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: May 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

Accepted for presentation at the Conference on Learning Theory (COLT) 2024.
*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.

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 function[1] $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [-1, 1]$, we are asked to identify a (possibly randomized) hypothesis $\widehat{h}$ achieving near-optimal *worst-case* loss across these data distributions, namely,[2]

$$\max_{1 \leqslant i \leqslant k} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i, \widehat{h}} \left[\ell\big(\widehat{h}, (x, y)\big)\right] \leqslant \min_{h\in\mathcal{H}} \max_{1 \leqslant i \leqslant k} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i} \left[\ell\big(h, (x, y)\big)\right] + \varepsilon \tag{1}$$

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

[1] For 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.

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

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 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 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 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 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 $O\big(\frac{\log(|\mathcal{H}|)+k\log(k/\delta)}{\varepsilon^2}\big)$ 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.

Paper	Sample complexity bound
Haghtalab et al. (2022)	$\frac{\log(\lvert\mathcal{H}\rvert)+k}{\varepsilon^2}$
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 $\big(k, d, \frac{1}{\varepsilon}, \frac{1}{\delta}\big)$.

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 Zakynthinou, 2018), a special scenario where the loss function takes the form of $\ell\big(h,(x,y)\big)=\mathbb{1}\{h(x)\neq y\}$ 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]

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

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

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

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 a (potentially large) threshold on the order of $\frac{d^2k^2}{d+k}$. 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 obliviates the need for computing an epsilon-net of $\mathcal{H}$ — yielding a sample complexity of

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

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 $\widetilde{O}\left(\frac{d+k}{\varepsilon^2}\right)$ 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).

[3]Let $\mathcal{X}=\big(k,d,\frac{1}{\varepsilon},\frac{1}{\delta}\big)$. Here and throughout, the notation $f(\mathcal{X})=O\big(g(\mathcal{X})\big)$ or $f(\mathcal{X})\lesssim g(\mathcal{X})$ (resp. $f(\mathcal{X})=\Omega\big(g(\mathcal{X})\big)$ or $f(\mathcal{X})\gtrsim g(\mathcal{X})$) mean that there exists some universal constant $C_1>0$ (resp. $C_2>0$) such that $f(\mathcal{X})\leqslant C_1 g(\mathcal{X})$ (resp. $f(\mathcal{X})\geqslant C_2 g(\mathcal{X})$); the notation $f(\mathcal{X})=\Theta\big(g(\mathcal{X})\big)$ or $f(\mathcal{X})\asymp g(\mathcal{X})$ mean $f(\mathcal{X})=O\big(g(\mathcal{X})\big)$ and $f(\mathcal{X})=\Omega\big(g(\mathcal{X})\big)$ hold simultaneously. The notation $\widetilde{O}(\cdot)$, $\widetilde{\Theta}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ are defined analogously except that all log factors in $\big(k,d,\frac{1}{\varepsilon},\frac{1}{\delta}\big)$ are hidden.

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 in the following theorem.[4]

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*

$$\max_{1\leqslant i\leqslant k} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i, h^{\mathsf{final}}}\left[\ell\big(h^{\mathsf{final}},(x,y)\big)\right] \leqslant \min_{h\in\mathcal{H}} \max_{1\leqslant i\leqslant k} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i}\left[\ell\big(h,(x,y)\big)\right] + \varepsilon,$$

provided that the total sample size exceeds

$$\frac{d+k}{\varepsilon^2}\mathrm{poly}\log\Big(k,d,\frac{1}{\varepsilon},\frac{1}{\delta}\Big). \tag{4}$$

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; see Section 7 for details.

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 $[N] \coloneqq \{1,\dots,N\}$ for any positive integer N. Let $\mathsf{conv}(\mathcal{A})$ 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 $v=[v_i]_{1\leqslant i\leqslant n}$ and $v'=[v_i']_{1\leqslant i\leqslant n}$ with the same dimension, we overload the notation by using $\max\{v,v'\} = \big[\max\{v_i,v_i'\}\big]_{1\leqslant i\leqslant n}$ to denote the coordinate-wise maximum of v and v'. Also we say $v\leqslant v'$ iff $v_i\leqslant v_i'$ for all $i\in[n]$. For any random variable X, we use $\mathsf{Var}[X]$ to denote its variance,

[4] Following the definition of the VC dimension, the hypotheses in $\mathcal{H}$ are assumed to be binary-valued for VC classes.

i.e., $\mathsf{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$. Let $\mathrm{e}_1^{\mathsf{basis}}, \dots, \mathrm{e}_k^{\mathsf{basis}}$ represent the standard basis vectors in $\mathbb{R}^k$. 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

$$\mathsf{KL}(P \,\|\, Q) := \mathbb{E}_Q\left[\frac{\mathrm{d}P}{\mathrm{d}Q} \log \frac{\mathrm{d}P}{\mathrm{d}Q}\right]. \tag{5}$$

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. When it comes to a VC class, we assume that the hypotheses in $\mathcal{H}$ are binary-valued, and that the VC dimension (Vapnik et al., 1994) of $\mathcal{H}$ is
$$\mathsf{VC\text{-}dim}(\mathcal{H}) = d. \tag{6}$$
- *Loss function.* Suppose we are given a loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [-1, 1]$, 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

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

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 $\mathcal{D}_{i_\tau}$.

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 for learning VC classes. Before proceeding, we find it convenient to introduce some notation concerning the loss under mixed distributions. Specifically, for any distribution $w = [w_i]_{1 \leqslant i \leqslant k} \in \Delta(k)$ 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:

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

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

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

Algorithm 1: Hedge for multi-distribution learning on VC classes (`MDL-Hedge-VC`)

1 **input:** k data distributions $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k\}$, 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 $T = \frac{20000 \log\left(\frac{k}{\delta}\right)}{\varepsilon^2}$, 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 $W_i^1 = 1$, $\widehat{w}_i^0 = 0$ and $n_i^0 = 0$; $\mathcal{S} = \varnothing$.

