{I}} w_i^{t_1} \in [p/2, p]$,
- (ii) $\sum_{i \in \mathcal{I}} w_i^{t_2} \geq p \exp(x)$,
- (iii) $\sum_{i \in \mathcal{I}} w_i^t \geq q$ for any $t_1 \leq t \leq t_2$.

We shall refer to t_1 as the starting point and t_2 as the ending point, and call \mathcal{I} the index set. Moreover, two segments (s_1, e_1) and (s_2, e_2) are said to be disjoint if $s_1 < e_1 \leq s_2 < e_2$ or $s_2 < e_2 \leq s_1 < e_1$.

Definition 6 (Configuration (restated)). *A configuration Conf is a set of intervals $\text{Conf} = \{[a_i, b_i]\}_{i=1}^m$ obeying $b_i > a_i$ for each $i \in [m]$ (note that repeated elements are allowed). A configuration Conf is said to be regular if, for any $i, j \in [m]$, one of the following three properties holds:*

- (a) $a_i < b_i \leq a_j < b_j$;
- (b) $a_j < b_j \leq a_i < b_i$;
- (c) $a_i = a_j, b_i = b_j$.

In addition, we shall take $\delta' = \delta/(32T^4 k^2)$ throughout this subsection, and we focus on any $j \in [1, \lfloor \log_2(k) \rfloor] - 2$.

D.1 Outline

Before delving into the proof, we first describe the high-level idea. In Lemma 17, we provide a lower bound on the length of a (p, q, x) segment. We then proceed to prove that: if $|\mathcal{W}_j|$ is large, then there exist many disjoint segments, thereby requiring the total length T to be large enough in order to contain these segments.

As discussed in Section 4.2, we will extract a regular configuration from a general configuration. At a high level, the proof consists of the following steps.

1. Identify a suitable segment for each $i \in \mathcal{W}_j$ (see Lemma 14 in Section D.2);
2. Identify some disjoint blocks such that (see Lemma 15 and Figure 3):
 - The segments in the same blocks have a common inner point;
 - The number of segments in these blocks is at least $\frac{1}{3(\log_2(T)+1)}$ times the number of all the segments, i.e., $\frac{1}{3(\log_2(T)+1)}|\mathcal{W}_j|$;
 - Continue the analysis on a single block in view of the fact that these blocks are disjoint;
3. For each single block, use the common inner point to align either the starting points or the ending points of at least half of these segments (see Lemma 16 and Figure 4);
4. Design a group of regular configurations (at most $O(\log_2(k))$ groups) such that at least one of the regular configurations contains enough segments with significant variation (see Lemma 16 and Figure 6).

In the sequel, we shall present the details of each of these steps.

D.2 Step 1: identifying segments for each distribution in \mathcal{W}_j

Recall that \mathcal{W}_j contains those distributions whose corresponding weight iterates obey $\max_{1 \leq t \leq T} w_i^t \in (2^{-j}, 2^{-j+1}]$ (cf. (69a)). As it turns out, for any $i \in \mathcal{W}_j$, one can find an $(\frac{1}{2^{j+1}}, \frac{1}{2^{j+2}}, \log(2))$ -segment, as stated in the lemma below. This basic fact allows one to link each distribution in \mathcal{W}_j with a segment of suitable parameters.

Lemma 14. *For each $i \in \mathcal{W}_j$, there exists $1 \leq s_i < e_i \leq T$, such that*

$$\frac{1}{2^{j+2}} < w_i^{s_i} \leq \frac{1}{2^{j+1}}, \quad w_i^{e_i} > \frac{1}{2^j}, \quad \text{and} \quad w_i^t > 2^{-(j+2)} \quad \forall t \in [s_i, e_i]. \quad (71)$$

In other words, there exists a $(\frac{1}{2^{j+1}}, \frac{1}{2^{j+2}}, \log(2))$ -segment (s_i, e_i) with the index set as $\{i\}$ (see Definition 5).

Proof. From the definition (69a) of \mathcal{W}_j , it is straightforward to find a time point e_i obeying $w_i^{e_i} > \frac{1}{2^j}$. It then remains to identify a valid point s_i . To this end, let us define

$$\tau = \max \{t \mid t \leq e_i, w_i^t \leq 2^{-(j+2)}\},$$

which is properly defined since $w_i^1 = 1/k \leq 2^{-(j+2)}$. With this choice in mind, we have

$$w_i^t > 2^{-(j+2)}, \quad \forall t \text{ obeying } \tau + 1 \leq t \leq e_i.$$

In addition, it follows from the update rule (cf. lines 5 and 15 of Algorithm 1) that

$$\begin{aligned} \log(w_i^{t+1}/w_i^t) &= \log(W_i^{t+1}/W_i^t) - \log\left(\sum_j W_j^{t+1}/\sum_j W_j^t\right) \\ &\leq \eta - \log\left(\sum_j W_j^{t+1}/\sum_j W_j^t\right) \leq 2\eta \leq 1/10, \end{aligned}$$

which in turn allows us to show that

$$w_i^{\tau+1} \leq w_i^\tau \exp(1/10) \leq \frac{1}{2^{j+2}} \cdot \exp(1/10) \leq \frac{1}{2^{j+1}}. \quad (72)$$

As a result, it suffices to choose $s_i = \tau + 1$, thus concluding the proof. \square

D.3 Step 2: extracting regular segments from irregular segments

Lemma 15. Recall the definition of \mathcal{W}_j in (69a). For each $i \in \mathcal{W}_j$, denote by (s_i, e_i) the segment identified in Lemma 14. Then there exist a group of disjoint subsets $\{\mathcal{W}_j^p\}_{p=1}^P$ of \mathcal{W}_j obeying

- (i) $\mathcal{W}_j^p \subseteq \mathcal{W}_j$, $\mathcal{W}_j^p \cap \mathcal{W}_j^{p'} = \emptyset$, $\forall p \neq p'$;
- (ii) $\sum_{p=1}^P |\mathcal{W}_j^p| \geq \frac{|\mathcal{W}_j|}{3(\log_2(T)+1)}$;
- (iii) Let $\tilde{s}_p = \min_{i \in \mathcal{W}_j^p} s_i$ and $\tilde{e}_p = \max_{i \in \mathcal{W}_j^p} e_i$ for each $1 \leq p \leq P$. One has $1 \leq \tilde{s}_1 < \tilde{e}_1 \leq \tilde{s}_2 < \tilde{e}_2 \leq \dots \leq \tilde{s}_P < \tilde{e}_P \leq T$ and $\max_{i \in \mathcal{W}_j^p} s_i \leq \min_{i \in \mathcal{W}_j^p} e_i$ for each $1 \leq p \leq P$.

In words, Lemma 15 reveals the existence a collection of disjoint subsets of \mathcal{W}_j such that (a) they account for a sufficiently large fraction of indices in \mathcal{W}_j , and (b) the starting points and end points of their associated segments can be well sorted in the sense that $\tilde{s}_1 < \tilde{e}_1 \leq \tilde{s}_2 < \tilde{e}_2 \leq \dots \leq \tilde{s}_P < \tilde{e}_P$.

Proof. For any integer $1 \leq x \leq \log_2(T) + 1$, define

$$\mathcal{W}_j(x) := \{i \in [k] \mid 2^{x-1} \leq e_i - s_i \leq 2^x\},$$

so that the length of each segment associated with $\mathcal{W}_j(x)$ lies within $[2^{x-1}, 2^x]$. Let x^* indicate the one that maximizes the cardinality of $\mathcal{W}_j(x)$:

$$x^* = \arg \max_{1 \leq x \leq \log_2(T)+1} |\mathcal{W}_j(x)|.$$

Given that there are at most $\log_2(T) + 1$ choices of x , the pigeonhole principle gives

$$|\mathcal{W}_j(x^*)| \geq \frac{|\mathcal{W}_j|}{\log_2(T) + 1}. \quad (73)$$

In the sequel, we intend to choose the subsets $\{\mathcal{W}_j^m\}_{m=1}^M$ from $\mathcal{W}_j(x^*)$.

To proceed, let us set

$$\kappa_1 := \min_{i \in \mathcal{W}_j(x^*)} e_i, \quad \mathcal{U}_j^1 := \{i \in \mathcal{W}_j(x^*) \mid s_i \leq \kappa_1\}, \quad (74a)$$

and then for each $o \geq 1$, take

$$\kappa_{o+1} := \min_{i \in \mathcal{W}_j(x^*) / \cup_{o'=1}^o \mathcal{U}_j^{o'}} e_i, \quad (74b)$$

$$\mathcal{U}_j^{o+1} := \{i \in \mathcal{W}_j(x^*) / \cup_{o'=1}^o \mathcal{U}_j^{o'} \mid s_i \leq \kappa_{o+1}\}. \quad (74c)$$

We terminate such constructions until $\cup_{o \geq 1} \mathcal{U}_j^o = \mathcal{W}_j(x^*)$. By construction, for each o , we have

$$s_{i_2} \leq \kappa_o \leq e_{i_1}, \quad \forall i_1, i_2 \in \mathcal{U}_j^o \iff \max_{i \in \mathcal{U}_j^o} s_i \leq \min_{i \in \mathcal{U}_j^o} e_i. \quad (75)$$

Let us look at the three groups of subsets of $\mathcal{W}_j(x^*)$: $\{\mathcal{U}_j^{3o-2}\}_{o \geq 1}$, $\{\mathcal{U}_j^{3o-1}\}_{o \geq 1}$ and $\{\mathcal{U}_j^{3o}\}_{o \geq 1}$. Clearly, there exists $\ell \in \{0, 1, 2\}$ such that $\sum_{o \geq 1} |\mathcal{U}_j^{3o-\ell}| \geq \frac{1}{3} \sum_{o \geq 1} |\mathcal{U}_j^o|$; without loss of generality, assume that

$$\sum_{o \geq 1} |\mathcal{U}_j^{3o-2}| \geq \frac{1}{3} \sum_{o \geq 1} |\mathcal{U}_j^o| = \frac{1}{3} |\mathcal{W}_j(x^*)|. \quad (76)$$

With the above construction in place, we would like to verify that $\{\mathcal{U}_j^{3o-2}\}_{o \geq 1}$ forms the desired group of subsets. First of all, Condition (i) holds directly from the definition of $\{\mathcal{U}_j^o\}_{o \geq 1}$. When it comes to Condition (ii), it follows from (76) and (73) that

$$\sum_{o \geq 1} |\mathcal{U}_j^{3o-2}| \geq \frac{1}{3} |\mathcal{W}_j(x^*)| \geq \frac{|\mathcal{W}_j|}{3(\log_2(T) + 1)}.$$

Regarding Condition (iii), it suffices to verify that

$$\max_{i \in \mathcal{U}_j^{3o-2}} e_i \leq \min_{i \in \mathcal{U}_j^{3o+1}} s_i \quad (77)$$

for any o . To do so, note that for each $o \geq 1$, there exists $i \in \mathcal{W}_j(x^*)$ such that $s_i \geq \kappa_o$ and $\kappa_{o+1} = e_i$. We can then deduce that

$$\kappa_{o+1} = e_i \geq s_i + 2^{x^*-1} \geq \kappa_o + 2^{x^*-1}. \quad (78)$$

It then follows that, for any $i \in \mathcal{U}_j^{3o+1}$, one has

$$s_i \geq \kappa_{3o} \geq \kappa_{3o-1} + 2^{x^*-1} \geq \kappa_{3o-2} + 2^{x^*}.$$

In addition, for any $\ell \in \mathcal{U}_j^{3o-2}$, it is seen that

$$e_\ell \leq s_\ell + 2^{x^*} \leq \kappa_{3o-2} + 2^{x^*}.$$

Putting all this together yields

$$\max_{i \in \mathcal{U}_j^{3o-2}} e_i \leq \kappa_{3o-2} + 2^{x^*} \leq \min_{i \in \mathcal{U}_j^{3o+1}} s_i.$$

The proof is thus complete. \square

Lemma 16. Recall the definition of \mathcal{W}_j in (69a). For each $i \in \mathcal{W}_j$, denote by (s_i, e_i) the segment identified in Lemma 14. Then there exists a group of subsets $\{\mathcal{V}_j^n\}_{n=1}^N$ satisfying the following properties:

- (i) $\mathcal{V}_j^n \subseteq \mathcal{W}_j$, $\mathcal{V}_j^n \cap \mathcal{V}_j^{n'} = \emptyset$, $\forall n \neq n'$;
- (ii) $\sum_{n=1}^N |\mathcal{V}_j^n| \geq \frac{|\mathcal{W}_j|}{24 \log_2(k)(\log_2(T)+1)}$;
- (iii) There exist $1 \leq \hat{s}_1 < \hat{e}_1 \leq \hat{s}_2 < \hat{e}_2 \leq \dots \leq \hat{s}_N < \hat{e}_N \leq T$, and $\{g_n\}_{n=1}^N \in [1, \infty)^N$, such that for each $1 \leq n \leq N$, (\hat{s}_n, \hat{e}_n) is a $\left(2^{-(j+1)} g_n |\mathcal{V}_j^n|, 2^{-(j+2)} |\mathcal{V}_j^n|, \frac{\log(2)}{2 \log_2(k)}\right)$ -segment with index set as \mathcal{V}_j^n . That is, the following properties hold for each $1 \leq n \leq N$:
 - (a) $\frac{g_n |\mathcal{V}_j^n|}{2^{j+2}} \leq \sum_{i \in \mathcal{V}_j^n} w_i^{\hat{s}_n} \leq \frac{g_n |\mathcal{V}_j^n|}{2^{j+1}}$;
 - (b) $\frac{g_n |\mathcal{V}_j^n|}{2^j} \cdot \exp\left(\frac{\log(2)}{2 \log_2(k)}\right) \leq \sum_{i \in \mathcal{V}_j^n} w_i^{\hat{e}_n}$;
 - (c) $\sum_{i \in \mathcal{V}_j^n} w_i^t \geq \frac{|\mathcal{V}_j^n|}{2^{j+2}}$ for any t obeying $\hat{s}_n \leq t \leq \hat{e}_n$.

Proof. We shall begin by presenting our construction of the subsets, followed by justification of the advertised properties. In what follows, we set $x = \log(2)$.

Our construction. Let $\{\mathcal{W}_j^p\}_{p=1}^P$ and $\{(\tilde{s}_p, \tilde{e}_p)\}_{p=1}^P$ be the construction in Lemma 15.