4 **for** $t = 1, 2, \ldots, T$ **do**

5   set $w^t = [w_i^t]_{1 \leqslant i \leqslant k}$ and $\widehat{w}^t = [\widehat{w}_i^t]_{1 \leqslant i \leqslant k}$, with $w_i^t \leftarrow \frac{W_i^t}{\sum_j W_j^t}$ and $\widehat{w}_i^t \leftarrow \widehat{w}_i^{t-1}$ for all $i \in [k]$.

  `/* recompute` $\widehat{w}^t$ `& draw new samples for` $\mathcal{S}$ `only if` w^t `changes sufficiently. */`

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

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

8   **for** $i = 1, \ldots, k$ **do**

9       $n_i^t \leftarrow \lceil T_1 \widehat{w}_i^t \rceil$;

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

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

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

$$\widehat{L}^t(h, w^t) := \sum_{i=1}^{k} \frac{w_i^t}{n_i^t} \cdot \sum_{j=1}^{n_i^t} \ell\big(h, (x_{i,j}, y_{i,j})\big) \tag{9}$$

  with $(x_{i,j}, y_{i,j})$ being the j-th datapoint from $\mathcal{D}_i$ in $\mathcal{S}$.

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

12   $\overline{w}_i^t \leftarrow \max_{1 \leqslant \tau \leqslant t} w_i^\tau$ for all $i \in [k]$.

13   **for** $i = 1, \ldots, k$ **do**

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

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

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

16 **output:** a randomized hypothesis h^{final} uniformly distributed over $\{h^t\}_{t=1}^T$.

Remark 1. *Note that in Algorithm 1, the quantities $\{n_i^t\}_{1 \leqslant t \leqslant T}$ are designed to be non-decreasing in t.*

Remark 2. *In line 11 of Algorithm 1, we also allow h^t to be ε_1-best response that obeys $\widehat{L}^t(h^t, w^t) \leqslant \min_{h \in \mathcal{H}} \widehat{L}^t(h, w^t) + \varepsilon_1$; our theory remains unchanged if we make this modification.*

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,

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

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,[5] in an attempt to achieve low regret even when the loss vectors are adversarially chosen. More precisely, we run

$$w_i^{t+1} \propto w_i^t \exp\left(\eta \widehat{r}_i^t\right), \qquad i \in [k], \tag{11}$$

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

$$\widehat{r}_i^t \approx \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i}\left[\ell\big(h^t, (x,y)\big)\right], \qquad i \in [k],$$

computed over another set of data samples.

In words, the min-player and the max-player in our algorithm follow the best-response dynamics and no-regret dynamics, respectively (note that the Hedge algorithm is known to be a no-regret algorithm (Roughgarden, 2016)). At the end of the algorithm, we output a randomized hypothesis h^{final} that is uniformly distributed over the hypothesis iterates $\{h^t\}_{1\leqslant t\leqslant T}$ 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) — which also alternates between best response and no-regret algorithm — 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 one of the main sources 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 $\widehat{L}^t(h, w^t)$ for each $h \in \mathcal{H}$ in each round t, with the aim of achieving uniform convergence $|\widehat{L}^t(h, w^t) - L(h, w^t)| \leqslant O(\varepsilon)$ 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 $\mathbb{E}_{(x,y)\in\mathcal{D}_i}\big[\ell(h^t, (x,y))\big]$, where the number of samples depends upon the running weights $\{w_i^\tau\}$. More precisely, we collect $\lceil k\overline{w}_i^t \rceil$ fresh samples from each $\mathcal{D}_i$, where $\overline{w}_i^t := \max_{1\leqslant\tau\leqslant t} w_i^\tau$ is the maximum weight assigned to $\mathcal{D}_i$ up to now. Informally speaking, this strategy is chosen carefully to ensure reduced variance of the estimators given a sample size budget. 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.

[5] Note 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 novelties

In this section, we highlight two technical novelties that empower 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 $\|\cdot\|_{1,\infty}$ norm of the iterates $\{w^t\}_{1\leqslant t\leqslant T}$ that dictates the sample efficiency (see Section 4.2).

4.1 Towards sample reuse: uniform concentration and a key quantity

Recall that in Algorithm 1, we invoke the empirical risk estimator $\widehat{L}^t(h, w^t)$ 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,

$$\widehat{L}_n(h, w) \coloneqq \sum\nolimits_{i=1}^k w_i \cdot \frac{1}{n_i} \sum\nolimits_{j=1}^{n_i} \ell\big(h, (x_{i,j}, y_{i,j})\big), \tag{12}$$

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

$$\mathsf{Var}\big(\widehat{L}_n(h, w)\big) \leqslant \sum\nolimits_{i=1}^k \frac{w_i^2}{n_i} \leqslant \Big(\sum\nolimits_{i=1}^k w_i\Big) \frac{1}{\min_i n_i/w_i} = \frac{1}{\min_i n_i/w_i}.$$

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

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

Armed with this result, we can extend it to accommodate all $h \in \mathcal{H}$ through the union bound. For a VC class with $\mathsf{VC\text{-}dim}(\mathcal{H}) = d$, 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 $\exp\big(\widetilde{O}(d)\big)$. Taking the union bound then yields

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

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 $\exp\big(\widetilde{O}(k)\big)$. In addition, given the trivial upper bound $n_i \leqslant T_1$ for all $i \in [k]$, we know that there exist at most $T_1^k = \exp\big(\widetilde{O}(k)\big)$ possible combinations of $\{n_i\}_{i\in[k]}$. We can then apply the union bound to show that

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

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 $w \in \Delta(k)$ of interest.

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 $\widehat{L}_n(h, w)$ (see (12)) of $L(h, w)$ for all $h \in \mathcal{H}$ and all $w \in \Delta(k)$ with high probability, provided that

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

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) to guaranteed that

$$\min_i \frac{n_i^t}{w_i^t} \gtrsim T_1 \geqslant \widetilde{\Omega}\left(\frac{k+d}{\varepsilon^2}\right) \tag{15}$$

for all $1 \leqslant t \leqslant T$. 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 (note that the latter approach clearly falls short of data efficiency). To conclude, to guarantee ε-uniform convergence for all rounds, it suffices to choose $T_1 = \widetilde{\Omega}\big(\frac{k+d}{\varepsilon^2}\big)$.

Finally, recall that $n_i^t \asymp T_1 \overline{w}_i^t$ for each $i \in [k]$ and $t \leqslant T$, with $\overline{w}_i^t := \max_{1 \leqslant \tau \leqslant t} w_i^\tau$; taking $n_i^t \asymp T_1 \overline{w}_i^t$ (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

$$\frac{1}{T_1} \sum\nolimits_{i=1}^k n_i^T \asymp \sum\nolimits_{i=1}^k \overline{w}_i^T =: \|\overline{w}^T\|_1. \tag{16}$$

This threshold $\|\overline{w}^T\|_1$ — or equivalently, the $\|\cdot\|_{1,\infty}$ norm of $\{w^t\}_{1 \leqslant t \leqslant T}$ — is a critical quantity that we wish to control. In particular, in the desirable scenario where $\|\overline{w}^T\|_1 \leqslant \widetilde{O}(1)$, the total sample size obeys $\sum_{i=1}^k n_i^T \asymp T_1 \|\overline{w}^T\|_1 = \widetilde{O}\big(\frac{k+d}{\varepsilon^2}\big)$.

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

Perhaps the most innovative (and most challenging) part of our analysis lies in controlling the $\|\cdot\|_{1,\infty}$ norm of $\{w_i^t\}_{1 \leqslant t \leqslant T}$, 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, let us look at a simpler minimax optimization problem w.r.t. the set of loss vectors in the convex hull of a set $\mathcal{Y} \subseteq \mathbb{R}^k$:

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

where the equivalence arises from von Neumann's minimax theorem (v. Neumann, 1928). Consider the following algorithm (cf. Algorithm 2) tailored to this minimax problem, assuming perfect knowledge about the loss vectors.[6]

Algorithm 2: The Hedge algorithm for bilinear games.

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

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

[6] Note 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).

Question: *can we bound* $\|\overline{w}^T\|_1 := \sum_{i=1}^k \max_{1 \leqslant t \leqslant T} w_i^t$ *in Algorithm 2 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.

4.2.1 First attempt: bounding the number of distributions with $\max_{1 \leqslant t \leqslant T} w_i^t = \Omega(1)$

Instead of bounding $\|\overline{w}^T\|_1$ directly, our first attempt is to look at those $i \in [k]$ with large $\max_{1 \leqslant t \leqslant T} w_i^t$ (more specifically, $\max_{1 \leqslant t \leqslant T} w_i^t \geqslant 1/4$) and show that:

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

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

$$\mathcal{W}_{\mathsf{large}} := \big\{ i \in [k] \mid \max_{1 \leqslant t \leqslant T} w_i^t \geqslant 1/4 \big\}. \tag{18}$$

To do so, note that some standard "continuity"-type argument tells us that: for a sufficiently small stepsize η, one can find, for each $i \in \mathcal{W}_{\mathsf{large}}$, a time interval $[s_i, e_i] \subseteq [0, T]$ obeying

$$1/16 \leqslant w_i^{s_i} \leqslant 1/8, \qquad w_i^{e_i} \geqslant 1/4, \qquad \text{and} \qquad w_i^t \geqslant 1/8 \quad \forall t \in (s_i, e_i]. \tag{19}$$

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

$$e_i - s_i \geqslant \Omega(1/\eta^2) = \Omega(1/\varepsilon^2) \qquad \forall i \in \mathcal{W}_{\mathsf{large}}. \tag{20}$$

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

$$8T \geqslant \sum_{i \in \mathcal{W}_{\mathsf{large}}} (e_i - s_i) \geqslant |\mathcal{W}_{\mathsf{large}}| \cdot \Omega\left(1/\varepsilon^2\right), \tag{21}$$

which combined with our choice of $T = \widetilde{O}(1/\varepsilon^2)$ (cf. line 2 of Algorithm 1) yields

$$|\mathcal{W}_{\mathsf{large}}| \leqslant O(T\varepsilon^2) = \widetilde{O}(1).$$

In words, the number of distributions with large $\max_{1 \leqslant t \leqslant T} w_i^t$ (i.e., $\max_{1 \leqslant t \leqslant T} w_i^t \geqslant 1/4$) is fairly small.

Proof sketch for (20). Let us briefly discuss the high-level proof ideas. Following standard analysis for the Hedge algorithm (e.g., Shalev-Shwartz (2012); Lattimore and Szepesvári (2020)), we can often obtain (under certain mild conditions)

$$\mathsf{KL}\big(w^{s_i} \,\|\, w^{e_i}\big) \leqslant O\big(\eta^2 (e_i - s_i)\big).$$

Combine this relation with basic properties about the KL divergence (see, e.g., Lemmas 10 and 11 in Appendix A) and the choice $\eta = \varepsilon/10$ to obtain

$$\begin{aligned} e_i - s_i &\geqslant \Omega\big(\eta^{-2} \mathsf{KL}\big(w^{s_i} \,\|\, w^{e_i}\big)\big) \geqslant \Omega\Big(\eta^{-2} \mathsf{KL}\big(\mathsf{Ber}(w_i^{s_i}) \,\|\, \mathsf{Ber}(w_i^{e_i})\big)\Big) \\ &\geqslant \Omega\left(\eta^{-2} \frac{\frac{1}{16} \cdot 2^2}{4}\right) = \Omega\left(1/\eta^2\right) = \Omega\left(1/\varepsilon^2\right). \end{aligned}$$

$\square$

4.2.2 More general cases: issues and solutions

Naturally, one would hope to generalize the arguments in Section 4.2.1 to cope with more general cases. More specifically, let us look at the following set

$$\mathcal{W}(p) := \left\{ i \in [k] \mid \max_{1 \leqslant t \leqslant T} w_i^t \in [2p, 4p] \right\}, \tag{22}$$

defined for each $p \in [0, 1]$. If one could show that

$$|\mathcal{W}(p)| = \widetilde{O}(1/p) \qquad \text{for each } p, \tag{23}$$

then a standard doubling argument would immediately lead to

$$\begin{aligned} \|\overline{w}^T\|_1 &= \sum_{i \in [k]} \max_{1 \leqslant t \leqslant T} w_i^t \approx \sum_{j=1}^{\log_2 k} \sum_{i \in \mathcal{W}(2^{-j})} \max_{1 \leqslant t \leqslant T} w_i^t \leqslant \sum_{j=1}^{\log_2 k} 4 \cdot 2^{-j} \cdot \left|\mathcal{W}(2^{-j})\right| \\ &\leqslant \sum_{j=1}^{\log_2 k} 4 \cdot 2^{-j} \cdot \widetilde{O}\left(\frac{1}{2^{-j}}\right) = \widetilde{O}(1). \end{aligned}$$