Step a): constructing $\widehat{\mathcal{W}}_j^p$. Consider any $1 \leq p \leq P$. Setting

$$t_{\text{mid}}^p := \min_{i \in \mathcal{W}_j^p} e_i,$$

we can derive, for each $i \in \mathcal{W}_j^p$, that

$$\max \left\{ \log \left(\frac{w_i^{e_i}}{w_i^{t_{\text{mid}}^p}} \right), \log \left(\frac{w_i^{t_{\text{mid}}^p}}{w_i^{s_i}} \right) \right\} \geq \frac{1}{2} \log \left(\frac{w_i^{e_i}}{w_i^{t_{\text{mid}}^p}} \right) + \frac{1}{2} \log \left(\frac{w_i^{t_{\text{mid}}^p}}{w_i^{s_i}} \right) = \frac{1}{2} \log \left(\frac{w_i^{e_i}}{w_i^{s_i}} \right) \geq \frac{x}{2},$$

where the last inequality holds since (s_i, e_i) is constructed to be a $(\frac{1}{2^{j+1}}, \frac{1}{2^{j+2}}, x)$ -segment (see Lemma 14). It then follows that

$$\sum_{i \in \mathcal{W}_j^p} \left(\mathbb{1} \left\{ \log \left(\frac{w_i^{e_i}}{w_i^{t_{\text{mid}}^p}} \right) \geq \frac{x}{2} \right\} + \mathbb{1} \left\{ \log \left(\frac{w_i^{e_i}}{w_i^{t_{\text{mid}}^p}} \right) \geq \frac{x}{2} \right\} \right) \geq |\mathcal{W}_j^p|.$$

Without loss of generality, we assume that

$$\sum_{i \in \mathcal{W}_j^p} \mathbb{1} \left\{ \log \left(\frac{w_i^{e_i}}{w_i^{t_{\text{mid}}^p}} \right) \geq \frac{x}{2} \right\} \geq \frac{|\mathcal{W}_j^p|}{2}, \quad (79)$$

This means that the set define below

$$\widehat{\mathcal{W}}_j^p := \left\{ i \in \mathcal{W}_j^p \mid \log(w_i^{e_i}/w_i^{t_{\text{mid}}^p}) \geq \frac{x}{2} \right\} \quad (80)$$

satisfies

$$|\widehat{\mathcal{W}}_j^p| \geq \frac{|\mathcal{W}_j^p|}{2}. \quad (81)$$

In what follows, we take¹⁰

$$Q := |\widehat{\mathcal{W}}_j^p|, \quad \tilde{\ell} := \max \{ \ell \geq 0 \mid 2^\ell \leq Q \} \quad \text{and} \quad \tilde{Q} := 2^{\tilde{\ell}}.$$

Without loss of generality, we assume

$$\widehat{\mathcal{W}}_j^p = \{1, 2, \dots, Q\} \quad \text{and} \quad e_1 \leq e_2 \leq \dots \leq e_Q. \quad (82)$$

Step b): constructing $\widehat{\mathcal{W}}_j^p(\ell)$. Let us take $e_0 = t_{\text{mid}}^p$, and employ $[e_0, e_k] \oplus a$ as a shorthand notation for $[e_a, e_{k+a}]$. We can then define a group of disjoint intervals of $[T]$ as follows:

$$\mathcal{K}_1 = \{[e_0, e_{2^{\tilde{\ell}-1}}]\}; \quad (83a)$$

$$\mathcal{K}_2 = \{[e_0, e_{2^{\tilde{\ell}-2}}], [e_0, e_{2^{\tilde{\ell}-2}}] \oplus 2^{\tilde{\ell}-1}\}; \quad (83b)$$

$$\mathcal{K}_3 = \{[e_0, e_{2^{\tilde{\ell}-3}}], [e_0, e_{2^{\tilde{\ell}-3}}] \oplus 2^{\tilde{\ell}-2}, [e_0, e_{2^{\tilde{\ell}-3}}] \oplus 2 \cdot 2^{\tilde{\ell}-2}, [e_0, e_{2^{\tilde{\ell}-3}}] \oplus 3 \cdot 2^{\tilde{\ell}-2}\}; \quad (83c)$$

...

$$\mathcal{K}_\ell = \{[e_0, e_{2^{\tilde{\ell}-\ell}}], [e_0, e_{2^{\tilde{\ell}-\ell}}] \oplus 2^{\tilde{\ell}-\ell+1}, [e_0, e_{2^{\tilde{\ell}-\ell}}] \oplus 2 \cdot 2^{\tilde{\ell}-\ell+1}, \dots, [e_0, e_{2^{\tilde{\ell}-\ell}}] \oplus (2^{\ell-1} - 1)2^{\tilde{\ell}-\ell+1}\}; \quad (83d)$$

...

$$\mathcal{K}_{\tilde{\ell}} = \{[e_{2^i}, e_{2^{i+1}}] \mid i = 0, 1, 2, \dots, 2^{\tilde{\ell}-1} - 1\}; \quad (83e)$$

$$\mathcal{K}_{\tilde{\ell}+1} = \{[e_{2^{i+1}}, e_{2^{i+2}}] \mid i = 0, 1, 2, \dots, 2^{\tilde{\ell}-1} - 1\}. \quad (83f)$$

For each $i \in [\tilde{Q} - 1]$ with binary form $\{i_\ell\}_{\ell=1}^{\tilde{\ell}}$ and $0 \leq \ell \leq \tilde{\ell}$, we define $\text{trunc}(i, \ell)$ to be the number with binary form $\{i_1, i_2, \dots, i_\ell, 0, 0, \dots, 0\}$. For example, $\text{trunc}(i, 0) = 0$ and $\text{trunc}(i, \tilde{\ell}) = i$.

From the definition (80) of $\widehat{\mathcal{W}}_j^p$, we know that for each $i \in [\tilde{Q} - 1]$,

$$\frac{x}{2} \leq \log \left(\frac{w_i^{e_i}}{w_i^{e_0}} \right) = \sum_{\ell=1}^{\tilde{\ell}} \log \left(\frac{w_i^{e_{\text{trunc}(i, \ell)}}}{w_i^{e_{\text{trunc}(i, \ell-1)}}} \right) = \sum_{\ell=1}^{\tilde{\ell}} \log \left(\frac{w_i^{e_{\text{trunc}(i, \ell)}}}{w_i^{e_{\text{trunc}(i, \ell-1)}}} \right) \mathbb{1} \{e_{\text{trunc}(i, \ell)} \neq e_{\text{trunc}(i, \ell-1)}\}, \quad (84)$$

¹⁰We assume $\tilde{Q} \geq 2$ without loss of generality. In the case $\tilde{Q} = 1$, we simply choose an arbitrary element in $\widehat{\mathcal{W}}_j^p$ as a single subset. In this way, we can collect at least $\frac{1}{4}|\widehat{\mathcal{W}}_j^p|$ segments.

which in turn implies that

$$\max_{1 \leq \ell \leq \tilde{\ell}} \log \left(\frac{w_i^{e_{\text{trunc}(i, \ell)}}}{w_i^{e_{\text{trunc}(i, \ell-1)}}} \right) \geq \frac{x}{2\tilde{\ell}}. \quad (85)$$

By defining

$$\tilde{\mathcal{W}}_j^p(\ell) := \left\{ i \in \widehat{\mathcal{W}}_j^p : \arg \max_{1 \leq \ell' \leq \tilde{\ell}} \log \left(\frac{w_i^{e_{\text{trunc}(i, \ell')}}}{w_i^{e_{\text{trunc}(i, \ell'-1)}}} \right) = \ell \right\}$$

for each¹¹ $1 \leq \ell \leq \tilde{\ell}$, we can demonstrate that

$$\sum_{\ell=1}^{\tilde{\ell}} |\tilde{\mathcal{W}}_j^p(\ell)| \geq \tilde{Q} - 1, \quad (86)$$

thus implying the existence of some $1 \leq \ell^* \leq \tilde{\ell}$ obeying

$$|\tilde{\mathcal{W}}_j^p(\ell^*)| \geq \frac{\tilde{Q} - 1}{\tilde{\ell}} \geq \frac{\tilde{Q}}{2\tilde{\ell}}. \quad (87)$$

Step c): constructing $\tilde{\mathcal{W}}_j^p(\ell, o)$, $\hat{s}(p, o)$ and $\hat{e}(p, o)$. By definition, for any i , if $\text{trunc}(i, \ell^*) \neq \text{trunc}(i, \ell^* - 1)$, then one has

$$[e_{\text{trunc}(i, \ell^*-1)}, e_{\text{trunc}(i, \ell^*)}] \in \mathcal{K}_{\ell^*},$$

where the set \mathcal{K}_ℓ has been defined in (83). In addition, from the construction of $\tilde{\mathcal{W}}_j^p(\ell^*)$ (see (87)), we know that $\text{trunc}(i, \ell^*) \neq \text{trunc}(i, \ell^* - 1)$ for any $i \in \tilde{\mathcal{W}}_j^p(\ell^*)$. For each $1 \leq o \leq 2^{\ell^*-1}$, define

$$\tilde{\mathcal{W}}_j^p(\ell^*, o) := \left\{ i \in \tilde{\mathcal{W}}_j^p(\ell^*) \mid [e_{\text{trunc}(i, \ell^*-1)}, e_{\text{trunc}(i, \ell^*)}] = [e_0, e_{2^{\tilde{\ell}-\ell^*}}] \oplus (o-1)2^{\tilde{\ell}-\ell^*+1} \right\}, \quad (88)$$

where we employ the notation ℓ^* and $\tilde{\ell}$ to abbreviate $\ell^*(p)$ and $\tilde{\ell}(p)$, respectively.

In addition, for any $1 \leq p \leq P$ and $1 \leq o \leq 2^{\ell^*(p)-1}$, we set

$$\hat{s}(p, o) = e_{(o-1)2^{\tilde{\ell}(p)-\ell^*(p)+1}}, \quad (89a)$$

$$\hat{e}(p, o) = e_{2^{\tilde{\ell}(p)-\ell^*(p)} + (o-1)2^{\tilde{\ell}(p)-\ell^*(p)+1}}. \quad (89b)$$

In words, $[\hat{s}(p, o), \hat{e}(p, o)]$ can be understood as the o -th interval in the set $\mathcal{K}_{\ell^*(p)}$.

Step d): construction output. With the above construction in mind, we would like to select

$$\left\{ \left\{ \tilde{\mathcal{W}}_j^p(\ell^*(p), o) \right\}_{o=1}^{2^{\ell^*(p)-1}} \right\}_{p=1}^P \quad \text{with intervals} \quad \left\{ \left\{ \hat{s}(p, o), \hat{e}(p, o) \right\}_{o=1}^{2^{\ell^*(p)-1}} \right\}_{p=1}^P$$

as the group of subsets we construct. With a slight abuse of notation, we use (p, o) as the index of the segments instead of n . In what follows, we verify the validity of this construction.

Verification of the advertised properties. We now proceed to justify the claimed properties.

Property (i). By construction, it is clearly seen that

$$\tilde{\mathcal{W}}_j^p(\ell^*(p), o) \subseteq \tilde{\mathcal{W}}_j^p(\ell^*(p)) \subseteq \widehat{\mathcal{W}}_j^p \subseteq \mathcal{W}_j^p \subseteq \mathcal{W}_j.$$

In addition, if

$$\tilde{\mathcal{W}}_j^{p_1}(\ell^*(p_1), o_1) \cap \tilde{\mathcal{W}}_j^{p_2}(\ell^*(p_2), o_2) \neq \emptyset,$$

then one has $\mathcal{W}_j^{p_2} \cap \mathcal{W}_j^{p_1} \neq \emptyset$, and as a result, $p_1 = p_2$ (otherwise it violates the condition that $\mathcal{W}_j^{p_2} \cap \mathcal{W}_j^{p_1} = \emptyset$ for $p_1 \neq p_2$). It also follows from the definition in (88) that $o_1 = o_2$. Therefore, for any (p_1, o_1) that does not equal (p_2, o_2) , we have $\tilde{\mathcal{W}}_j^{p_1}(\ell^*(p_1), o_1) \cap \tilde{\mathcal{W}}_j^{p_2}(\ell^*(p_2), o_2) = \emptyset$.

¹¹Without loss of generality, we assume the $\arg \max$ function is a single-valued function.

Property (ii). By construction, we have

$$\sum_{o=1}^{2^{\ell^*(p)-1}} \left| \widetilde{\mathcal{W}}_j^p(\ell^*(p), o) \right| = \left| \widetilde{\mathcal{W}}_j^p(\ell^*(p)) \right| \geq \frac{|\widehat{\mathcal{W}}_j^p|}{4 \log_2(|\widehat{\mathcal{W}}_j^p|)} \geq \frac{|\mathcal{W}_j^p|}{8 \log_2(|\widehat{\mathcal{W}}_j^p|)}, \quad (90)$$

where we have made use of (87) and (81). Summing over p and applying Lemma 15 yield

$$\sum_{p=1}^P \sum_{o=1}^{2^{\ell^*(p)-1}} \left| \widetilde{\mathcal{W}}_j^p(\ell^*(p), o) \right| \geq \sum_{p=1}^P \frac{|\mathcal{W}_j^p|}{8 \log_2(k)} \geq \frac{|\mathcal{W}_j|}{24 \log_2(k) (\log_2(T) + 1)}. \quad (91)$$

Property (iii)(a). Let us set the parameters $\left\{ \{g(p, o)\}_{o=1}^{2^{\ell^*(p)-1}} \right\}_{p=1}^P$ as follows:

$$g(p, o) = \frac{\sum_{i \in \widetilde{\mathcal{W}}_j(\ell^*(p), o)} w_i^{\widehat{s}(p, o)}}{2^{-(j+2)} \cdot \left| \widetilde{\mathcal{W}}_j^p(\ell^*(p), o) \right|} \geq 1.$$

Then Property (iii)(a) is satisfied since

$$\sum_{i \in \widetilde{\mathcal{W}}_j(\ell^*(p), o)} w_i^{\widehat{s}(p, o)} = \frac{g(p, o) \left| \widetilde{\mathcal{W}}_j^p(\ell^*(p), o) \right|}{2^{j+2}}.$$

Property (iii)(b). For any $i \in \widehat{\mathcal{W}}_j^p \subseteq \mathcal{W}_j^p$, we have

$$s_i \leq e_{\text{trunc}(i, \ell-1)} \leq e_i \quad \text{for any } 1 \leq \ell \leq \widetilde{\ell}(p),$$

which is valid since $\max_{i \in \mathcal{W}_j^p} s_i \leq \min_{i \in \mathcal{W}_j^p} e_i$ (see Lemma 15) and (82). It then holds that

$$s_i \leq \widehat{s}(p, o) \leq e_i \quad \text{for any } i \in \widehat{\mathcal{W}}_j^p.$$

Also, the definition of (s_i, e_i) (see Lemma 14) tells us that $w_i^{\widehat{s}(p, o)} \geq 2^{-(j+2)}$.

Also, by construction, we know that for any $i \in \widetilde{\mathcal{W}}_j^p(\ell^*(p), o)$,

$$\log \left(\frac{w_i^{\widehat{e}(p, o)}}{w_i^{\widehat{s}(p, o)}} \right) \geq \frac{x}{2\widetilde{\ell}(p)} \quad \text{and} \quad w_i^{\widehat{s}(p, o)} \geq 2^{-(j+2)}.$$

Recalling that $x = \log(2)$, one can further derive

$$\begin{aligned} \sum_{i \in \widetilde{\mathcal{W}}_j^p(\ell^*(p), o)} w_i^{\widehat{s}(p, o)} &\geq 2^{-(j+2)} \cdot \left| \widetilde{\mathcal{W}}_j^p(\ell^*(p), o) \right| \\ \log \left(\frac{\sum_{i \in \widetilde{\mathcal{W}}_j^p(\ell^*(p), o)} w_i^{\widehat{e}(p, o)}}{\sum_{i \in \widetilde{\mathcal{W}}_j^p(\ell^*(p), o)} w_i^{\widehat{s}(p, o)}} \right) &\geq \log \left(\frac{\sum_{i \in \widetilde{\mathcal{W}}_j^p(\ell^*(p), o)} w_i^{\widehat{s}(p, o)} \cdot \exp \left(\frac{x}{2\widetilde{\ell}(p)} \right)}{\sum_{i \in \widetilde{\mathcal{W}}_j^p(\ell^*(p), o)} w_i^{\widehat{s}(p, o)}} \right) = \frac{x}{2\widetilde{\ell}(p)} \geq \frac{\log(2)}{2 \log_2(k)}. \end{aligned}$$

Property (iii)(c). Note that for any t obeying $\widehat{s}(p, o) \leq t \leq \widehat{e}(p, o)$, and any $i \in \widetilde{\mathcal{W}}_j^p(\ell^*(p), o)$, it holds that $s_i \leq \widehat{s}(p, o) \leq t \leq \widehat{e}(p, o) \leq e_i$. As a result, we have

$$\sum_{i \in \widetilde{\mathcal{W}}_j^p(\ell^*(p), o)} w_i^t \geq \left| \widetilde{\mathcal{W}}_j^p(\ell^*(p), o) \right| \cdot 2^{-(j+2)}.$$

Proper ordering. To finish up, it remains to verify that the intersection of $[\hat{s}(p_1, o_1), \hat{e}(p_1, o_1)]$ and $[\hat{s}(p_2, o_2), \hat{e}(p_2, o_2)]$ is either empty or contains only the boundary points, unless $(p_1, o_1) = (p_2, o_2)$. To show this, note that in the case where $p_1 \neq p_2$ (assuming $p_1 < p_2$), we have

$$\tilde{s}_{p_1} \leq \hat{s}(p_1, o_1) < \hat{e}(p_1, o_1) \leq \tilde{e}_{p_1} \leq \tilde{s}_{p_2} \leq \hat{s}(p_2, o_2) < \hat{e}(p_2, o_2),$$

which arises from Lemma 15. Also, in the case where $p_1 = p_2 = p$ and $o_1 \neq o_2$ (assuming $o_1 < o_2$), we have

$$\hat{s}(p, o_1) < \hat{e}(p, o_1) < \hat{s}(p, o_2) < \hat{e}(p, o_2),$$

which comes from the construction (89).

We have thus completed the proof of this lemma. \square

D.4 Step 3: bounding the length of segments

Recall the definition of *segment* in Definition 1, as well as the definition (60) of v^t as follows

$$v^t := L(h^t, w^t) - \text{OPT}. \quad (92)$$

We have the following lemma to bound the length of segments.