A technical issue. Nevertheless, simply repeating the arguments in Section 4.2.1 fails to deliver the desirable bound (23) on $|\mathcal{W}(p)|$ when p is small. Briefly speaking, for each $i \in \mathcal{W}(p)$, let $[s_i, e_i]$ represent a time interval (akin to (19)) such that

$$p/2 \leqslant w_i^{s_i} \leqslant p, \qquad w_i^{e_i} \geqslant 2p \qquad \text{and} \qquad w_i^t \geqslant p \quad \text{for any } s_i < t \leqslant e_i. \tag{24}$$

Repeating the heuristic arguments in Section 4.2.1 leads to

$$e_i - s_i \geqslant \Omega\big(\eta^{-2} \mathsf{KL}\big(w^{s_i} \,\|\, w^{e_i}\big)\big) \geqslant \Omega\Big(\eta^{-2} \mathsf{KL}\big(\mathsf{Ber}(w_i^{s_i}) \,\|\, \mathsf{Ber}(w_i^{e_i})\big)\Big) \geqslant \Omega(\eta^{-2} p) = \Omega\left(p/\varepsilon^2\right). \tag{25}$$

Given that each t is contained within at most $1/(p/2)$ intervals associated with $\mathcal{W}(p)$, repeat the arguments for (21) to derive

$$\frac{1}{p/2} \cdot T \geqslant \sum_{i \in \mathcal{W}(p)} (e_i - s_i) \geqslant |\mathcal{W}(p)| \cdot \Omega\left(p/\varepsilon^2\right) \qquad \Longrightarrow \qquad |\mathcal{W}(p)| \leqslant \widetilde{O}\left(\frac{1}{p^2}\right).$$

This bound, however, is clearly loose compared to the desirable one in (23).

Our solution. To address this issue, we make two key observations below that inspire our approach:

- *Shared intervals.* For the interval $[s_i, e_i]$ associated with i as defined above, if there exist other indices sharing the same interval $[s_i, e_i]$ (in the sense that relations analogous to (24) are satisfied for other indices), then it is plausible to improve the bound. For instance, if a set $\mathcal{M}_i \subseteq [k]$ of indices share the same interval $[s_i, e_i]$, then one can follow the heuristic argument in (25) to obtain

$$e_i - s_i \geqslant \Omega\big(\eta^{-2} \mathsf{KL}\big(w^{s_i} \,\|\, w^{e_i}\big)\big) \geqslant \Omega\left(\eta^{-2} \mathsf{KL}\left(\mathsf{Ber}\Big(\sum_{j \in \mathcal{M}_i} w_j^{s_i}\Big) \,\Big\|\, \mathsf{Ber}\Big(\sum_{j \in \mathcal{M}_i} w_j^{e_i}\Big)\right)\right) \geqslant \Omega\left(p|\mathcal{M}_i|/\varepsilon^2\right), \tag{26}$$

  which clearly strengthens the original bound (25) if $|\mathcal{M}_i|$ is large.

- *Disjoint intervals.* Consider the special case where $\mathcal{W}(p)$ can be divided into subsets $\{\mathcal{V}_n\}_{n=1}^N$ obeying
  - (i) for each $n \in [N]$, all indices in $\mathcal{V}_n$ share the same interval $[s_n, e_n]$ (defined analogously as (24));
  - (ii) the intervals $\{[s_n, e_n]\}_{n=1}^N$ are *disjoint*.

Then one can derive the desired bound on $|\mathcal{W}(p)|$. More precisely, it follows from (26) that

$$e_n - s_n \geqslant \Omega\left(p|\mathcal{V}_n|/\varepsilon^2\right), \tag{27}$$

which together with the disjointness property yields

$$|\mathcal{W}_j| = \sum_{n=1}^{N} |\mathcal{V}_n| \leqslant \sum_{n=1}^{N} O\left(\frac{(e_n - s_n)\varepsilon^2}{p}\right) \leqslant O\left(\frac{T\varepsilon^2}{p}\right) = \widetilde{O}\left(\frac{1}{p}\right). \tag{28}$$

In light of the above discussion, it is helpful to (a) merge those indices that share similar intervals, and (b) identify disjoint intervals whose associated indices can cover a good fraction of $\mathcal{W}_j$.

Motivated by the aforementioned observation about "shared intervals," we introduce the notion of "*segments*" to facilitate analysis.

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) $\sum_{i\in\mathcal{I}} w_i^{t_2} \geqslant p\exp(x)$,

(iii) $\sum_{i\in\mathcal{I}} w_i^{t} \geqslant q$ *for any* $t_1 \leqslant t \leqslant t_2$.

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

This definition allows one to pool indices with similar intervals together.

In general, however, it is common for two segments to be overlapping (see Figure 2 in Appendix C.5), which precludes us from directly invoking our aforementioned observation about "disjoint intervals." To address this issue, our strategy is to extract out shared sub-segments[7] of (a subset of) these segments in a meticulous manner. Encouragingly, it is possible to find such sub-segments that taken collectively cover a good fraction (i.e., $\frac{1}{\mathrm{poly}\log(k,T)}$) of all segments, meaning that we do not have to discard too many segments. The construction is built upon careful analysis of these segments, and will be elucidated in Appendix C.

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 $h^t = \arg\min_{h\in\mathcal{H}} \widehat{L}^t(h, w^t)$ (cf. line 11 of Algorithm 1); in words, h^t is the minimizer of the empirical loss function $\widehat{L}^t(\cdot, w^t)$, 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; here, we recall that $\varepsilon_1 = \varepsilon/100$ (cf. line 2 of Algorithm 1).

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

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

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

Proof. See Appendix B.1. ☐

[7] A sub-segment refers to a sub-interval of a segment, as illustrated in Figure 6.

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 $L(h^t, w^t) \leqslant \min_{h\in\mathcal{H}} L(h, w^t) + \varepsilon_1$ in the t-th round for each $1 \leqslant t \leqslant T$. With probability exceeding $1 - \delta/4$, the hypothesis h^{final} output by Algorithm 1 is ε-optimal in the sense that*

$$\max_{1\leqslant i\leqslant k} L(h^{\mathsf{final}}, \mathrm{e}_i^{\mathsf{basis}}) \leqslant \min_{h\in\mathcal{H}} \max_{1\leqslant i\leqslant k} L(h, \mathrm{e}_i^{\mathsf{basis}}) + \varepsilon. \tag{30}$$

Here, we recall that $\mathrm{e}_i^{\mathsf{basis}}$ indicates the i-th standard basis vector.

Proof. See Appendix B.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 in Algorithm 1. Towards this end, recall that $\overline{w}_i^T = \max_{1\leqslant t\leqslant T} w_i^t$ for each $i \in [k]$. Recognizing that $\widehat{w}_i^t \leqslant \overline{w}_i^t$ for each $t \in [T]$ and $i \in [k]$, we can bound the total sample size by

$$\begin{aligned}
\text{(sample size)} \qquad T_1 \sum_{i=1}^{k} \widehat{w}_i^T + k + T\left(k \sum_{i=1}^{k} \overline{w}_i^T + k\right) &\leqslant \left(T_1 \|\overline{w}^T\|_1 + kT\|\overline{w}^T\|_1\right) + k(T+1) \\
&\lesssim \frac{d \log\left(\frac{kd}{\varepsilon}\right) + k \log\left(\frac{k}{\delta\varepsilon}\right)}{\varepsilon^2} \cdot \|\overline{w}^T\|_1,
\end{aligned} \tag{31}$$

where the last relation follows from our choices $T \asymp \frac{\log(k/\delta)}{\varepsilon^2}$ and $T_1 \asymp \frac{k\log(k/\varepsilon) + d\log(kd/\varepsilon) + \log(1/\delta)}{\varepsilon^2}$ (cf. line 2 of Algorithm 1) and the basic property that $\|\overline{w}^T\|_1 \geqslant \sum_i w_i^1 = 1$. Consequently, everything then comes down to bounding $\|\overline{w}^T\|_1$, for which we resort to the following lemma.

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

$$\|\overline{w}^T\|_1 \leqslant O\left(\log^8\left(\frac{k}{\delta\varepsilon}\right)\right).$$

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

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

$$O\left(\frac{d \log\left(\frac{kd}{\varepsilon}\right) + k \log\left(\frac{k}{\delta\varepsilon}\right)}{\varepsilon^2} \cdot \log^8\left(\frac{k}{\delta\varepsilon}\right)\right),$$

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

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 $d \geqslant 2\log(8k)$. 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 $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [-1, 1]$,*

such that: for any algorithm $\mathcal{G}$ that finds $h \in \mathcal{H}$ obeying

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

with probability exceeding $3/4$, its sample size — denoted by $M(\mathcal{G})$ — must exceed $\mathbb{E}[M(\mathcal{G})] \geqslant \frac{dk}{240000\varepsilon^2}$.

In words, the sample complexity when a deterministic output is required scales at least with $\Omega\big(\frac{dk}{\varepsilon^2}\big)$, which is considerably larger than the sample complexity $\widetilde{O}\big(\frac{d+k}{\varepsilon^2}\big)$ achievable via Algorithm 1 (which outputs a randomized hypothesis). The proof of this theorem is deferred to Appendix D.

7 Extension: learning Rademacher classes

In this section, we adapt our algorithm and theory to accommodate MDL for Rademacher classes. Note that when learning Rademacher classes, the hypotheses in $\mathcal{H}$ do not need to be binary-valued.

7.1 Preliminaries: Rademacher complexity

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

Definition 2 (Rademacher complexity). *Given a distribution $\mathcal{D}$ supported on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ and a positive integer n, the Rademacher complexity is defined as*

$$\mathsf{Rad}_n(\mathcal{D}) := \mathop{\mathbb{E}}_{\{z_i\}_{i=1}^n} \left[\mathop{\mathbb{E}}_{\{\sigma_i\}_{i=1}^n} \left[\max_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(h, z_i) \right] \right], \tag{33}$$

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 $1 \leqslant i \leqslant n$.

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

$$\mathcal{D}(w) := \sum_{i=1}^k w_i \mathcal{D}_i \tag{34}$$

for any probability vector $w = [w_i]_{1\leqslant i\leqslant k} \in \Delta(k)$, we can state our assumption as follows.

Assumption 1. *For each $n \geqslant 1$, there exists a quantity $C_n > 0$ (known to the learner) such that*

$$C_n \geqslant \sup_{w\in\Delta(k)} \mathsf{Rad}_n\big(\mathcal{D}(w)\big). \tag{35}$$

For instance, if the hypothesis class $\mathcal{H}$ obeys $\mathsf{VC\text{-}dim}(\mathcal{H}) \leqslant d$, then it is well-known that Assumption 1 holds with the choice (see, e.g., Mohri et al. (2018))

$$C_n = \sqrt{\frac{2d\log(en/d)}{n}}.$$

Remark 3. *One might raise a natural question about Assumption 1: can we use* $\widetilde{C}_n := \max_{i\in[k]} \mathsf{Rad}_n(\mathcal{D}_i)$ *instead of* C_n *without incuring a worse sample complexity? The answer is, however, negative. In fact, the Rademacher complexity* $\mathsf{Rad}_n(\mathcal{D}(w))$ *is not convex in* w*, and hence we fail to use* $\max_i \mathsf{Rad}_n(\mathcal{D}_i)$ *to bound* $\max_{w\in\Delta(k)} \mathsf{Rad}_n(\mathcal{D}(w))$*. The interested reader is referred to Appendix E.4 for more details.*

To facilitate analysis, we find it helpful to introduce another notion called weighted Rademacher complexity.