Lemma 17. *Assume the conditions in Lemma 3 hold. Suppose (t_1, t_2) is a (p, q, x) -segment satisfying $p \geq 2q > 0$. Then one has*

$$t_2 - t_1 \geq \frac{x}{2\eta}. \quad (93)$$

Moreover, if

$$\frac{qx^2}{50(\log_2(k) + 1)^2} \geq \max \left\{ 2\eta \log \left(\frac{1}{\delta'} \right), \frac{1}{k} \right\} \quad (94)$$

holds, then with probability exceeding $1 - 6T^4 k \delta'$, at least one of the following two claims holds:

(1) there exists $1 \leq \hat{j} \leq \log_2(1/\eta) + 1$ such that

$$\sum_{\tau=t_1}^{t_2-1} \mathbf{1}\{-v^\tau \geq 2^{-\hat{j}}\} \geq \frac{qx^2 \cdot 2^{\hat{j}-1}}{100(\log_2(1/\eta) + 1)^2 \eta} \geq \frac{qx^2 \cdot 2^{\hat{j}-1}}{100(\log_2(k) + 1)^2 \eta}; \quad (95)$$

(2) the length of the segment satisfies

$$t_2 - t_1 \geq \frac{qx^2}{100(\log_2(1/\eta) + 1)^2 \eta^2} \geq \frac{qx^2}{100(\log_2(k) + 1)^2 \eta^2}. \quad (96)$$

Proof. See Section D.6. \square

D.5 Putting all this together

With the above lemmas in place, we are positioned to establish Lemma 13. Denote by $\{\mathcal{V}_j^n\}_{n=1}^N$ and $\{(\hat{s}_n, \hat{e}_n)\}_{n=1}^N$ the construction in Lemma 16. We divide $\{\mathcal{V}_j^n\}_{n=1}^N$ into two parts:

- The first part consists of those \mathcal{V}_j^n obeying

$$2^{-(j+2)} |\mathcal{V}_j^n| \cdot \frac{\log^2(2)}{50(\log_2(k) + 1)^2 \log_2^2(k)} \geq \max \{2 \log(1/\delta') \eta, 1/k\}, \quad (97)$$

which we shall denote as $\{\overline{\mathcal{V}}_j^n\}_{n=1}^{\overline{N}}$ in the sequel.

- The second part consists of the remaining sets, which we denote as $\{\mathcal{V}_j^n\}_{n=1}^{\bar{N}}$ in the sequel.

In view of Lemma 17, for each $1 \leq n \leq \bar{N}$, at least one of the following two claims holds true:

- there exists $1 \leq \hat{j} \leq \log_2(1/\eta) + 1$ such that

$$\sum_{\tau=\hat{s}_n}^{\hat{e}_n-1} \mathbb{1}\{-v^\tau \geq 2^{-\hat{j}}\} \geq \frac{2^{-(j+2)} |\bar{\mathcal{V}}_j^n| \log^2(2) \cdot 2^{\hat{j}-1}}{200 \log_2^2(k) (\log_2(k) + 1)^2 \eta}$$

- the length of the segment $[\hat{s}_n, \hat{e}_n]$ obeys

$$\hat{e}_n - \hat{s}_n \geq \frac{2^{-(j+2)} |\bar{\mathcal{V}}_j^n| \log^2(2)}{200 \log_2^2(k) (\log_2(k) + 1)^2 \eta^2}.$$

As a consequence, for each $1 \leq n \leq \bar{N}$, we have

$$\begin{aligned} (\hat{e}_n - \hat{s}_n)\eta + \sum_{\tau=\hat{s}_n}^{\hat{e}_n-1} \sum_{\hat{j}=1}^{\log_2(1/\eta)+1} \mathbb{1}\{-v^\tau \geq 2^{-\hat{j}}\} 2^{-(\hat{j}-1)} &= (\hat{e}_n - \hat{s}_n)\eta + \sum_{\hat{j}=1}^{\log_2(1/\eta)+1} \sum_{\tau=\hat{s}_n}^{\hat{e}_n-1} \mathbb{1}\{-v^\tau \geq 2^{-\hat{j}}\} 2^{-(\hat{j}-1)} \\ &\geq \frac{2^{-(j+2)} |\bar{\mathcal{V}}_j^n| \log^2(2)}{200 \log_2^2(k) (\log_2(k) + 1)^2 \eta}. \end{aligned} \quad (98)$$

By observing that

$$\sum_{\hat{j}=1}^{\log_2(1/\eta)+1} \mathbb{1}\{x \geq 2^{-\hat{j}}\} \cdot 2^{-(\hat{j}-1)} \leq 4x$$

holds for any $x \geq 0$, we can combine this fact with (98) to derive

$$(\hat{e}_n - \hat{s}_n)\eta + \sum_{\tau=\hat{s}_n}^{\hat{e}_n-1} 4 \cdot (-v^\tau) \geq \frac{2^{-(j+2)} |\bar{\mathcal{V}}_j^n| \log^2(2)}{200 \log_2^2(k) (\log_2(k) + 1)^2 \eta}. \quad (99)$$

Summing over n and taking advantage of the property $1 \leq \hat{s}_1 \leq \hat{e}_1 \leq \hat{s}_2 \leq \dots \leq \hat{e}_n \leq T$ (see Lemma 16) give

$$T\eta + 4 \sum_{t=1}^T (-v^t) \geq \sum_{n=1}^{\bar{N}} \left((\hat{e}_n - \hat{s}_n)\eta + \sum_{\tau=\hat{s}_n}^{\hat{e}_n-1} 4 \cdot (-v^\tau) \right) \geq \frac{2^{-(j+2)} \sum_{n=1}^{\bar{N}} |\bar{\mathcal{V}}_j^n| \log^2(2)}{200 \log_2^2(k) (\log_2(k) + 1)^2 \eta}. \quad (100)$$

Moreover, it follows from (66) that

$$\sum_{t=1}^T (-v^t) \leq 100 \left(\frac{\log(k)}{\eta} + \eta T + 4\sqrt{T \log(1/\delta')} \right), \quad (101)$$

which taken together with (100) gives

$$\begin{aligned} \sum_{n=1}^{\bar{N}} |\bar{\mathcal{V}}_j^n| &\leq \frac{3200(\log_2(k) + 1)^4 \eta \cdot 2^{j+2}}{\log^2(2)} \cdot 100 \left(\frac{\log(k)}{\eta} + \eta T + 4\sqrt{T \log(1/\delta')} \right) \\ &\quad + \frac{800T(\log_2(k) + 1)^4 \cdot 2^{j+2} \eta^2}{\log^2(2)}. \end{aligned} \quad (102)$$

In addition, in view of the first part of Lemma 17, we can demonstrate that

$$\hat{e}_n - \hat{s}_n \geq \frac{\log(2)}{4\eta \log_2(k)},$$

which combined with the property $1 \leq \hat{s}_1 \leq \hat{e}_1 \leq \hat{s}_2 \leq \dots \leq \hat{e}_n \leq T$ (see Lemma 16) gives

$$\underline{N} \leq \sum_{n=1}^N (\hat{e}_n - \hat{s}_n) \frac{4\eta \log_2(k)}{\log(2)} \leq T \cdot \frac{4\eta \log_2(k)}{\log(2)}. \quad (103)$$

As a result, it can be readily seen that

$$\begin{aligned} \sum_{n=1}^N |\mathcal{V}_j^n| &\leq \underline{N} \cdot \left\{ \frac{50(\log_2(k) + 1)^2 \log_2^2(k)}{\log^2(2)} 2^{j+2} \max\{2 \log(1/\delta')\eta, 1/k\} \right\} \\ &\leq \frac{4T\eta \log_2(k)}{\log(2)} \cdot 2^{j+2} \frac{50(\log_2(k) + 1)^2 \log_2^2(k)}{\log^2(2)} \cdot \max\{2 \log(1/\delta')\eta, 1/k\} \\ &\leq \frac{1600 \cdot 2^j T \eta^2 (\log_2(k) + 1)^2 \log_2^3(k) \log(1/\delta')}{\log^3(2)} + \frac{800 \cdot 2^j T \eta (\log_2(k) + 1)^2 \log_2^3(k)}{k \log^3(2)}, \end{aligned} \quad (104)$$

where the first inequality comes from the definition of \mathcal{V}_j^n (cf. the complement condition of (97)), and the second inequality arises from (103).

To finish up, note that $2 \log(1/\delta')\eta \geq 1/k$ according to our parameter choice. Thus, combining (102) and (104), we arrive at

$$\sum_{n=1}^N |\mathcal{V}_j^n| \leq 3200000(\log_2(k) + 1)^3 (\log_2(k) + \log(1/\delta'))^3 \cdot 2^j. \quad (105)$$

It then follows from Property (ii) of Lemma 16 that

$$\begin{aligned} |\mathcal{W}_j| &\leq 24 \log_2(k) (\log_2(T) + 1) \left(\sum_{n=1}^N |\mathcal{V}_j^n| \right) \\ &\leq 8 \cdot 10^7 \cdot ((\log_2(k) + 1)^4 (\log_2(k) + \log(1/\delta'))^3 (\log_2(T) + 1)) \cdot 2^j, \end{aligned}$$

thereby completing the proof.

D.6 Proof of Lemma 17

Throughout this proof, we find it convenient to denote $Z^t = \sum_{i=1}^k W_i^t$.

Part 1. We start by proving the first claim (93). Recall that $[t_1, t_2]$ is assumed to be a (p, q, x) -segment. From the definition of the segment (see Definition 5), there exists $i \in [k]$ such that

$$\log\left(\frac{w_i^{t_2}}{w_i^{t_1}}\right) \geq x.$$

Given that $W_i^{t_2} = W_i^{t_1} \exp(\eta \sum_{\tau=t_1}^{t_2-1} \hat{r}_i^\tau)$ and $w_t = W_t/Z_t$ (see lines 15 and 5 of Algorithm 1), the above inequality can be equivalently expressed as

$$\eta \sum_{\tau=t_1}^{t_2-1} \hat{r}_i^\tau - \log(Z^{t_2}/Z^{t_1}) \geq x. \quad (106)$$

Moreover, recognizing that

$$\log(Z^{t_2}/Z^{t_1}) = \log\left(\frac{\sum_{i \in [k]} W_i^{t_1} \exp(\eta \sum_{\tau=t_1}^{t_2-1} \hat{r}_i^\tau)}{\sum_{i \in [k]} W_i^{t_1}}\right) \geq -\eta(t_2 - t_1)$$

and $\hat{r}_i^\tau \leq 1$ for any $1 \leq \tau \leq T$, we can use (106) to show that

$$x \leq 2(t_2 - t_1)\eta, \quad (107)$$

from which the claimed inequality (93) follows.

Part 2. We now turn to the remaining claims of Lemma 17. For each hypothesis $h \in \mathcal{H}$, let us introduce the following vector $v_h \in \mathbb{R}^k$:

$$v_h = [v_{h,i}]_{i \in [k]} \quad \text{with} \quad v_{h,i} = L(h, e_i) - \text{OPT}. \quad (108)$$

Given the ε_1 -optimality of h^t (see Lemma 3), we have the following property that holds for any $1 \leq \tau, t \leq T$:

$$\langle v_{h^\tau}, w^t \rangle \geq \min_{h \in \mathcal{H}} \langle v_h, w^t \rangle \geq \langle v_{h^t}, w^t \rangle - \varepsilon_1 = v^t - \varepsilon_1, \quad (109)$$

where we recall the definition of v^t in (92). In the sequel, we divide the proof into a couple of steps.

Step 1: decomposing the KL divergence between w^t and w^{t_2} . Let us write

$$W_i^t = \exp \left(\eta \sum_{\tau=1}^t \eta \hat{r}_i^\tau \right) = \exp \left(\eta \sum_{\tau=1}^t (v_{h^\tau, i} + \text{OPT} + \xi_i^\tau) \right) \quad \text{with} \quad \xi_i^\tau = \hat{r}_i^\tau - v_{h^\tau, i} - \text{OPT},$$

where $\xi_i^\tau = \hat{r}_i^\tau - L(h^\tau, e_i)$ is clearly a zero-mean random variable. Define

$$\Delta_{t_1, t_2} = \sum_{\tau=t_1}^{t_2-1} \xi^\tau.$$

Taking $W^t = [W_i^t]_{i \in [k]}$ and denoting by $\log(x/y)$ the vector $\{\log(x_i/y_i)\}_{i \in [k]}$ for two k -dimensional vectors (x, y) , one can then deduce that

$$\left\langle \frac{1}{\eta} \log \left(\frac{W^{t_2}}{W^{t_1}} \right) - \Delta_{t_1, t_2}, w^t \right\rangle - (t_2 - t_1) \text{OPT} = \sum_{\tau=t_1}^{t_2-1} \langle v_{h^\tau}, w^t \rangle \geq (t_2 - t_1)(v^t - \varepsilon_1), \quad (110)$$

where the last inequality results from (109).

Recall that $Z^t = \sum_{i=1}^k W_i^t$ and $w_i^t = \frac{W_i^t}{Z^t}$. By taking $t_1 = t$, we can derive from (110) that

$$\left\langle \log \left(\frac{w^{t_2}}{w^t} \right) - \eta \Delta_{t, t_2}, w^t \right\rangle + \log \left(\frac{Z^{t_2}}{Z^t} \right) - \eta(t_2 - t) \text{OPT} \geq \eta(t_2 - t)(v^t - \varepsilon_1). \quad (111)$$

As it turns out, this inequality allows us to bound the KL divergence between w^t and w^{t_2} as follows:

$$\begin{aligned} \text{KL}(w^t \parallel w^{t_2}) &:= \left\langle w^t, \log \left(\frac{w^t}{w^{t_2}} \right) \right\rangle \\ &\leq \log(Z^{t_2}/Z^t) - \eta(t_2 - t) \text{OPT} - \eta \langle w^t, \Delta_{t, t_2} \rangle + \eta(t_2 - t)(\varepsilon_1 - v^t). \end{aligned} \quad (112)$$

In what follows, we shall cope with the right-hand side of (112).

Step 2: bounding the term $\log(Z^{t_2}/Z^t)$. We start with bounding $\log(Z^{t_2}/Z^t)$. With probability exceeding $1 - 2T^2 k \delta'$, it holds that

$$\begin{aligned} \log(Z^{t_2}/Z^t) &= \sum_{\tau=t}^{t_2-1} \log(Z^{\tau+1}/Z^\tau) = \sum_{\tau=t}^{t_2-1} \log \left(\sum_{i \in [k]} \frac{W_i^\tau \exp(\eta \hat{r}_i^\tau)}{\sum_{j \in [k]} W_j^\tau} \right) \\ &\stackrel{(i)}{=} \sum_{\tau=t}^{t_2-1} \log \left(\sum_{i=1}^k w_i^\tau \exp(\eta \hat{r}_i^\tau) \right) \stackrel{(ii)}{\leq} \sum_{\tau=t}^{t_2-1} \log \left(\sum_{i=1}^k w_i^\tau + \sum_{i=1}^k w_i^\tau (\eta \hat{r}_i^\tau) + 2 \sum_{i=1}^k w_i^\tau \eta^2 (\hat{r}_i^\tau)^2 \right) \\ &\stackrel{(iii)}{\leq} \sum_{\tau=t}^{t_2-1} \log \left(1 + \eta \sum_{i=1}^k w_i^\tau \hat{r}_i^\tau + 2\eta^2 \right) \leq \sum_{\tau=t}^{t_2-1} \left(\eta \sum_{i=1}^k w_i^\tau \hat{r}_i^\tau + 2\eta^2 \right) \end{aligned}$$

$$\begin{aligned}
& \stackrel{(iv)}{=} \eta \sum_{\tau=t}^{t_2-1} v^\tau + \eta(t_2-t)\text{OPT} + \eta \sum_{\tau=t}^{t_2-1} \sum_{i=1}^k \langle w_i^\tau, \hat{r}_i^\tau - v_{h^\tau, i} - \text{OPT} \rangle + 2(t_2-t)\eta^2 \\
& \stackrel{(v)}{\leq} \eta(t_2-t)\varepsilon_1 + \eta(t_2-t)\text{OPT} + \eta \sum_{\tau=t}^{t_2-1} \sum_{i=1}^k \langle w_i^\tau, \hat{r}_i^\tau - v_{h^\tau, i} - \text{OPT} \rangle + 2(t_2-t)\eta^2. \tag{113}
\end{aligned}$$

Here, (i) comes from line 5 of Algorithm 1, (ii) follows from the elementary inequality $\exp(x) \leq 1 + x + 2x^2$ for any $x \leq 1$, (iii) is valid since $\sum_i w_i^\tau = 1$ and $|\hat{r}_i^\tau| \leq 1$, (iv) holds due to the fact that $v^t = \langle w^t, v_{h^t} \rangle$, and (v) arises from the fact that $v^\tau \leq \varepsilon_1$ (see (61)).

Step 3: bounding the weighted sum of $\{\xi_i^\tau\}$. Next, we intend to control the two random terms below:

$$\eta \sum_{\tau=t}^{t_2-1} \sum_{i=1}^k \langle w_i^\tau, \hat{r}_i^\tau - v_{h^\tau, i} - \text{OPT} \rangle = \eta \sum_{\tau=t}^{t_2-1} \sum_{i=1}^k w_i^\tau \xi_i^\tau, \tag{114a}$$

$$\eta \langle w^t, \Delta_{t, t_2} \rangle = \eta \sum_{\tau=t}^{t_2-1} \sum_{i=1}^k w_i^t \xi_i^\tau. \tag{114b}$$

Let \mathcal{F}^τ denote what happens before the τ -th round in Algorithm 1. Two properties are worth noting.