Definition 3 (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*

$$\widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k} := \mathop{\mathbb{E}}_{\{z_i^j\}_{j=1}^{n_i}, \forall i\in[k]} \left[\mathop{\mathbb{E}}_{\{\sigma_i^j\}_{j=1}^{n_i}, \forall i\in[k]} \left[\frac{1}{\sum_{i=1}^k n_i} \max_{h\in\mathcal{H}} \sum_{i=1}^k \sum_{j=1}^{n_i} \sigma_i^j \ell(h, z_i^j) \right] \right], \tag{36}$$

where $\big\{\{z_i^j\}_{j=1}^{n_k}\big\}_{i=1}^k$ *are independently generated with each* z_i^j *drawn from* $\mathcal{D}_i$*, and* $\big\{\{\sigma_i^j\}_{j=1}^{n_i}\big\}_{i=1}^k$ *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* $\widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k} = \widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k}(\mathcal{D})$*.*

The following two lemmas provide useful properties about the weighted Rademacher complexity.

Lemma 4. *For any two groups of positive integers* $\{n_i\}_{i=1}^k$ *and* $\{m_i\}_{i=1}^k$*, it holds that*

$$\begin{aligned}\left(\sum_{i=1}^k n_i\right) \widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k} &\leqslant \left(\sum_{i=1}^k (m_i + n_i)\right) \widetilde{\mathsf{Rad}}_{\{m_i+n_i\}_{i=1}^k} \\ &\leqslant \left(\sum_{i=1}^k n_i\right) \widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k} + \left(\sum_{i=1}^k m_i\right) \widetilde{\mathsf{Rad}}_{\{m_i\}_{i=1}^k}.\end{aligned} \tag{37}$$

Proof. See Appendix E.2. □

Lemma 5. *Consider any* $\{n_i\}_{i=1}^k$ *obeying* $n_i \geqslant 12\log(2k)$ *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*

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

Proof. See Appendix E.3. □

7.2 Algorithm and sample complexity

We are now positioned to introduce our algorithm that learns a Rademacher class in the presence of multiple distributions, which is also based on a Hedge-type strategy to learn a convex-concave game; see Algorithm 3 for full details. Its main distinction from Algorithm 1 lies in the subroutine to learn h^t (see lines 7-12 in Algorithm 3) as well as the choice of T_1 (see line 2 in Algorithm 3). More precisely, to compute the estimator $\widehat{L}^t(h, w^t)$ 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

$$n_i^{t,\mathsf{rad}} = \min\left\{ \lceil T_1 w_i^t + 12\log(2k) \rceil, T_1 \right\}$$

samples from $\mathcal{D}_i$ for each i, where T_1 is taken to be

$$T_1 = \min\left\{ t \geqslant \frac{4000(k\log(k/\varepsilon_1) + \log(1/\delta))}{\varepsilon_1^2} \,\Big|\, C_t \leqslant \frac{\varepsilon_1}{4800} \right\}. \tag{38}$$

Here, T_1 needs to take advantage of the quantities $\{C_n\}$ (cf. Assumption 1) that upper bound the associated Rademacher complexity.

Equipped with this algorithm, we are ready to establish the following theoretical guarantees.

Algorithm 3: Hedge for multi-distribution learning on Rademacher Classes (`MDL-Hedge-Rad`)

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 $1-\delta$, quantities $\{C_n\}_{n\geqslant 1}$ as in Assumption 1.

hyper-parameters: stepsize $\eta = \frac{1}{100}\varepsilon$, number of rounds $T = \frac{20000\log\left(\frac{k}{\delta}\right)}{\varepsilon^2}$, auxiliary accuracy level $\varepsilon_1 = \frac{1}{100}\varepsilon$, auxiliary sub-sample-size $T_1 = \min\left\{t \geqslant \frac{4000(k\log(k/\varepsilon_1)+\log(1/\delta))}{\varepsilon_1^2} \,\middle|\, C_t \leqslant \frac{\varepsilon_1}{4800}\right\}$.

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

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

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

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

`/* recompute` $\widehat{w}^t$ `& draw new samples for` $\mathcal{S}_{\mathrm{w}}$ `only if` w^t `changes sufficiently. */`

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

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

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

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

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

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

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

$$\widehat{L}^t(h, w^t) := \sum_{i=1}^{k} \frac{w_i^t}{n_i^{t,\mathsf{rad}}} \cdot \sum_{j=1}^{n_i^{t,\mathsf{rad}}} \ell\big(h, (x_{i,j}, y_{i,j})\big) \tag{39}$$

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

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

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

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

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

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

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

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

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

$$\max_{1\leqslant i\leqslant k} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i, h^{\mathsf{final}}}\left[\ell\big(h^{\mathsf{final}}, (x,y)\big)\right] \leqslant \min_{h\in\mathcal{H}} \max_{1\leqslant i\leqslant k} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i}\left[\ell\big(h, (x,y)\big)\right] + \varepsilon. \tag{40}$$

In particular, the total number of samples collected by Algorithm 3 is bounded by

$$\left(\frac{k}{\varepsilon^2} + \min\left\{n \mid C_n \leqslant c_1\varepsilon\right\}\right) \operatorname{poly}\log\left(k, \frac{1}{\varepsilon}, \frac{1}{\delta}\right)$$

with probability exceeding $1-\delta$, where $c_1 > 0$ is some sufficiently small constant, and T_1 is defined in (38).

In contrast to the VC classes, the sample complexity for learning Rademacher classes entails a term related to the Rademacher complexity — namely, $\min\left\{n \mid C_n \leqslant c_1\varepsilon\right\}$ — as opposed to the term d/ε^2 concerning the VC-dimension. When it comes to the special case where $\mathsf{VC\text{-}dim}(\mathcal{H}) \leqslant d$, taking $C_n \leqslant \sqrt{\frac{2d\log(en/d)}{n}}$ in (38) leads to $T_1 = \widetilde{O}\left(\frac{d+k}{\varepsilon^2}\right)$, which recovers the sample complexity bound of $\widetilde{O}\left(\frac{d+k}{\varepsilon^2}\right)$ derived in Theorem 1 for VC classes.

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

Lemma 6. *Suppose Assumption 1 holds. With probability at least $1-\delta/4$, the iterates of Algorithm 3 satisfy*

$$L(h^t, w^t) \leqslant \min_{h\in\mathcal{H}} L(h, w^t) + \varepsilon_1 \tag{41}$$

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

The proof of Lemma 6 is postponed to Appendix E. Combine this with Lemma 2 and Lemma 3 to show that: the total number of samples collected in Algorithm 3 is upper bounded by

$$T_1 \operatorname{poly}\log\left(k, \frac{1}{\varepsilon}, \frac{1}{\delta}\right) \leqslant \left(\frac{k}{\varepsilon^2} + \min\left\{n \mid C_n \leqslant \frac{1}{4800}\varepsilon\right\}\right) \operatorname{poly}\log\left(k, \frac{1}{\varepsilon}, \frac{1}{\delta}\right)$$

as claimed. □

8 Discussion

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

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

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 martingales.

Lemma 7 (Freedman's inequality (Freedman, 1975))**.** *Let $(M_n)_{n\geqslant 0}$ be a martingale obeying $M_0 = 0$. Define $V_n := \sum_{k=1}^{n} \mathbb{E}[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}]$ for each $n \geqslant 0$, where $\mathcal{F}_k$ denotes the σ-algebra generated by $(M_1, M_2, \ldots, M_k)$. Suppose that $M_k - M_{k-1} \leqslant 1$ for all $k \geqslant 1$. Then for any $x > 0$ and $y > 0$, one has*

$$\mathbb{P}\big(M_n \geqslant nx, V_n \leqslant ny\big) \leqslant \exp\left(-\frac{nx^2}{2(y + \frac{1}{3}x)}\right). \tag{42}$$

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

$$\mathbb{P}\left\{\exists n \in \mathbb{N}, \sum_{k=1}^{n} X_k \geqslant 3\sum_{k=1}^{n} Y_k + l \log\frac{1}{\delta}\right\} \leqslant \delta,$$
$$\mathbb{P}\left\{\exists n \in \mathbb{N}, \sum_{k=1}^{n} Y_k \geqslant 3\sum_{k=1}^{n} X_k + l \log\frac{1}{\delta}\right\} \leqslant \delta.$$

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

Lemma 9 (Mcdiarmid's inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables, with X_i supported on $\mathcal{X}_i$. Let $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ be a function such that: for any $i \in [n]$ and any $\{x_1, \ldots, x_n\} \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$,*

$$\sup_{x_i' \in \mathcal{X}_i} \big|f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)\big| \leqslant c$$

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

$$\mathbb{P}\Big\{\big|f(X_1, X_2, \cdots, X_n) - \mathbb{E}\big[f(X_1, X_2, \cdots, X_n)\big]\big| \geqslant \varepsilon\Big\} \leqslant 2\exp\left(-\frac{2\varepsilon^2}{nc^2}\right).$$

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

$$\mathsf{KL}\big(\mu(X_1) \,\|\, \mu(X_2)\big) \geqslant \mathsf{KL}\big(\mu\big(f(X_1)\big) \,\|\, \mu\big(f(X_2)\big)\big), \tag{43}$$

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

$$\mathsf{KL}\left(\mathsf{Ber}(y) \,\|\, \mathsf{Ber}(y')\right) \geqslant \frac{qx^2}{4},$$

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

Proof. To begin with, the function defined below satisfies

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

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

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

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

$$\begin{aligned}
\mathsf{KL}\left(\mathsf{Ber}(y) \,\|\, \mathsf{Ber}(y')\right) &= f(y,y') = f(y,y) + \int_y^{y'} \frac{\partial f(y,z)}{\partial z}\mathrm{d}z = \int_y^{y'} \frac{z-y}{z(1-z)}\mathrm{d}z \\
&\geqslant \frac{1}{y'}\int_y^{y'} (z-y)\mathrm{d}z \geqslant \frac{(y'-y)^2}{2y'} \\
&\geqslant \frac{(y'-y)(1-\exp(-x))}{2} \\
&\geqslant \frac{y(\exp(x)-1)^2}{4} \geqslant \frac{qx^2}{4},
\end{aligned}$$

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

Finally, let us present a basic property related to Rademacher random variables, which will play a useful role in understanding the Rademacher complexity.

Lemma 12. *Let $\mathcal{L}$ be a set of vectors in $\mathbb{R}^n$. Let $w^1, w^2 \in \mathbb{R}^n$ be two vectors obeying $|w_i^1| \leqslant |w_i^2|$ for all $i \in [n]$. Then it holds that*

$$\mathop{\mathbb{E}}_{\{\sigma_i\}_{i=1}^n}\left[\max_{f\in\mathcal{L}} \sum_{i=1}^n \sigma_i w_i^1 f_i\right] \leqslant \mathop{\mathbb{E}}_{\{\sigma_i\}_{i=1}^n}\left[\max_{f\in\mathcal{L}} \sum_{i=1}^n \sigma_i w_i^2 f_i\right], \tag{44}$$

where $\{\sigma_i\}$ is a collection of independent Rademacher random variables obeying $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$.

Proof. Clearly, it suffices to prove (44) for the special case where $w_i^1 = w_i^2$ for $1 \leqslant i \leqslant n-1$, and $|w_n^1| \leqslant |w_n^2|$. Fixing σ_i for $1 \leqslant i \leqslant n-1$, we can deduce that