- The variance of ξ_i^τ is at most $O\left(\frac{1}{k\bar{w}_i^\tau}\right)$, according to the update rule (see line 14 in Algorithm 1);
- $\{\xi_i^\tau\}_{i \in [k]}$ are independent conditioned on \mathcal{F}^τ .

Let us develop bounds on the two quantities in (114) below.

- Letting $q^\tau = \sum_{i=1}^k w_i^t \xi_i^\tau$, one sees that

$$|q^\tau| \leq 1, \quad \mathbb{E}[q^\tau | \mathcal{F}^\tau] = 0 \quad \text{and} \quad \text{Var}[q^\tau | \mathcal{F}^\tau] \leq \sum_{i=1}^k \frac{(w_i^t)^2}{k\bar{w}_i^\tau} \leq \sum_{i=1}^k \frac{w_i^t}{k} = \frac{1}{k}. \tag{115}$$

By virtue of Freedman's inequality (cf. Lemma 8), with probability at least $1 - \delta'$ one has

$$\left| \sum_{\tau=t}^{t_2-1} \sum_{i=1}^k w_i^t \xi_i^\tau \right| \leq 2\sqrt{\frac{t_2-t}{k} \log(2/\delta') + 2\log(2/\delta')}; \tag{116}$$

- Regarding the other term, by letting $\hat{q}^\tau = \sum_{i=1}^k w_i^\tau \xi_i^\tau$, we have

$$|\hat{q}^\tau| \leq 1, \quad \mathbb{E}[\hat{q}^\tau | \mathcal{F}^\tau] = 0 \quad \text{and} \quad \text{Var}[\hat{q}^\tau | \mathcal{F}^\tau] \leq \sum_{i=1}^k \frac{(w_i^\tau)^2}{k\bar{w}_i^\tau} \leq \sum_{i=1}^k \frac{w_i^\tau}{k} = \frac{1}{k}.$$

Invoke Freedman's inequality (cf. Lemma 8) once again to show that, with probability exceeding $1 - \delta'$,

$$\left| \sum_{\tau=t}^{t_2-1} \sum_{i=1}^k w_i^\tau \xi_i^\tau \right| \leq 2\sqrt{\frac{t_2-t}{k} \log(2/\delta') + 2\log(2/\delta')}. \tag{117}$$

Step 4: bounding the KL divergence between w^t and w^{t_2} . Combining (112), (113), (116) and (117), and applying the union bound over (t, t_2) , we can demonstrate that with probability at least $1 - 6T^4 k \delta'$,

$$\begin{aligned}
\text{KL}(w^t \| w^{t_2}) & \leq 2(t_2-t)\eta\varepsilon_1 - (t_2-t)\eta v^t \\
& \quad + 4\eta\sqrt{\frac{(t_2-t)\log(2/\delta')}{k}} + 2(t_2-t)\eta^2 + 4\eta\log(2/\delta') \tag{118}
\end{aligned}$$

holds for any $1 \leq t < t_2 \leq T$. The analysis below then operates under the condition that (118) holds for any $1 \leq t < t_2 \leq T$.

Step 5: connecting the KL divergence with the advertised properties. Set

$$\tau_{\tilde{j}} := \min\{\tau \mid t_1 \leq \tau \leq t_2 - 1, -v^\tau \leq 2^{-(\tilde{j}-1)}\}, \quad 1 \leq \hat{j} \leq j_{\max} := \lfloor \log_2(1/\eta) + 1 \rfloor; \quad (119a)$$

$$\tau_{j_{\max}+1} := t_2. \quad (119b)$$

By definition, we know that $\tau_1 = t_1$ and $\tau_{j_2} \geq \tau_{j_1}$ for $j_2 \geq j_1$. Let \mathcal{I} be the index set of this segment $[t_1, t_2]$, and set $y_j := \sum_{i \in \mathcal{I}} w_i^{\tau_j}$. We then claim that there exists $1 \leq \tilde{j} \leq j_{\max}$ such that

$$\log \left(\frac{y_{\tilde{j}+1}}{y_{\tilde{j}}} \right) \geq \frac{x}{\log_2(1/\eta) + 1} \quad (120)$$

$$\geq \frac{x}{\log_2(k) + 1}, \quad (121)$$

where the last inequality is valid since $1/\eta = 100/\varepsilon \leq k$ (given our assumption that $\varepsilon \gtrsim 1/k$).

Proof of (120). Suppose that none of $1 \leq \tilde{j} \leq j_{\max}$ satisfies (120). Then for any $1 \leq \hat{j} \leq j_{\max}$, it holds that $\log \left(\frac{y_{\hat{j}+1}}{y_{\hat{j}}} \right) \leq \frac{x}{\log_2(1/\eta) + 1}$, which implies that $y_{\hat{j}} \geq y_{\hat{j}+1} \exp \left(-\frac{x}{\log_2(1/\eta) + 1} \right)$. As a result, we have

$$y_1 \geq y_{j_{\max}+1} \cdot \exp \left(-j_{\max} \cdot \frac{x}{\log_2(1/\eta) + 1} \right) > p,$$

thus leading to contradiction (as according to the definition of the (p, q, x) -segment, one has $y_1 \leq p$). \square

Now, assume that \tilde{j} satisfies (120). From the definition of the (p, q, x) -segment, we have $y_{\tilde{j}} \geq q$. It follows from (118) that

$$\begin{aligned} \text{KL}(w^{\tau_{\tilde{j}}} \parallel w^{\tau_{\tilde{j}+1}}) &\leq 2(\tau_{\tilde{j}+1} - \tau_{\tilde{j}})\eta\varepsilon_1 + (\tau_{\tilde{j}+1} - \tau_{\tilde{j}})\eta 2^{-(\tilde{j}-1)} \\ &\quad + 4\eta \sqrt{\frac{(\tau_{\tilde{j}+1} - \tau_{\tilde{j}}) \log(2/\delta')}{k}} + 2(\tau_{\tilde{j}+1} - \tau_{\tilde{j}})\eta^2 + 4\eta \log(2/\delta'). \end{aligned} \quad (122)$$

Since $\log \left(\frac{y_{\tilde{j}+1}}{y_{\tilde{j}}} \right) \geq \frac{x}{\log_2(k) + 1}$ and $y_{\tilde{j}} \geq q$, we can invoke Lemma 11 and Lemma 12 to show that

$$\text{KL}(w^{\tau_{\tilde{j}}} \parallel w^{\tau_{\tilde{j}+1}}) \geq \text{KL}(\text{Ber}(y_{\tilde{j}}) \parallel \text{Ber}(y_{\tilde{j}+1})) \geq \frac{qx^2}{4(\log_2(k) + 1)^2},$$

where $\text{Ber}(x)$ denote the Bernoulli distribution with mean $x \in [0, 1]$. As a result, we can obtain

$$\begin{aligned} \frac{qx^2}{4(\log_2(k) + 1)^2} &\leq 2(\tau_{\tilde{j}+1} - \tau_{\tilde{j}})\eta\varepsilon_1 + (\tau_{\tilde{j}+1} - \tau_{\tilde{j}})\eta 2^{-(\tilde{j}-1)} \\ &\quad + 4\eta \sqrt{\frac{(\tau_{\tilde{j}+1} - \tau_{\tilde{j}}) \log(2/\delta')}{k}} + 2(\tau_{\tilde{j}+1} - \tau_{\tilde{j}})\eta^2 + 4\eta \log(2/\delta'), \end{aligned} \quad (123)$$

which in turn results in

$$\begin{aligned} \tau_{\tilde{j}+1} - \tau_{\tilde{j}} &\geq \min \left\{ \frac{qx^2}{100(\log_2(k) + 1)^2} \min \left\{ \frac{1}{\eta\varepsilon_1}, \frac{2^{\tilde{j}-1}}{\eta}, \frac{1}{\eta^2} \right\}, \frac{kq^2x^4}{10000\eta^2 \log(1/\delta')(\log_2(k) + 1)^4} \right\} \\ &= \frac{qx^2}{100(\log_2(k) + 1)^2} \min \left\{ \frac{2^{\tilde{j}-1}}{\eta}, \frac{1}{\eta^2} \right\} \end{aligned} \quad (124)$$

$$\geq \frac{qx^2 \cdot 2^{\tilde{j}-1}}{100(\log_2(k) + 1)^2 \eta}. \quad (125)$$

Here, to see why (124) holds, it suffices to note that

$$\frac{qx^2}{100(\log_2(k) + 1)^2} \cdot \frac{2^{\tilde{j}-1}}{\eta} \leq \frac{kq^2x^4}{10000\eta^2 \log(1/\delta')(\log_2(k) + 1)^4},$$

a property that arises from the fact that $2^{\tilde{j}-1} \leq 1/\eta = \frac{100}{\varepsilon} \leq k$ and the assumption that $\frac{qx^2}{50(\log_2(k)+1)^2} \geq 2\log(1/\delta')\eta$.

Additionally, we learn from the definition (119) that

$$\sum_{\tau=t_1}^{t_2-1} \mathbb{1}\{-v^\tau \geq 2^{-\hat{j}}\} \geq \sum_{\tau=t_1}^{t_2-1} \mathbb{1}\{-v^\tau > 2^{-\tilde{j}}\} \geq \tau_{\tilde{j}+1} - t_1 \geq \tau_{\tilde{j}+1} - \tau_{\tilde{j}} \quad \text{for any } 1 \leq \hat{j} \leq \log_2(k) + 1.$$

This taken collectively with (125) gives

$$\sum_{\tau=t_1}^{t_2-1} \mathbb{1}\{-v^\tau \geq 2^{-\tilde{j}}\} \geq \tau_{\tilde{j}+1} - \tau_{\tilde{j}} \geq \frac{qx^2 \cdot 2^{\tilde{j}-1}}{100(\log_2(k) + 1)^2 \eta}.$$

The proof is completed by noting that $\tau_{j_{\max}+1} = t_2$.

E Missing proofs for lower bounds

E.1 Restatement and proof of Lemma 4

Lemma 4. [Restatement] Fix $m \geq 0$ and $\tilde{i} \in [k]$. Suppose $\mathbb{P}_{\mathcal{G}'}[h_{\text{out}} = h^*, M_i \leq m] \geq \frac{1}{2}$. It then holds that for any $h \in \mathcal{H}_{\tilde{i}}$

$$\mathbb{P}_{\mathcal{G}'}\{h_{\text{out}} = h, M_{\tilde{i}} \leq m\} \geq \frac{1}{2} \mathbb{P}_{\mathcal{G}'}\{h_{\text{out}} = h^*, M_{\tilde{i}} \leq m\} \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2). \quad (126)$$

Moreover, it holds that $m \geq \frac{\log(N_0/4)}{30000\varepsilon^2}$.

Proof. For $v \in \{-1, 1\}^m$, we let

$$n^+(v) = |\{p : v_p = 1\}|$$

denote the number of 1's in the coordinates of v . Let \mathcal{V} be a subset of $\{-1, 1\}^{2m}$ defined as $\mathcal{V} := \{v^1, v^2 \in \{-1, 1\}^m \mid n^+(v^1) - n^+(v^2) \leq 4\sqrt{m} + 2m\varepsilon\}$. Let $\mathbb{P}_{\mathcal{C}}[\cdot]$ denote the probability distribution of $\{\{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m\}$ and $\mathbb{P}_{\mathcal{C}'}[\cdot]$ denote the probability distribution of $\{\{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m, \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m\}$. Recall the definition of $\{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m$ and $\{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m$. By Hoeffding's inequality, we have that $\mathbb{P}_{\mathcal{C}}[\mathcal{V}] \geq \frac{3}{4}$. Also noting that the distribution of \mathcal{V} is independent of the algorithm \mathcal{G}' , we have that

$$\mathbb{P}_{\mathcal{G}'}\left[h_{\text{out}} = h^*, M_{\tilde{i}} \leq m, \left\{\{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m\right\} \in \mathcal{V}\right] \geq \frac{3}{4} - \left(1 - \frac{3}{4}\right) = \frac{1}{2}. \quad (127)$$

That is,

$$\sum_{v=\{v^1, v^2\} \in \mathcal{V}} \mathbb{P}_{\mathcal{G}'}\left[h_{\text{out}} = h^*, M_{\tilde{i}} \leq m, \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^1, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^2\right] \geq \frac{1}{2}.$$

In addition, for any $v = \{v^1, v^2\} \in \mathcal{V}$, we have

$$\begin{aligned} \mathbb{P}_{\mathcal{C}'}[v] &= \mathbb{P}_{\mathcal{C}}[v] \cdot (1 - 8\varepsilon)^{n^+(v^1) - n^+(v^2)} (1 + 8\varepsilon)^{n^+(v^2) - n^+(v^1)} \\ &= \mathbb{P}_{\mathcal{C}}[v] \left(\frac{1 - 8\varepsilon}{1 + 8\varepsilon}\right)^{n^+(v^1) - n^+(v^2)} \end{aligned}$$

$$\begin{aligned}
&\geq \mathbb{P}_C[v] \exp(-20(n^+(v^1) - n^+(v^2))\varepsilon) \\
&\geq \mathbb{P}_C[v] \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2).
\end{aligned}$$

As a result, we can demonstrate that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h, M_{\tilde{i}} \leq m, \{x_{j_{h^*, \tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^1, \{x_{j_{h, \tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^2 \right] \\
&= \mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h, M_{\tilde{i}} \leq m | \{x_{j_{h^*, \tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^1, \{x_{j_{h, \tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^2 \right] \mathbb{P}_{C'}[v] \\
&\geq \mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h, M_{\tilde{i}} \leq m | \{x_{j_{h^*, \tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^1, \{x_{j_{h, \tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^2 \right] \mathbb{P}_C[v] \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2) \\
&= \mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h^*, M_{\tilde{i}} \leq m | \{x_{j_{h, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^1, \{x_{j_{h^*, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^2 \right] \mathbb{P}_C[v] \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2) \quad (128) \\
&= \mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h^*, M_{\tilde{i}} \leq m, \{x_{j_{h, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^1, \{x_{j_{h^*, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^2 \right] \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2).
\end{aligned}$$

Here, (128) results from Lemma 18. We present the detailed computation as follows. Let v_ℓ^o be the ℓ -th coordinate of v^1 for $1 \leq \ell \leq m$ and $o = 1, 2$.

$$\begin{aligned}
&\mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h, M_{\tilde{i}} \leq m | \{x_{j_{h^*, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^1, \{x_{j_{h, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^2 \right] \\
&= \sum_{m'=1}^m \mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h, M_{\tilde{i}} = m' | \{x_{j_{h^*, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^1, \{x_{j_{h, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^2 \right] \\
&= \sum_{m'=1}^m \mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h, M_{\tilde{i}} = m' | \{x_{j_{h^*, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^{m'} = \{v_\ell^1\}_{\ell=1}^{m'}, \{x_{j_{h, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^{m'} = \{v_\ell^2\}_{\ell=1}^{m'} \right] \quad (129)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{m'=1}^m \mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h^*, M_{\tilde{i}} = m' | \{x_{j_{h, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^{m'} = \{v_\ell^1\}_{\ell=1}^{m'}, \{x_{j_{h^*, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^{m'} = \{v_\ell^2\}_{\ell=1}^{m'} \right] \\
&= \sum_{m'=1}^m \mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h^*, M_{\tilde{i}} = m' | \{x_{j_{h, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^1, \{x_{j_{h^*, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^2 \right] \quad (130) \\
&= \mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h^*, M_{\tilde{i}} \leq m | \{x_{j_{h, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^1, \{x_{j_{h^*, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^2 \right].
\end{aligned}$$

Here, (129) and (130) hold since for $h' = h, h^*$, the event $\{h_{\text{out}} = h', M_{\tilde{i}} = m'\}$ is independent of $\{x^\ell(\tilde{i})\}_{\ell \geq m'+1}$. Taking the sum over $v = \{v^1, v^2\} \in \mathcal{V}$, we obtain

$$\begin{aligned}
&\mathbb{P}_{\mathcal{G}'}[h_{\text{out}} = h, M_{\tilde{i}} \leq m] \\
&\geq \sum_{v \in \mathcal{V}} \mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h, M_{\tilde{i}} \leq m, \{x_{j_{h^*, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^1, \{x_{j_{h, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^2 \right] \\
&\geq \sum_{v \in \mathcal{V}} \mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h^*, M_{\tilde{i}} \leq m, \{x_{j_{h, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^1, \{x_{j_{h^*, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m = v^2 \right] \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2) \\
&= \mathbb{P}_{\mathcal{G}'} \left[h_{\text{out}} = h^*, M_{\tilde{i}} \leq m, \left\{ \{x_{j_{h, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m, \{x_{j_{h^*, \tilde{i}}}^\ell(\tilde{i})\}_{\ell=1}^m \right\} \in \mathcal{V} \right] \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2) \\
&\geq \frac{1}{2} \mathbb{P}_{\mathcal{G}'}[h_{\text{out}} = h^*, M_{\tilde{i}} \leq m] \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2)
\end{aligned}$$

as claimed in (126).

Summing over all $h \in \mathcal{H}_{\tilde{i}}$, we reach

$$1 \geq \mathbb{P}_{\mathcal{G}'}[h_{\text{out}} \in \mathcal{H}_{\tilde{i}}, M_{\tilde{i}} \leq m] \geq \frac{N_0}{2} \mathbb{P}_{\mathcal{G}'}[h_{\text{out}} = h^*, M_{\tilde{i}} \leq m] \exp(-40m\varepsilon^2 - 80\sqrt{m}\varepsilon), \quad (131)$$

which reveals that

$$\frac{1}{2} \leq \mathbb{P}_{\mathcal{G}'}[h_{\text{out}} = h^*, M_{\tilde{i}} \leq m] \leq \frac{2}{N_0} \exp(40m\varepsilon^2 + 80\sqrt{m}\varepsilon). \quad (132)$$

Consequently, it is seen that

$$40m\varepsilon^2 + 80\sqrt{m\varepsilon} \geq \log(N_0/4),$$

which implies that $m \geq \min\{\frac{\log(N_0/4)}{80\varepsilon^2}, \frac{\log^2(N_0/4)}{30000\varepsilon^2}\} \geq \frac{\log(N_0/4)}{30000\varepsilon^2}$ \square

E.2 Statement and proof of Lemma 18

Lemma 18. Fix $\tilde{i} \in [k]$, and let $x^l(i)$ be the l -th sample from D_i for any $i \in [k]$ and $l \geq 1$. Fix $h \in \mathcal{H}_{\tilde{i}}$, $m > 0$ and $v^1, v^2 \in \{-1, 1\}^{2m}$.

$$\begin{aligned} & \mathbb{P}_{\mathcal{G}'} \left\{ h_{\text{out}} = h^*, M_i = m \mid \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^1, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^2 \right\} \\ &= \mathbb{P}_{\mathcal{G}'} \left\{ h_{\text{out}} = h, M_i = m \mid \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^2, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^1 \right\}. \end{aligned}$$

Proof. Let $\bar{\sigma}$ be the permutation over \mathcal{H} with $\bar{\sigma}(h^*) = h, \bar{\sigma}(h) = h^*$ and $\bar{\sigma}(h') = h'$ for all $h' \notin \{h, h^*\}$. It then holds that $\bar{\sigma}^{-1} = \bar{\sigma}$.

Fix $\{m_i\}_{i=1}^k$ and $X(i) = \{X^\ell(i)\}_{\ell=1}^{m_i} \in \{-1, 0, 1\}^{kNm_i}$ for $i \in [k]$. Let $x(i) = \{x^l(i)\}_{l=1}^{m_i}$ be the datapoints of the first m_i calls to $\text{Query}(D_i)$. With a slight abuse of notations, we let $\sigma(x(i)) = \{\sigma(x^l(i))\}_{l=1}^{m_i}$ for each $i \in [k]$.

It follows from Lemma 19 that

$$\begin{aligned} & \mathbb{P}_{\mathcal{G}, \mathcal{H}} [h_{\text{out}} = h, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k \mid x(i) = X(i), \forall i \in [k]] \\ &= \mathbb{P}_{\mathcal{G}, \sigma(\mathcal{H})} [h_{\text{out}} = \sigma^{-1}(h), \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k \mid \sigma^{-1}(x(i)) = X(i), \forall i \in [k]], \end{aligned}$$

which implies that

$$\begin{aligned} & \mathbb{P}_{\mathcal{G}, \sigma(\mathcal{H})} [h_{\text{out}} = h^*, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k, \sigma(x(i)) = X(i), \forall i \in [k]] \\ &= \mathbb{P}_{\mathcal{G}, \bar{\sigma}\sigma(\mathcal{H})} [h_{\text{out}} = h, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k, \bar{\sigma}\sigma(x(i)) = X(i), \forall i \in [k]] \cdot \frac{\mathbb{P}[\sigma(x(i)) = X(i), \forall i \in [k]]}{\mathbb{P}[\bar{\sigma}\sigma(x(i)) = X(i), \forall i \in [k]]} \\ &= \mathbb{P}_{\mathcal{G}, \bar{\sigma}\sigma(\mathcal{H})} [h_{\text{out}} = h, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k, \bar{\sigma}\sigma(x(i)) = X(i), \forall i \in [k]] \cdot \frac{\mathbb{P}[\sigma(x(\tilde{i})) = X(\tilde{i})]}{\mathbb{P}[\bar{\sigma}\sigma(x(\tilde{i})) = X(\tilde{i})]} \\ &= \mathbb{P}_{\mathcal{G}, \bar{\sigma}\sigma(\mathcal{H})} [h_{\text{out}} = h, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k, \bar{\sigma}\sigma(x(i)) = X(i), \forall i \in [k]] \\ & \quad \cdot \frac{\mathbb{P}[\{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} = \{X_{j_{\sigma(h),\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}}, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} = \{X_{j_{\sigma(h^*),\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}}]}{\mathbb{P}[\{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} = \{X_{j_{\sigma(h),\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}}, \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} = \{X_{j_{\sigma(h^*),\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}}]}. \end{aligned} \tag{133}$$

Re-arrange the equation to arrive at

$$\begin{aligned} & \mathbb{P}_{\mathcal{G}, \sigma(\mathcal{H})} \left[h_{\text{out}} = h^*, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k, \sigma(x(i)) = X(i), \forall i \in [k] \right. \\ & \quad \left. \mid \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} = \{X_{j_{\sigma(h),\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}}, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} = \{X_{j_{\sigma(h^*),\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} \right] \\ &= \mathbb{P}_{\mathcal{G}, \bar{\sigma}\sigma(\mathcal{H})} \left[h_{\text{out}} = h, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k, \bar{\sigma}\sigma(x(i)) = X(i), \forall i \in [k] \right. \\ & \quad \left. \mid \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} = \{X_{j_{\sigma(h),\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}}, \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} = \{X_{j_{\sigma(h^*),\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} \right]. \end{aligned}$$

Taking the sum over all possible choices of $\{X(i)\}_{i \neq \tilde{i}}, \left\{\{X_{j_{h',i'}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}}\right\}_{i' \in [k], h' \notin \{h, h^*\}}, \left\{\{X_{j_{h',i'}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}}\right\}_{h' \in \{h, h^*\}, i' \neq \tilde{i}}$ and $\{m_i\}_{i \neq \tilde{i}}$, we obtain

$$\begin{aligned} & \mathbb{P}_{\mathcal{G}, \sigma(\mathcal{H})} \left[h_{\text{out}} = h^*, M_{\tilde{i}} = m_{\tilde{i}} \right. \\ & \quad \left. \left| \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} = \{X_{j_{\sigma(h),\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}}, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} = \{X_{j_{\sigma(h^*),\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} \right] \right. \\ &= \mathbb{P}_{\mathcal{G}, \bar{\sigma}(\mathcal{H})} \left[h_{\text{out}} = h, M_{\tilde{i}} = m_{\tilde{i}} \right. \\ & \quad \left. \left| \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} = \{X_{j_{\sigma(h),\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}}, \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} = \{X_{j_{\sigma(h^*),\tilde{i}}}^l(\tilde{i})\}_{l=1}^{m_{\tilde{i}}} \right] \right] \end{aligned}$$

for any $X(\tilde{i}) \in \{-1, 0, 1\}^{kN m_{\tilde{i}}}$.

Fix $m_{\tilde{i}} = m$, and choose $\{X_{j_{\sigma(h),\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v_1, \{X_{j_{\sigma(h^*),\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v_2$. We then have

$$\begin{aligned} & \mathbb{P}_{\mathcal{G}', \mathcal{H}} \left[h_{\text{out}} = h^*, M_i = m \mid \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^1, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^2 \right] \\ &= \frac{1}{|\Pi_{\mathcal{H}}|} \sum_{\sigma \in \Pi_{\mathcal{H}}} \mathbb{P}_{\mathcal{G}, \sigma(\mathcal{H})} \left[h_{\text{out}} = h^*, M_i = m \mid \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^1, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^2 \right] \\ &= \frac{1}{|\Pi_{\mathcal{H}}|} \sum_{\sigma \in \Pi_{\mathcal{H}}} \mathbb{P}_{\mathcal{G}, \bar{\sigma}(\mathcal{H})} \left[h_{\text{out}} = \bar{\sigma}^{-1}(h^*), M_i = m \mid \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^2, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^1 \right] \\ &= \frac{1}{|\Pi_{\mathcal{H}}|} \sum_{\sigma \in \Pi_{\mathcal{H}}} \mathbb{P}_{\mathcal{G}, \sigma(\mathcal{H})} \left[h_{\text{out}} = h, M_i = m \mid \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^2, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^1 \right] \\ &= \mathbb{P}_{\mathcal{G}', \mathcal{H}} \left[h_{\text{out}} = h, M_i = m \mid \{x_{j_{h,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^2, \{x_{j_{h^*,\tilde{i}}}^l(\tilde{i})\}_{l=1}^m = v^1 \right]. \end{aligned}$$

The proof is thus completed. \square

E.3 Statement and proof of Lemma 19

Lemma 19. Fix $\{m_i\}_{i \in [k]}$, $\sigma \in \Pi_{\mathcal{H}}$ and $X \in \{-1, 0, 1\}^{kN \sum_{i=1}^k m_i}$. Let $\{X(i)\}_{i \in [k]}$ and $\{x(i)\}$ be defined as Lemma 18. We then have

$$\begin{aligned} & \mathbb{P}_{\mathcal{G}, \mathcal{H}} \left[h_{\text{out}} = h, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k \mid x(i) = X(i), \forall i \in [k] \right] \\ &= \mathbb{P}_{\mathcal{G}, \sigma(\mathcal{H})} \left[h_{\text{out}} = \sigma^{-1}(h), \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k \mid \sigma^{-1}(x(i)) = X(i), \forall i \in [k] \right]. \end{aligned} \quad (134)$$

Proof. Let $\mathcal{H}' = \sigma(\mathcal{H})$. Let $h_p(\cdot)$ denote the p -th hypothesis in the hypothesis set. Then one has

$$\begin{aligned} & \mathbb{P}_{\mathcal{G}, \mathcal{H}} \left[h_{\text{out}} = h_p(\mathcal{H}), \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k \mid x(i) = X(i), \forall i \in [k] \right] \\ &= \mathbb{P}_{\mathcal{G}, \mathcal{H}} \left[h_{\text{out}} = h_p(\mathcal{H}), \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k \mid \left\{ \{x_{j_{h_{p'}(\mathcal{H}), i}}^l(i')\}_{p'=1, i=1}^{|\mathcal{H}|, k} \}_{l=1}^{m_{i'}} \right\}_{i'=1}^k = X \right] \\ &= \mathbb{P}_{\mathcal{G}, \mathcal{H}'} \left[h_{\text{out}} = h_p(\mathcal{H}'), \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k \mid \left\{ \{x_{j_{h_{p'}(\mathcal{H}'), i}}^l(i')\}_{p'=1, i=1}^{|\mathcal{H}'|, k} \}_{l=1}^{m_{i'}} \right\}_{i'=1}^k = X \right] \\ &= \mathbb{P}_{\mathcal{G}, \sigma(\mathcal{H})} \left[h_{\text{out}} = h_p(\sigma(\mathcal{H})), \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k \mid \sigma^{-1}(x(i)) = X(i), \forall i \in [k] \right] \\ &= \mathbb{P}_{\mathcal{G}, \sigma(\mathcal{H})} \left[h_{\text{out}} = \sigma^{-1}(h_p(\mathcal{H})), \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k \mid \sigma^{-1}(x(i)) = X(i), \forall i \in [k] \right]. \end{aligned} \quad (135)$$

$$\begin{aligned} &= \mathbb{P}_{\mathcal{G}, \sigma(\mathcal{H})} \left[h_{\text{out}} = h_p(\sigma(\mathcal{H})), \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k \mid \sigma^{-1}(x(i)) = X(i), \forall i \in [k] \right] \\ &= \mathbb{P}_{\mathcal{G}, \sigma(\mathcal{H})} \left[h_{\text{out}} = \sigma^{-1}(h_p(\mathcal{H})), \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k \mid \sigma^{-1}(x(i)) = X(i), \forall i \in [k] \right]. \end{aligned} \quad (136)$$

Here, (135) holds since the algorithm \mathcal{G} cannot distinguish \mathcal{H} from \mathcal{H}' using its own randomness. \square

F Missing proofs for Rademacher classes

F.1 Restatement and proof of Lemma 7

Lemma 7. [Restatement] *By running Algorithm 3, with probability at least $1 - \delta/4$, it holds that*

$$L(h^t, w^t) \leq \min_{h \in \mathcal{H}} L(h, w^t) + \varepsilon_1 \quad (137)$$

for any $1 \leq t \leq T$.

Proof. We will follow the notation in the proof of Lemma 1. Fix $n = \{n_i\}_{i=1}^k$ such that $n_i \geq 12 \log(2k)$ for all $i \in [k]$ and $w \in \Delta(k)$. Let $\kappa = \min_i \frac{n_i}{w_i}$. Recall that $(x_{i,j}, y_{i,j})$ is the j -th sample from D_i .

Define

$$F(n, w) := \mathbb{E}_{\{(x_{i,j}, y_{i,j})\}_{j=1}^{n_i}\}_{i=1}^k \left[\max_{h \in \mathcal{H}} \left(\sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} \ell(h, (x_{i,j}, y_{i,j})) - \sum_{i=1}^k w_i L(h, e_i) \right) \right].$$

Let $\{(\tilde{x}_{i,j}, \tilde{y}_{i,j})\}_{j=1}^{n_i}\}_{i=1}^k$ be a group of ghost samples for $\{(x_{i,j}, y_{i,j})\}_{j=1}^{n_i}\}_{i=1}^k$. We then have

$$\begin{aligned} F(n, w) &\leq \mathbb{E}_{\{(x_{i,j}, y_{i,j})\}_{j=1}^{n_i}\}_{i=1}^k, \{(\tilde{x}_{i,j}, \tilde{y}_{i,j})\}_{j=1}^{n_i}\}_{i=1}^k \left[\max_{h \in \mathcal{H}} \left(\sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} (\ell(h, (x_{i,j}, y_{i,j})) - \ell(h, (\tilde{x}_{i,j}, \tilde{y}_{i,j}))) \right) \right] \\ &= \mathbb{E}_{\{(x_{i,j}, y_{i,j})\}_{j=1}^{n_i}\}_{i=1}^k, \{(\tilde{x}_{i,j}, \tilde{y}_{i,j})\}_{j=1}^{n_i}\}_{i=1}^k, \{\sigma_i^j\}_{j=1}^{n_i}\}_{i=1}^k \left[\max_{h \in \mathcal{H}} \left(\sum_{i=1}^k \frac{w_i}{n_i} \sum_{j=1}^{n_i} \sigma_i^j (\ell(h, (x_{i,j}, y_{i,j})) - \ell(h, (\tilde{x}_{i,j}, \tilde{y}_{i,j}))) \right) \right] \\ &\leq 2 \mathbb{E}_{\{(x_{i,j}, y_{i,j})\}_{j=1}^{n_i}\}_{i=1}^k, \{\sigma_i^j\}_{j=1}^{n_i}\}_{i=1}^k \left[\max_{h \in \mathcal{H}} \left(\sum_{i=1}^k \frac{1}{\kappa} \sum_{j=1}^{n_i} \sigma_i^j \ell(h, (x_{i,j}, y_{i,j})) \right) \right] \quad (138) \end{aligned}$$

$$\leq 2 \frac{\sum_{i=1}^k n_i}{\kappa} \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k}. \quad (139)$$

Here, (138) follows from Lemma 20, as stated below.

Lemma 20. *Let \mathcal{L} be a subset of \mathbb{R}^n . Let $w^1, w^2 \in \mathbb{R}^n$ be such that $|w_i^1| \leq |w_i^2|$ for all $i \in [n]$. Then it holds*

$$\mathbb{E}_{\{\sigma_i\}_{i=1}^n \sim \text{iid} \{\pm 1\}} \left[\max_{f \in \mathcal{L}} \sum_{i=1}^n \sigma_i w_i^1 f_i \right] \leq \mathbb{E}_{\{\sigma_i\}_{i=1}^n \sim \text{iid} \{\pm 1\}} \left[\max_{f \in \mathcal{L}} \sum_{i=1}^n \sigma_i w_i^2 f_i \right]. \quad (140)$$

Proof. It suffices to prove (140) under the case where $w_i^1 = w_i^2$ for $1 \leq i \leq n-1$, and $|w_n^1| \leq |w_n^2|$. Fix σ_i for $1 \leq i \leq n-1$.

$$\begin{aligned} &\mathbb{E}_{\sigma_n \sim \{\pm 1\}} \left[\max_{f \in \mathcal{L}} \sum_{i=1}^n \sigma_i w_i^1 f_i \right] \\ &= \frac{1}{2} \max_{f \in \mathcal{L}} \left(\sum_{i=1}^{n-1} \sigma_i w_i^1 f_i + w_n^1 f_n \right) + \frac{1}{2} \max_{f \in \mathcal{L}} \left(\sum_{i=1}^{n-1} \sigma_i w_i^1 f_i - w_n^1 f_n \right) \\ &= \frac{1}{2} \max_{f^1, f^2 \in \mathcal{L}} \left(\sum_{i=1}^{n-1} \sigma_i w_i^1 (f_i^1 + f_i^2) + w_n^1 (f_n^1 - f_n^2) \right) \\ &= \frac{1}{2} \max_{f^1, f^2 \in \mathcal{L}} \left(\sum_{i=1}^{n-1} \sigma_i w_i^2 (f_i^1 + f_i^2) + w_n^1 (f_n^1 - f_n^2) \right) \\ &\leq \frac{1}{2} \max_{f^1, f^2 \in \mathcal{L}} \left(\sum_{i=1}^{n-1} \sigma_i w_i^2 (f_i^1 + f_i^2) + |w_n^2| (f_n^1 - f_n^2) \right) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \max_{f^1, f^2 \in \mathcal{L}} \left(\sum_{i=1}^{n-1} \sigma_i w_i^2 (f_i^1 + f_i^2) + w_n^2 (f_n^1 - f_n^2) \right) \\
&= \frac{1}{2} \max_{f \in \mathcal{L}} \left(\sum_{i=1}^{n-1} \sigma_i w_i^2 f_i + w_n^2 f_n \right) + \frac{1}{2} \max_{f \in \mathcal{L}} \left(\sum_{i=1}^{n-1} \sigma_i w_i^2 f_i - w_n^2 f_n \right) \\
&= \mathbb{E}_{\sigma_n \sim \{\pm 1\}} \left[\max_{f \in \mathcal{L}} \sum_{i=1}^n \sigma_i w_i^2 f_i \right].
\end{aligned}$$

The proof is thus completed by taking expectation over $\{\sigma_i\}_{i=1}^{n-1}$. \square

By virtue of Lemma 10 with the choice $c = 1/\kappa$, we obtain

$$\begin{aligned}
&\mathbb{P}_{\{(x_{i,j}, y_{i,j})\}_{j=1}^{n_i}\}_{i=1}^k} \left[\left| \max_{h \in \mathcal{H}} \left(\sum_{i=1}^k \frac{w_i}{n_i} \ell(h, (x_{i,j}, y_{i,j})) \right) - \sum_{i=1}^k w_i L(h, e_i) \right| \geq \varepsilon \right] \\
&\leq 2 \exp \left(- \frac{2\kappa^2 \varepsilon^2}{\sum_{i=1}^k n_i} \right). \tag{141}
\end{aligned}$$

According to (141), for any $\delta' \in (0, 1]$, with probability at least $1 - \delta'$, it holds that

$$\max_{h \in \mathcal{H}} \left(\sum_{i=1}^k \frac{w_i}{n_i} \ell(h, (x_{i,j}, y_{i,j})) - \sum_{i=1}^k w_i L(h, e_i) \right) \tag{142}$$

$$\leq 2 \frac{\sum_{i=1}^k n_i}{\kappa} \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} + \frac{\sum_{i=1}^k n_i}{\kappa} \sqrt{\frac{\log(2/\delta')}{2 \sum_{i=1}^k n_i}}. \tag{143}$$

Replacing (ℓ, L) with $(-\ell, -L)$ and using similar arguments, we can show that, with probability at least $1 - 2\delta'$,

$$\begin{aligned}
&\max_{h \in \mathcal{H}} \left| \sum_{i=1}^k \frac{w_i}{n_i} \ell(h, (x_{i,j}, y_{i,j})) - \sum_{i=1}^k w_i L(h, e_i) \right| \\
&\leq 2 \frac{\sum_{i=1}^k n_i}{\kappa} \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} + \frac{\sum_{i=1}^k n_i}{\kappa} \sqrt{\frac{\log(2/\delta')}{2 \sum_{i=1}^k n_i}}. \tag{144}
\end{aligned}$$

Now we fix $\kappa \geq 0$. Define that

$$\tilde{\mathcal{L}} = \left\{ n = \{n_i\}_{i=1}^k, w = \{w_i\}_{i=1}^k \in \Delta_{\varepsilon_1/(8k)}(k) \mid T_1 w_i \leq 2n_i, 12 \log(2k) \leq n_i \leq T_1, \forall i \in [k], \sum_{i=1}^k n_i \leq 2T_1 \right\}.$$

By (144), for any $\delta' > 0$, with probability at least $1 - \delta'$, for any $\{n, w\} \in \tilde{\mathcal{L}}$, we have that

$$\begin{aligned}
&\max_{h \in \mathcal{H}} \left| \sum_{i=1}^k \frac{w_i}{n_i} \ell(h, (x_{i,j}, y_{i,j})) - \sum_{i=1}^k w_i L(h, e_i) \right| \\
&\leq 2 \frac{\sum_{i=1}^k n_i}{T_1} \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} + \frac{\sum_{i=1}^k n_i}{T_1} \sqrt{\frac{\log(|\tilde{\mathcal{L}}|) + \log(2/\delta')}{2 \sum_{i=1}^k n_i}} \\
&\leq 4 \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} + 4 \sqrt{\frac{\log(2|\tilde{\mathcal{L}}|) + \log(2/\delta')}{T_1}} \\
&\leq 4 \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} + 4 \sqrt{\frac{2k \log(16kT_1/\varepsilon_1) + \log(2/\delta')}{T_1}}
\end{aligned}$$

$$\leq 600C_{T_1} + 4\sqrt{\frac{2k \log(16kT_1/\varepsilon_1) + \log(2/\delta')}{T_1}}.$$

Here, the last inequality follows from Assumption 1, Lemma 5, Lemma 6 and the fact that $\sum_{i=1}^k n_i \geq \frac{T_1}{2}$.

Note that in Algorithm 3,

$$\hat{L}^t(h, w^t) = \sum_{i=1}^k \frac{w_i^t}{\tilde{n}_i^t} \cdot \sum_{j=1}^{\tilde{n}_i^t} \ell(h, (x_{i,j}, y_{i,j})).$$

By the definition that $\tilde{n}_i^t = \min\{\lceil T_1 w_i^t + 12 \log(2k) \rceil, T_1\}$ for $i \in [k]$, we have that $T_1 w_i^t \leq \tilde{n}_i^t - 1$ and $12 \log(2k) \leq \tilde{n}_i^t \leq T_1$ for all $i \in [k]$. In addition,

$$\sum_{i=1}^k \tilde{n}_i^t \leq T_1 + k + 12k \log(2k) \leq 2T_1 - 2.$$

Therefore, there exists some $\tilde{w}^t \in \Delta(k)$ such that $\{\{\tilde{n}_i^t\}_{i=1}^k, \tilde{w}^t\} \in \tilde{\mathcal{L}}$ and $\|w^t - \tilde{w}^t\|_1 \leq \frac{\varepsilon_1}{8k}$ for each $1 \leq t \leq T$. Choose $\delta' = \delta/4$. We then obtain that: with probability at least $1 - \delta/4$, it holds that

$$\max_{h \in \mathcal{H}} |\hat{L}^t(h, w^t) - L(h, w^t)| \leq 600C_{T_1} \leq \frac{\varepsilon_1}{2}$$

for any $1 \leq t \leq T$. Here the last inequality is by definition of T_1 .

Finally, the fact that $h^t = \arg \min_{h \in \mathcal{H}} \hat{L}^t(h, w^t)$ allows one to derive

$$L(h^t, w^t) \leq \hat{L}^t(h^t, w^t) + \frac{\varepsilon_1}{2} = \min_{h \in \mathcal{H}} \hat{L}^t(h, w^t) + \frac{\varepsilon_1}{2} \leq \min_{h \in \mathcal{H}} L(h, w^t) + \varepsilon_1,$$

which concludes the proof. \square

F.2 Restatement and proof of Lemma 5

Lemma 5.[Restatement] *For any two groups of positive integers $\{n_i\}_{i=1}^k$ and $\{m_i\}_{i=1}^k$, it holds that*

$$\begin{aligned} \left(\sum_{i=1}^k n_i \right) \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} &\leq \left(\sum_{i=1}^k (m_i + n_i) \right) \widetilde{\text{Rad}}_{\{m_i + n_i\}_{i=1}^k} \\ &\leq \left(\sum_{i=1}^k n_i \right) \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} + \left(\sum_{i=1}^k m_i \right) \widetilde{\text{Rad}}_{\{m_i\}_{i=1}^k}. \end{aligned} \quad (145)$$

Proof. In what follows, assume that each z_i^j obeys $z_i^j \sim \mathcal{D}_i$, and each σ_i^j is a zero-mean Rademacher random variable. Direct computation then gives

$$\begin{aligned} &\left(\sum_{i=1}^k n_i \right) \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} \\ &= \mathbb{E}_{\{z_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\mathbb{E}_{\{\sigma_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\max_{h \in \mathcal{H}} \sum_{i=1}^k \sum_{j=1}^{n_i} \sigma_i^j \ell(h, z_i^j) \right] \right] \\ &\stackrel{(i)}{=} \mathbb{E}_{\{z_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\mathbb{E}_{\{\sigma_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\max_{h \in \mathcal{H}} \mathbb{E}_{\{z_i^j\}_{j=n_i+1}^{n_i+m_i}, \{\sigma_i^j\}_{j=n_i+1}^{n_i+m_i}, \forall i \in [k]} \left[\sum_{i=1}^k \sum_{j=1}^{n_i+m_i} \sigma_i^j \ell(h, z_i^j) \right] \right] \right] \\ &\stackrel{(ii)}{\leq} \mathbb{E}_{\{z_i^j\}_{j=1}^{n_i+m_i}, \forall i \in [k]} \left[\mathbb{E}_{\{\sigma_i^j\}_{j=1}^{n_i+m_i}, \forall i \in [k]} \left[\max_{h \in \mathcal{H}} \sum_{i=1}^k \sum_{j=1}^{n_i+m_i} \sigma_i^j \ell(h, z_i^j) \right] \right] \end{aligned}$$

$$\begin{aligned}
&= \left(\sum_{i=1}^k (n_i + m_i) \right) \widetilde{\text{Rad}}_{\{n_i + m_i\}_{i=1}^k} \\
&\stackrel{\text{(iii)}}{\leq} \mathbb{E}_{\{z_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\mathbb{E}_{\{\sigma_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\max_{h \in \mathcal{H}} \sum_{i=1}^k \sum_{j=1}^{n_i} \sigma_i^j \ell(h, z_i^j) \right] \right] \\
&\quad + \mathbb{E}_{\{z_i^j\}_{j=n_i+1}^{n_i+m_i}, \forall i \in [k]} \left[\mathbb{E}_{\{\sigma_i^j\}_{j=n_i+1}^{n_i+m_i}, \forall i \in [k]} \left[\max_{h \in \mathcal{H}} \sum_{i=1}^k \sum_{j=n_i+1}^{n_i+m_i} \sigma_i^j \ell(h, z_i^j) \right] \right] \\
&= \left(\sum_{i=1}^k n_i \right) \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} + \left(\sum_{i=1}^k m_i \right) \widetilde{\text{Rad}}_{\{m_i\}_{i=1}^k}.
\end{aligned}$$

Here, (i) is valid due to the zero-mean property of $\{\sigma_i^j\}$, (ii) comes from Jensen's inequality, and (iii) holds since $\max_x (f_1(x) + f_2(x)) \leq \max_x f_1(x) + \max_x f_2(x)$. \square

F.2.1 Restatement and proof of Lemma 6

Lemma 6.[Restatement] *Consider any $\{n_i\}_{i=1}^k$ obeying $n_i \geq 12 \log(2k)$ for each $i \in [k]$. By taking $w \in \Delta^k$ with $w_i = \frac{n_i}{\sum_{l=1}^k n_l}$, one has*

$$\widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k} \leq 72 \text{Rad}_{\sum_{i=1}^k n_i}(\mathcal{D}(w)).$$

Proof. Set $\tilde{n} = \sum_{i=1}^k n_i$. Let $\{X_j\}_{j=1}^{\tilde{n}}$ be n i.i.d. multinomial random variables with parameter $\{w_i\}_{i=1}^k$, and take $\hat{n}_i = \sum_{j=1}^{\tilde{n}} \mathbb{1}\{X_j = i\}$ for each $i \in [k]$. From (38) and Definition 3, it is easily seen that

$$\text{Rad}_{\tilde{n}}(\mathcal{D}(w)) = \mathbb{E}_{\{X_i\}_{i=1}^{\tilde{n}}} \left[\mathbb{E}_{\{z_i^j\}_{j=1}^{\hat{n}_i}, \forall i \in [k]} \left[\mathbb{E}_{\{\sigma_i^j\}_{j=1}^{\hat{n}_i}, \forall i \in [k]} \left[\frac{1}{\tilde{n}} \max_{h \in \mathcal{H}} \sum_{i=1}^k \sum_{j=1}^{\hat{n}_i} \sigma_i^j \ell(h, z_i^j) \right] \right] \right],$$

where each z_i^j is independently drawn from \mathcal{D}_i , and each σ_i^j is an independent Rademacher random variable.

In addition, Lemma 9 tells us that: for any $i \in [k]$, one has

$$\hat{n}_i \geq \frac{1}{3} n_i - 2 \log(2k) \geq \frac{1}{6} n_i \quad \implies \quad \hat{n}_i \geq \left\lceil \frac{1}{6} n_i \right\rceil =: \tilde{n}_i \quad (146)$$

with probability exceeding $1 - \frac{1}{2k}$. By defining \mathcal{E} to be the event where $\hat{n}_i \geq n_i/6$ for all $i \in [k]$, we can immediately see from the union bound that

$$\mathbb{P}(\mathcal{E}) \geq 1/2.$$

Consequently, we can derive

$$\begin{aligned}
\text{Rad}_{\tilde{n}}(\mathcal{D}(w)) &\geq \mathbb{P}(\mathcal{E}) \cdot \mathbb{E}_{\{X_i\}_{i=1}^{\tilde{n}}} \left[\mathbb{E}_{\{z_i^j\}_{j=1}^{\hat{n}_i}, \forall i \in [k]} \left[\mathbb{E}_{\{\sigma_i^j\}_{j=1}^{\hat{n}_i}, \forall i \in [k]} \left[\frac{1}{\tilde{n}} \max_{h \in \mathcal{H}} \sum_{i=1}^k \sum_{j=1}^{\hat{n}_i} \sigma_i^j \ell(h, z_i^j) \mid \mathcal{E} \right] \right] \right] \\
&\geq \frac{1}{2} \cdot \frac{\sum_{i=1}^k \tilde{n}_i}{\tilde{n}} \widetilde{\text{Rad}}_{\{\tilde{n}_i\}_{i=1}^k} \\
&\geq \frac{1}{12} \cdot \frac{1}{6} \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k}, \quad (147)
\end{aligned}$$

where the last two inequalities both follow from Lemma 19. This concludes the proof. \square

F.3 Necessity of Assumption 1

In this section, we will discuss the necessity of Assumption 1 in comparison to the following weaker assumption, the latter of which only assumes that the Rademacher complexity on each \mathcal{D}_i is well-bounded.

Assumption 3. For each $n \geq 1$, there exists a universal constant $C_n \geq 0$ such that

$$\text{Rad}_n(\mathcal{D}_i) \leq C_n \quad (148)$$

for all $i \in [k]$.

Formally, we have the following results.

Lemma 21. Let $w^0 = [1/k, 1/k, \dots, 1/k]^\top$. There exist a group of distributions $\{\mathcal{D}_i\}_{i=1}^k$ and a hypothesis set \mathcal{H} such that

$$\text{Rad}_n(\mathcal{D}(w^0)) \geq \Omega\left(\frac{1}{k} \sum_{i=1}^k \text{Rad}_{n/k}(\mathcal{D}_i)\right) \quad (149)$$

for $n \geq 12k \log(k)$.

Proof. Without loss of generality, we set $\mathcal{Y} = \{0\}$ and $\ell(h, (x, y)) = h(x) - y = h(x)$. We can then regard \mathcal{D}_i as a distribution over \mathcal{X}_i because there is only one element in \mathcal{Y} .

Pick k subsets of \mathcal{X} as $\{\mathcal{X}_i\}_{i=1}^k$. For each $i \in [k]$, we choose the distribution \mathcal{D}_i to be an arbitrary distribution supported by \mathcal{X}_i . In addition, we define \mathcal{H}_i to be a set of hypothesis $h(x) = 0$ for all $x \notin \mathcal{X}_i$ for each $i \in [k]$. For $\{h_i\}_{i=1}^k$ such that $h_i \in \mathcal{H}_i$, we define $\text{joint}(\{h_i\}_{i=1}^k)$ to be the hypothesis h such that $h(x) = h_i(x)$, $\forall x \in \mathcal{X}_i, i \in [k]$ and $h(x) = 0$ for $x \notin \cup_i \mathcal{X}_i$. Now we construct the hypothesis set \mathcal{H} as

$$\mathcal{H} = \{\text{joint}(\{h_i\}_{i=1}^k) | h_i \in \mathcal{H}_i, \forall i \in [k]\}. \quad (150)$$

Recalling the definition of $\widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k}(\{\mathcal{D}_i\}_{i=1}^k)$, we see that

$$\begin{aligned} \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k}(\{\mathcal{D}_i\}_{i=1}^k) &= \mathbb{E}_{\{x_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\mathbb{E}_{\{\sigma_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\frac{1}{\sum_{i=1}^k n_i} \max_{h \in \mathcal{H}} \sum_{i=1}^k \sum_{j=1}^{n_i} \sigma_i^j h(x_i^j) \right] \right] \\ &= \mathbb{E}_{\{z_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\mathbb{E}_{\{\sigma_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^k \max_{h_i \in \mathcal{H}_i} \sum_{j=1}^{n_i} \sigma_i^j h_i(x_i^j) \right] \right] \\ &= \frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^k n_i \mathbb{E}_{\{z_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\mathbb{E}_{\{\sigma_i^j\}_{j=1}^{n_i}} \left[\frac{1}{n_i} \max_{h_i \in \mathcal{H}_i} \sum_{j=1}^{n_i} \sigma_i^j h_i(x_i^j) \right] \right] \\ &= \frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^k n_i \text{Rad}_{n_i}(\mathcal{D}_i). \end{aligned} \quad (151)$$

Here (151) is by the definition of \mathcal{H} . By taking $n_i = \frac{n}{k}$ for all $i \in [k]$ and Lemma 6, we reach

$$\frac{1}{k} \sum_{i=1}^k \text{Rad}_{n/k}(\mathcal{D}_i) = \widetilde{\text{Rad}}_{\{n_i\}_{i=1}^k}(\{\mathcal{D}_i\}_{i=1}^k) \leq 72 \text{Rad}_n(\mathcal{D}(w^0)) \quad (152)$$

□

By virtue of Lemma 21, if we set $C_n \sim \sqrt{\frac{d}{n}}$ in Assumption 3, the best upper bound on is $\text{Rad}_n(\mathcal{D}(w^0)) \sim \sqrt{\frac{dk}{n}}$, which implies that more samples are needed to learn the mixed distribution w^0 . Moreover, under the construction in Lemma 21, by further assuming $\min_{h_i \in \mathcal{H}_i} \mathbb{E}_{x \sim \mathcal{D}_i} [h_i(x)] = \frac{1}{2}$ for all $i \in [k]$, to find h such that

$$\max_i \mathbb{E}_{x \sim \mathcal{D}_i} [h(x)] \leq \frac{1}{2} + \varepsilon, \quad (153)$$

we need to find $h_i \in \mathcal{H}_i$ such that

$$\mathbb{E}_{x \sim \mathcal{D}_i}[h_i(x)] \leq \frac{1}{2} + \varepsilon \quad (154)$$

for all $i \in [k]$. Following this intuition, we can construct a counter example under Assumption 3 as follows.

Theorem 5. *There exist a group of distributions $\{\mathcal{D}_i\}_{i=1}^k$ and a hypothesis set \mathcal{H} such that Assumption 3 holds with $C_n = O\left(\sqrt{\frac{d}{n}}\right)$, and it takes at least $\tilde{\Omega}\left(\frac{dk}{\varepsilon^2}\right)$ samples to find some $h \in \mathcal{H}$ obeying*

$$\max_{i \in [k]} L(h, e_i) \leq \min_{h' \in \mathcal{H}} \max_{i \in [k]} L(h', e_i) + \varepsilon.$$

Proof. With the construction in Lemma 21, it suffices to find some \mathcal{H}' and \mathcal{D}' such that the following three conditions hold.

1. The following inequality holds:

$$\text{Rad}_n(\mathcal{D}', \mathcal{H}') := \frac{1}{n} \mathbb{E}_{\{x^j\}_{j=1}^n \stackrel{\text{iid}}{\sim} \mathcal{D}', \{\sigma^j\}_{j=1}^n \stackrel{\text{iid}}{\sim} \{\pm 1\}} \left[\max_{h' \in \mathcal{H}'} \sum_{j=1}^n \sigma^j h'(x^j) \right] \leq C_n; \quad (155)$$

2. $\min_{h' \in \mathcal{H}} \mathbb{E}_{x \sim \mathcal{D}'}[h(x)] = \frac{1}{2};$

3. It takes at least $\tilde{\Omega}\left(\frac{d}{\varepsilon^2}\right)$ samples to find some h such that $\mathbb{E}_{x \sim \mathcal{D}'}[h(x)] \leq \frac{1}{2} + \varepsilon.$

This construction is also straightforward. We set $N = 2^d$ and $\mathcal{X}' = \{0, 1\}^N$. Let \mathcal{D}' to be the distribution $\mathbb{P}_{\mathcal{D}'}[x] = \Pi_{n=1}^N \mathbb{P}_{\mathcal{D}_n}[x_n]$, where $\mathbb{P}_{\mathcal{D}_n^*}[x_{n^*}] = \frac{1}{2} \mathbb{I}[x_{n^*} = 1] + \frac{1}{2} \mathbb{I}[x_{n^*} = 0]$ for some n^* , $\mathbb{P}_{\mathcal{D}_n}[x_n] = (\frac{1}{2} + 2\varepsilon) \mathbb{I}[x_{n^*} = 1] + (\frac{1}{2} - 2\varepsilon) \mathbb{I}[x_{n^*} = 0]$ for all $n \neq n^*$. Then we set $\mathcal{H}' = \{h^n\}_{n=1}^N$ with $h^n(x) = x_n$ for each $n \in [N]$. It is then easy to verify that the first two conditions hold. Regarding the third condition, following the arguments in Theorem 2, we need at least $\tilde{\Omega}\left(\frac{d}{\varepsilon^2}\right)$ i.i.d. samples from \mathcal{D}' to identify n^* . The proof is thus completed. \square

G Proof of Theorem 4

Algorithm 4: Hedge for Multi-group Learning

- 1 **Input:** The hypothesis set \mathcal{H} , the group set \mathcal{G} , target accuracy level ε , target success rate $1 - \delta$, minimal probability γ .
 - 2 **Initialization:** $T = \frac{10000 \log(|\mathcal{G}|/\delta)}{\gamma \varepsilon^2}$, $\eta = \frac{1}{10} \varepsilon \gamma$, $N = \frac{20000(\log(|\mathcal{G}|/\delta) + d \log(d/\varepsilon))}{\gamma \varepsilon^2}$, $W_g^1 = 1$ for all $g \in \mathcal{G}$.
 - 3 Call Query(\mathcal{D}) for N times to get N i.i.d. samples from \mathcal{D} , denoted by $\{(x_i, y_i)\}_{i=1}^N$.
 - 4 $f_g \leftarrow \min_{h \in \mathcal{H}} \frac{1}{\sum_{i=1}^N \mathbb{I}[x_i \in g]} \sum_{i=1}^N \mathbb{I}[x_i \in g] \ell(h(x_i), y_i)$, $\forall g \in \mathcal{G}$,
 - 5 $N_g \leftarrow \sum_{i=1}^N \mathbb{I}[x_i \in g]$, $p_g \leftarrow \frac{N_g}{N}$, $\forall g \in \mathcal{G}$
 - 6 **for** $t = 1, 2, \dots, T$ **do**
 - 7 $w_g^t \leftarrow \frac{W_g^t}{\sum_{g \in \mathcal{G}} W_g^t}$, $\forall g \in \mathcal{G}$,
 - 8 $h^t \leftarrow \arg \min_{h \in \mathcal{H}} \left(\sum_{g \in \mathcal{G}} w_g^t \cdot \left(\frac{1}{N_g} \sum_{i=1}^N \mathbb{I}[x_i \in g] \ell(h(x_i), y_i) - f_g \right) \right)$,
 - 9 Call Query(\mathcal{D}) to get a data point (\hat{x}_t, \hat{y}_t) ,
 - 10 $\hat{r}_g^t \leftarrow \frac{1}{p_g} \mathbb{I}[\hat{x}_t \in g] (\ell(h^t(\hat{x}_t), \hat{y}_t) - f_g)$, $\forall g \in \mathcal{G}$,
 - 11 $W_g^{t+1} \leftarrow W_g^t \cdot \exp(\eta \hat{r}_g^t)$, $\forall g \in \mathcal{G}$;
 - 12 **Return:** A randomized hypothesis h^{final} as a uniform distribution over $\{h^t\}_{t=1}^T$.
-

Throughout this section, we let \mathcal{F}_t be the event field before the t -th round. Clearly, the number of samples used in Algorithm 4 is bounded by $T + N = O\left(\frac{(\log(|\mathcal{G}|/\delta) + d \log(d/\varepsilon))}{\gamma \varepsilon^2}\right)$. So it suffices to prove the optimality of h^{final} . Formally, we have the lemma below.

Lemma 22. *With probability at least $1 - \delta$, it holds that*

$$\frac{1}{T} \sum_{t=1}^T L_{\mathcal{D}}(h^t | g) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h | g) + \varepsilon \quad (156)$$

for any $g \in \mathcal{G}$.

Proof. We first show that $p_g \geq \eta$ with high probability. Using Lemma 9, for fixed $g \in \mathcal{G}$, we have that

$$\mathbb{P} \left[Np_g \leq \frac{1}{3}NP_g - \frac{1}{3} \log(1/\delta') \right] \leq \delta'. \quad (157)$$

By choosing $\delta' = \frac{\delta}{64|\mathcal{G}|}$, with probability at least $1 - \frac{\delta}{64}$, it holds that

$$p_g \geq \frac{1}{3}P_g - \frac{1}{3N} \log(1/\delta') \geq \frac{1}{6}P_g$$

for any $g \in \mathcal{G}$.

Let $W^t = \sum_{g \in \mathcal{G}} W_g^t$. Because $\eta \hat{r}_g^t \leq \frac{\eta}{p_g} \leq 6\varepsilon \leq 1$ for any proper (g, t) pair, we have

$$\begin{aligned} \log(W_g^{T+1}) &= \eta \sum_{t=1}^T \hat{r}_g^t, \\ \log \left(\frac{W^{T+1}}{W^t} \right) &= \log \left(\sum_{g \in \mathcal{G}} w_g^t \exp(\eta \hat{r}_g^t) \right) \leq \log \left(\sum_{g \in \mathcal{G}} w_g^t (1 + \eta \hat{r}_g^t + \eta^2 (\hat{r}_g^t)^2) \right) \leq \sum_{g \in \mathcal{G}} w_g^t (\eta \hat{r}_g^t + \eta^2 (\hat{r}_g^t)^2). \end{aligned}$$

Observe that

$$\log(W_g^{T+1}) \leq \log(W^{T+1}) = \sum_{t=1}^T \log(W^{t+1}/W^t) + \log(|\mathcal{G}|)$$

for any $\tilde{g} \in \mathcal{G}$,

$$\eta \sum_{t=1}^T \hat{r}_{\tilde{g}}^t \leq \eta \sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t \hat{r}_g^t + \eta^2 \sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t (\hat{r}_g^t)^2 + \log(|\mathcal{G}|).$$

Dividing both side by η , we reach

$$\sum_{t=1}^T \hat{r}_{\tilde{g}}^t - \sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t \hat{r}_g^t \leq \frac{\log(|\mathcal{G}|)}{\eta} + \eta \sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t (\hat{r}_g^t)^2. \quad (158)$$

By Lemma 24, with probability at least $1 - \frac{\delta}{16}$,

$$\sum_{t=1}^T \hat{r}_{\tilde{g}}^t - \sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t \hat{r}_g^t \leq \frac{\log(|\mathcal{G}|)}{\eta} + \frac{3\eta T}{\gamma} \leq \frac{5T\varepsilon}{16} \quad (159)$$

for any $\tilde{g} \in \mathcal{G}$. By (159) and Lemma 25, with probability at least $1 - \frac{3\delta}{4}$,

$$\sum_{t=1}^T (L_{\mathcal{D}}(h^t | g) - f_g) \leq \sum_{t=1}^T \hat{r}_g^t + \frac{T\varepsilon}{8} \leq \sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t \hat{r}_g^t + \frac{7T\varepsilon}{16} \leq \frac{7T\varepsilon}{8} \quad (160)$$

for any $g \in \mathcal{G}$. By Lemma 23, with probability at least $1 - \delta$, we have that

$$\sum_{t=1}^T L_{\mathcal{D}}(h^t | g) - T \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h | g) \leq T\varepsilon,$$

which means that

$$\frac{1}{T} \sum_{t=1}^T L_{\mathcal{D}}(h^t|g) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h|g) \leq \varepsilon.$$

The proof is completed. \square

Lemma 23. *with probability at least $1 - \frac{\delta}{16}$, it holds that*

$$\left| f_g - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h|g) \right| \leq \frac{\varepsilon}{8}$$

for any $g \in \mathcal{G}$.

Proof. Fix $g \in \mathcal{G}$. Let $\{(x'_i, y'_i)\}_{i=1}^N$ to be the ghost samples, another group of N i.i.d. datapoints from \mathcal{D} which is independent of $\{(x_i, y_i)\}_{i=1}^N$. Let $\lambda \in (0, 1/2]$ be a positive real number.

Then we have that

$$\begin{aligned} & \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^N} \left[\exp \left(\lambda \max_{h \in \mathcal{H}} \left(\sum_{i=1}^N \mathbb{1}[x_i \in g] \ell(h(x_i), y_i) - N \mathbb{E}_{(x, y) \sim \mathcal{D}} [\mathbb{1}[x \in g] \ell(h(x), y)] \right) \right) \right] \\ & \leq \mathbb{E}_{\{(x_i, y_i), (x'_i, y'_i)\}_{i=1}^N} \left[\exp \left(\lambda \max_{h \in \mathcal{H}} \left(\sum_{i=1}^N (\mathbb{1}[x_i \in g] \ell(h(x_i), y_i) - \mathbb{1}[x'_i \in g] \ell(h(x'_i), y'_i)) \right) \right) \right] \end{aligned} \quad (161)$$

$$\begin{aligned} & = \mathbb{E}_{\{(x_i, y_i), (x'_i, y'_i)\}_{i=1}^N} \mathbb{E}_{\{\sigma_i\}_{i=1}^N \stackrel{\text{i.i.d.}}{\sim} \{\pm 1\}} \left[\exp \left(\lambda \max_{h \in \mathcal{H}} \sum_{i=1}^N \sigma_i (\mathbb{1}[x_i \in g] \ell(h(x_i), y_i) - \mathbb{1}[x'_i \in g] \ell(h(x'_i), y'_i)) \right) \right] \\ & \leq \mathbb{E}_{\{(x_i, y_i), (x'_i, y'_i)\}_{i=1}^N} \left[\exp \left(\sum_{i=1}^N 2\lambda^2 (\mathbb{1}[x_i \in g] + \mathbb{1}[x'_i \in g])^2 \right) \right] \\ & \leq \mathbb{E}_{\{(x_i, y_i), (x'_i, y'_i)\}_{i=1}^N} \left[\exp \left(\sum_{i=1}^N 4\lambda^2 (\mathbb{1}[x_i \in g] + \mathbb{1}[x'_i \in g]) \right) \right] \\ & \leq \exp(16N\lambda^2 P_g). \end{aligned} \quad (162)$$

Here, (161) arises from Jensen's inequality. As a result, we can deduce that

$$\begin{aligned} & \mathbb{P} \left[\max_{h \in \mathcal{H}} \left(\sum_{i=1}^N \mathbb{1}[x_i \in g] \ell(h(x_i), y_i) - N \mathbb{E}_{(x, y) \sim \mathcal{D}} [\mathbb{1}[x \in g] \ell(h(x), y)] \right) > NP_g \varepsilon / 32 \right] \\ & \leq \min_{\lambda \in (0, 1/2]} \exp(4N\lambda^2 P_g - \lambda NP_g \varepsilon / 32) \\ & \leq \exp \left(-\frac{NP_g \varepsilon^2}{16384} \right) \end{aligned} \quad (163)$$

$$\begin{aligned} & \leq \exp \left(-\frac{N\gamma \varepsilon^2}{16384} \right) \\ & \leq \frac{\delta}{64|\mathcal{G}|}. \end{aligned} \quad (164)$$

Using similar arguments, we can obtain

$$\mathbb{P} \left[\max_{h \in \mathcal{H}} \left(-\sum_{i=1}^N \mathbb{1}[x_i \in g] \ell(h(x_i), y_i) + N \mathbb{E}_{(x, y) \sim \mathcal{D}} [\mathbb{1}[x \in g] \ell(h(x), y)] \right) > NP_g \varepsilon / 32 \right] \leq \frac{\delta}{64|\mathcal{G}|} \quad (165)$$

As a consequence, we see that

$$\mathbb{P} \left[\exists h, \left| \frac{1}{NP_g} \sum_{i=1}^N \mathbb{1}[x_i \in g] \ell(h(x_i), y_i) - L_{\mathcal{D}}(h|g) \right| \geq \varepsilon / 32 \right] \leq \frac{\delta}{32|\mathcal{G}|}. \quad (166)$$

Recall that $p_g = \frac{\sum_{i=1}^N \mathbb{1}[x_i \in g]}{N}$. Applying Chernoff's inequality gives

$$\mathbb{P} \left[|NP_g - Np_g| \geq 2\sqrt{NP_g \log(128|\mathcal{G}|/\delta)} \right] \leq \frac{\delta}{64|\mathcal{G}|}, \quad (167)$$

which means that

$$\mathbb{P} \left[|P_g - p_g| \geq \frac{P_g \varepsilon}{32} \right] \leq \frac{\delta}{64|\mathcal{G}|}. \quad (168)$$

By (166) and (168), and taking a union bound over \mathcal{G} , we obtain that with probability at least $1 - \frac{\delta}{16}$,

$$\left| \frac{1}{Np_g} \sum_{i=1}^N \mathbb{1}[x_i \in g] \ell(h(x_i), y_i) - L_{\mathcal{D}}(h|g) \right| \leq \frac{\varepsilon}{16} \quad (169)$$

for any $h \in \mathcal{H}, g \in \mathcal{G}$. From the definition of h_g , we have

$$\left| f_g - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h|g) \right| \leq 2 \max_{h \in \mathcal{H}} \left| \frac{1}{Np_g} \sum_{i=1}^N \mathbb{1}[x_i \in g] \ell(h(x_i), y_i) - L_{\mathcal{D}}(h|g) \right| \leq \frac{\varepsilon}{8}.$$

This concludes the proof. \square

Lemma 24. *With probability at least $1 - \frac{\delta}{16}$, it holds that*

$$\sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t (\hat{r}_g^t)^2 \leq \frac{3T}{\gamma}. \quad (170)$$

Proof. By definition, we have

$$\sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t (\hat{r}_g^t)^2 \leq \sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t \frac{\mathbb{1}[x_t \in g]}{p_g^2}. \quad (171)$$

Let $X_t = \sum_{g \in \mathcal{G}} w_g^t \frac{\mathbb{1}[x_t \in g]}{p_g^2}$ and recall \mathcal{F}_t is the event field before the t -th round. Then $X_t \leq \frac{1}{\gamma^2}$, and $\mathbb{E}[X_t^2 | \mathcal{F}_t] = \sum_{g, g'} w_g^t w_{g'}^t \frac{\mathbb{E}[\mathbb{1}[\hat{x}_t \in g, \hat{x}_t \in g']]}{p_g^2 p_{g'}^2} \leq \sum_{g, g'} w_g^t w_{g'}^t \frac{1}{(p_g p_{g'})^{3/2}} \leq \frac{1}{\gamma^3}$. Freedman's inequality (Lemma 8) reveals that, with probability at least $1 - \frac{\delta}{16}$,

$$\sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t (\hat{r}_g^t)^2 \leq \sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t \frac{1}{p_g} + 2\sqrt{\frac{T \log(16/\delta)}{\gamma^3}} \leq \frac{3T}{\gamma}, \quad (172)$$

where the last inequality arises from the fact that $T \geq \frac{\log(16/\delta)}{\gamma}$. \square

Lemma 25. *Assume the events in Lemma 23 hold. With probability at least $1 - \frac{\delta}{4}$, it holds that*

$$\left| \sum_{t=1}^T \hat{r}_g^t - \sum_{t=1}^T (L_{\mathcal{D}}(h^t|g) - f_g) \right| \leq \frac{T\varepsilon}{8}, \quad \forall g \in \mathcal{G}; \quad (173)$$

$$\sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t \hat{r}_g^t \leq \frac{7T\varepsilon}{16}. \quad (174)$$

Proof. Let us begin with the first inequality. By definition,

$$\sum_{t=1}^T \hat{r}_g^t - \sum_{t=1}^T (L_{\mathcal{D}}(h^t|g) - f_g)$$

$$= \sum_{t=1}^T \left(\frac{1}{p_g} \mathbb{1}[\hat{x}_t \in g] \ell(h(\hat{x}_t), \hat{y}_t) - L_{\mathcal{D}}(h^t|g) \right) + \sum_{t=1}^T \left(\frac{1}{p_g} \mathbb{1}[\hat{x}_t \in g] - 1 \right) f_g. \quad (175)$$

Defining $\hat{X}_t = \frac{1}{p_g} \mathbb{1}[\hat{x}_t \in g] \ell(h(\hat{x}_t), \hat{y}_t)$. We then see that $\mathbb{E}[\hat{X}_t | \mathcal{F}_t] = \frac{P_g}{p_g} L_{\mathcal{D}}(h^t|g)$ and $\mathbb{E}[\hat{X}_t^2 | \mathcal{F}_t] \leq \frac{P_g}{p_g^2}$. According to Freedman's inequality (Lemma 8), we know that with probability at least $1 - \frac{\delta}{64}$,

$$\left| \sum_{t=1}^T \left(\frac{1}{p_g} \mathbb{1}[\hat{x}_t \in g] \ell(h(\hat{x}_t), \hat{y}_t) - \frac{P_g}{p_g} L_{\mathcal{D}}(h^t|g) \right) \right| \leq \frac{T\varepsilon}{32}. \quad (176)$$

for any $g \in \mathcal{G}$. Combine (176) with (168) to show that: with probability at least $1 - \frac{\delta}{32}$,

$$\left| \sum_{t=1}^T \left(\frac{1}{p_g} \mathbb{1}[\hat{x}_t \in g] \ell(h(\hat{x}_t), \hat{y}_t) - L_{\mathcal{D}}(h^t|g) \right) \right| \leq \frac{T\varepsilon}{16} \quad (177)$$

holds for any $g \in \mathcal{G}$.

Similarly, we have with probability exceeding $1 - \frac{\delta}{32}$

$$\sum_{t=1}^T \left(\frac{1}{p_g} \mathbb{1}[\hat{x}_t \in g] - 1 \right) f_g \leq \frac{T\varepsilon}{16}.$$

Therefore with probability at least $1 - \frac{\delta}{16}$,

$$\left| \sum_{t=1}^T \hat{r}_g^t - \sum_{t=1}^T (L_{\mathcal{D}}(h^t|g) - f_g) \right| \leq \frac{T\varepsilon}{8}, \quad (178)$$

holds for any $g \in \mathcal{G}$.

As for the second inequality, we define $\tilde{X}_t = \sum_{g \in \mathcal{G}} w_g^t \left(\frac{1}{p_g} \mathbb{1}[\hat{x}_t \in g] (\ell(h^t(\hat{x}_t), \hat{y}_t) - f_g) \right)$. It then follows that $\mathbb{E}[\tilde{X}_t | \mathcal{F}_t] = \sum_{g \in \mathcal{G}} w_g^t \frac{P_g}{p_g} (L_{\mathcal{D}}(h^t|g) - f_g) \leq \frac{6}{\gamma}$ and

$$\mathbb{E}[\tilde{X}_t^2 | \mathcal{F}_t] \leq \sum_{g', g''} w_g^t w_{g''}^t \frac{\mathbb{E}[\mathbb{1}[\hat{x}_t \in g, \hat{x}_t \in g'']]}{p_g p_{g''}} \leq \sum_{g', g''} w_g^t w_{g''}^t \frac{1}{\sqrt{p_g p_{g''}}} \leq \frac{6}{\gamma}.$$

Invoking Freedman's inequality (cf. Lemma 8), we can demonstrate that, with probability at least $1 - \frac{\delta}{64}$,

$$\sum_{t=1}^T \left(\tilde{X}_t - \sum_{g \in \mathcal{G}} w_g^t \frac{P_g}{p_g} (L_{\mathcal{D}}(h^t|g) - f_g) \right) \leq \frac{T\varepsilon}{32}. \quad (179)$$

As a result, taking this together with (168) gives

$$\sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t \hat{r}_g^t \leq \sum_{t=1}^T \sum_{g \in \mathcal{G}} w_g^t (L_{\mathcal{D}}(h^t|g) - f_g) + \frac{T\varepsilon}{16}. \quad (180)$$

Recall the definition that

$$h^t = \arg \min_{h \in \mathcal{H}} \left(\sum_{g \in \mathcal{G}} w_g^t \left(\frac{1}{N_g} \sum_{i=1}^N \mathbb{1}[x_i \in g] \ell(h(x_i), y_i) - f_g \right) \right). \quad (181)$$

By Lemma 23, one has

$$\max_{h \in \mathcal{H}} \left| \frac{1}{N p_g} \sum_{i=1}^N \mathbb{1}[x_i \in g] \ell(h(x_i), y_i) - L_{\mathcal{D}}(h|g) \right| \leq \frac{\varepsilon}{16} \quad (182)$$

for any $g \in \mathcal{G}$. It follows from the definition that $Np_g = N_g$, and consequently,

$$\begin{aligned} \sum_{g \in \mathcal{G}} w_g^t \left(\frac{1}{Np_g} \sum_{i=1}^N \mathbb{1}[x_i \in g] \ell(h^t(x_i), y_i) - f_g \right) &\leq \sum_{g \in \mathcal{G}} w_g^t \left(\frac{1}{Np_g} \sum_{i=1}^N \mathbb{1}[x_i \in g] \ell(h^*(x_i), y_i) - f_g \right) \\ &\leq \sum_{g \in \mathcal{G}} w_g^t (L_{\mathcal{D}}(h^* | g) - f_g) + \frac{\varepsilon}{16}. \end{aligned} \quad (183)$$

In view of (183), Assumption 2 and Lemma 23, we have

$$\sum_{g \in \mathcal{G}} w_g^t \left(\frac{1}{Np_g} \sum_{i=1}^N \mathbb{1}[x_i \in g] \ell(h^t(x_i), y_i) - f_g \right) \leq \frac{5}{16} \varepsilon. \quad (184)$$

Moreover, we also have

$$\sum_{g \in \mathcal{G}} w_g^t (L_{\mathcal{D}}(h^t | g) - f_g) \leq \sum_{g \in \mathcal{G}} w_g^t \left(\frac{1}{Np_g} \sum_{i=1}^N \mathbb{1}[x_i \in g] \ell(h^t(x_i), y_i) - f_g \right) + \frac{\varepsilon}{16},$$

which implies that

$$\sum_{g \in \mathcal{G}} w_g^t (L_{\mathcal{D}}(h^t | g) - f_g) \leq \frac{3\varepsilon}{8}. \quad (185)$$

The advertised result thus follows by combining (180) and (185). \square

References

- Abernethy, J., Awasthi, P., Kleindessner, M., Morgenstern, J., Russell, C., and Zhang, J. (2022). Active sampling for min-max fairness. *International Conference on Machine Learning*, pages 53–65.
- Arora, S., Hazan, E., and Kale, S. (2012). The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164.
- Asi, H., Carmon, Y., Jambulapati, A., Jin, Y., and Sidford, A. (2021). Stochastic bias-reduced gradient methods. *Advances in Neural Information Processing Systems*, 34:10810–10822.
- Awasthi, P., Haghtalab, N., and Zhao, E. (2023). Open problem: The sample complexity of multi-distribution learning for VC classes. In *Conference on Learning Theory (COLT)*, volume 195, pages 5943–5949.
- Blum, A., Haghtalab, N., Phillips, R. L., and Shao, H. (2021a). One for one, or all for all: Equilibria and optimality of collaboration in federated learning. In *International Conference on Machine Learning*, pages 1005–1014.
- Blum, A., Haghtalab, N., Procaccia, A. D., and Qiao, M. (2017). Collaborative PAC learning. *Advances in Neural Information Processing Systems*, 30.
- Blum, A., Heinecke, S., and Reyzin, L. (2021b). Communication-aware collaborative learning. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 6786–6793.
- Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965.
- Bühlmann, P. and Meinshausen, N. (2015). Magging: maximin aggregation for inhomogeneous large-scale data. *Proceedings of the IEEE*, 104(1):126–135.

- Carmon, Y. and Hausler, D. (2022). Distributionally robust optimization via ball oracle acceleration. *Advances in Neural Information Processing Systems*, 35:35866–35879.
- Chen, J., Zhang, Q., and Zhou, Y. (2018). Tight bounds for collaborative PAC learning via multiplicative weights. *Advances in Neural Information Processing Systems*, 31.
- Deng, Y., Kamani, M. M., and Mahdavi, M. (2020). Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 33:15111–15122.
- Du, W., Xu, D., Wu, X., and Tong, H. (2021). Fairness-aware agnostic federated learning. In *SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM.
- Dudík, M., Haghtalab, N., Luo, H., Schapire, R. E., Syrgkanis, V., and Vaughan, J. W. (2020). Oracle-efficient online learning and auction design. *Journal of the ACM (JACM)*, 67(5):1–57.
- Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. (2021). Outcome indistinguishability. In *ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108.
- Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118.
- Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.
- Guo, Z. (2023). Statistical inference for maximin effects: Identifying stable associations across multiple studies. *Journal of the American Statistical Association*, pages 1–17.
- Haghtalab, N., Jordan, M., and Zhao, E. (2022). On-demand sampling: Learning optimally from multiple distributions. *Advances in Neural Information Processing Systems*, 35:406–419.
- Haghtalab, N., Jordan, M., and Zhao, E. (2023). A unifying perspective on multi-calibration: Game dynamics for multi-objective learning. In *Advances in Neural Information Processing Systems*.
- Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938.
- Hazan, E. (2022). *Introduction to online convex optimization*. MIT Press.
- Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948.
- Hu, W., Niu, G., Sato, I., and Sugiyama, M. (2018). Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037.
- Kar, A., Prakash, A., Liu, M.-Y., Cameracci, E., Yuan, J., Rusiniak, M., Acuna, D., Torralba, A., and Fidler, S. (2019). Meta-sim: Learning to generate synthetic datasets. In *IEEE/CVF International Conference on Computer Vision*, pages 4551–4560.
- Mansour, Y., Mohri, M., and Rostamizadeh, A. (2008). Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21.
- Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.
- Mohri, M., Sivek, G., and Suresh, A. T. (2019). Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625.
- Nguyen, H. and Zakynthinou, L. (2018). Improved algorithms for collaborative PAC learning. *Advances in Neural Information Processing Systems*, 31.

- Peng, B. (2023). The sample complexity of multi-distribution learning. *arXiv preprint arXiv:2312.04027*.
- Rothblum, G. N. and Yona, G. (2021). Multi-group agnostic PAC learnability. In *International Conference on Machine Learning*, pages 9107–9115.
- Roughgarden, T. (2016). *Twenty lectures on algorithmic game theory*. Cambridge University Press.
- Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.
- Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. (2020). An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356.
- Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.
- Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.
- Tosh, C. J. and Hsu, D. (2022). Simple and near-optimal algorithms for hidden stratification and multi-group learning. In *International Conference on Machine Learning*, pages 21633–21657.
- v. Neumann, J. (1928). Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320.
- Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.
- Vapnik, V., Levin, E., and Le Cun, Y. (1994). Measuring the VC-dimension of a learning machine. *Neural computation*, 6(5):851–876.
- Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.
- Wang, Z., Bühlmann, P., and Guo, Z. (2023). Distributionally robust machine learning with multi-source data. *arXiv preprint arXiv:2309.02211*.
- Xiong, X., Guo, Z., and Cai, T. (2023). Distributionally robust transfer learning. *arXiv preprint arXiv:2309.06534*.
- Zhang, J., Menon, A. K., Veit, A., Bhojanapalli, S., Kumar, S., and Sra, S. (2020). Coping with label shift via distributionally robust optimisation. In *International Conference on Learning Representations*.
- Zhang, Z., Ji, X., and Du, S. (2022). Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory (COLT)*, pages 3858–3904.
- Zhao, S., Li, B., Xu, P., and Keutzer, K. (2020). Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*.