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

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

B Proofs of auxiliary lemmas for VC classes

B.1 Proof of Lemma 1

For ease of presentation, suppose there exists a dataset $\widetilde{\mathcal{S}}$ containing T_1 independent samples drawn from each distribution $\mathcal{D}_i$ $(1 \leqslant i \leqslant k)$, 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 $\widetilde{\mathcal{S}}$ that is drawn from $\mathcal{D}_i$.
- For each set of integers $n = \{n_i\}_{i=1}^k \in \mathbb{N}^k$, we define $\widetilde{\mathcal{S}}(n)$ to be the dataset containing $\big\{(x_{i,j}, y_{i,j})\big\}_{1 \leqslant j \leqslant n_i}$ for all $i \in [k]$; namely, it comprises, for each $i \in [k]$, the first n_i samples in $\widetilde{\mathcal{S}}$ that are drawn from $\mathcal{D}_i$.
- We shall also let $\widetilde{\mathcal{S}}^+(n) = \big\{\big\{\big(x^+_{i,j}, y^+_{i,j}\big)\big\}_{j=1}^{n_i}\big\}_{i=1}^k$ be an *independent copy* of $\widetilde{\mathcal{S}}(n)$, where for each $i \in [k]$, $\big\{\big(x^+_{i,j}, y^+_{i,j}\big)\big\}$ 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 $0 \leqslant n_i \leqslant T_1$ for all $i \in [k]$, and any fixed $w \in \Delta(k)$. For any quantity $\lambda \in \big[0, \min_{i \in [k]} \frac{n_i}{w_i}\big]$, if we take

$$E(\lambda, n, w) := \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}(n)}\left[\max_{h \in \mathcal{H}} \exp\left(\lambda\left\{\sum_{i=1}^k w_i \frac{1}{n_i}\sum_{i=1}^{n_i} \ell\big(h, (x_{i,j}, y_{i,j})\big) - L(h, w)\right\}\right)\right] \tag{45}$$

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

$$\begin{aligned}
E(\lambda, n, w) &:= \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}(n)}\left[\max_{h \in \mathcal{H}} \exp\left(\lambda\left\{\sum_{i=1}^k \frac{w_i}{n_i}\sum_{i=1}^{n_i} \ell\big(h, (x_{i,j}, y_{i,j})\big) - L(h, w)\right\}\right)\right] \\
&= \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}(n)}\left[\max_{h \in \mathcal{H}} \exp\left(\lambda\left\{\sum_{i=1}^k \frac{w_i}{n_i}\sum_{i=1}^{n_i} \ell\big(h, (x_{i,j}, y_{i,j})\big) - \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}^+(n)}\left[\sum_{i=1}^k \frac{w_i}{n_i}\sum_{i=1}^{n_i} \ell\big(h, (x^+_{i,j}, y^+_{i,j})\big)\right]\right\}\right)\right] \\
&\leqslant \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}(n)}\left[\max_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}^+(n)}\left[\exp\left(\lambda\left\{\sum_{i=1}^k \frac{w_i}{n_i}\sum_{j=1}^{n_i} \Big(\ell\big(h, (x_{i,j}, y_{i,j})\big) - \ell\big(h, (x^+_{i,j}, y^+_{i,j})\big)\Big)\right\}\right)\right]\right] \\
&\leqslant \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}(n), \widetilde{\mathcal{S}}^+(n)}\left[\max_{h \in \mathcal{H}} \exp\left(\lambda\left\{\sum_{i=1}^k \frac{w_i}{n_i}\sum_{j=1}^{n_i} \Big(\ell\big(h, (x_{i,j}, y_{i,j})\big) - \ell\big(h, (x^+_{i,j}, y^+_{i,j})\big)\Big)\right\}\right)\right],
\end{aligned} \tag{46}$$

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

Next, let $\sigma(n) := \big\{\{\sigma_{i,j}\}_{j=1}^{n_i}\big\}_{i=1}^k$ 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 $\mathcal{C} = \big\{(x_{i,j}, y_{i,j})\big\} \bigcup \big\{(x^+_{i,j}, y^+_{i,j})\big\}$, we obtain

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}(n), \widetilde{\mathcal{S}}^+(n)}\left[\max_{h \in \mathcal{H}} \exp\left(\lambda\left\{\sum_{i=1}^k \frac{w_i}{n_i}\sum_{j=1}^{n_i} \Big(\ell\big(h, (x_{i,j}, y_{i,j})\big) - \ell\big(h, (x^+_{i,j}, y^+_{i,j})\big)\Big)\right\}\right)\right] \\
&= \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}(n), \widetilde{\mathcal{S}}^+(n)}\left[\mathop{\mathbb{E}}_{\sigma(n)}\left[\max_{h \in \mathcal{H}} \exp\left(\lambda\left\{\sum_{i=1}^k \frac{w_i}{n_i}\sum_{j=1}^{n_i} \sigma_{i,j}\Big(\ell\big(h, (x_{i,j}, y_{i,j})\big) - \ell\big(h, (x^+_{i,j}, y^+_{i,j})\big)\Big)\right\}\right) \,\Big|\, \mathcal{C}\right]\right].
\end{aligned} \tag{47}$$

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 $\mathsf{VC\text{-}dim}(\mathcal{H}) \leqslant d$ tells us that the cardinality of the following set obeys

$$\big|\mathcal{H}(\mathcal{C})\big| \leqslant (|\mathcal{C}| + 1)^d \leqslant \big(|\widetilde{\mathcal{S}}| + |\widetilde{\mathcal{S}}^+| + 1\big)^d \leqslant (2kT_1 + 1)^d, \tag{48}$$

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

$$\mathcal{H}(\mathcal{C}) := \Big\{ \big(h(x_{1,1}), h(x_{1,1}^{+}), h(x_{1,2}), h(x_{1,2}^{+}), \cdots \big) \mid h \in \mathcal{H} \Big\}. \tag{49}$$

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,

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

With these in place, we can demonstrate that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\sigma(n)} \left[\max_{h\in\mathcal{H}} \exp\left(\lambda \left\{ \sum_{i=1}^{k} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \sigma_{i,j} \Big(\ell\big(h,(x_{i,j},y_{i,j})\big) - \ell\big(h,(x_{i,j}^{+},y_{i,j}^{+})\big) \Big) \right\} \right) \,\Big|\, \mathcal{C} \right] \\
&= \mathop{\mathbb{E}}_{\sigma(n)} \left[\max_{h\in\mathcal{H}_{\min,\mathcal{C}}} \exp\left(\lambda \left\{ \sum_{i=1}^{k} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \sigma_{i,j} \Big(\ell\big(h,(x_{i,j},y_{i,j})\big) - \ell\big(h,(x_{i,j}^{+},y_{i,j}^{+})\big) \Big) \right\} \right) \,\Big|\, \mathcal{C} \right] \\
&\leqslant \mathop{\mathbb{E}}_{\sigma(n)} \left[\sum_{h\in\mathcal{H}_{\min,\mathcal{C}}} \exp\left(\lambda \left\{ \sum_{i=1}^{k} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \sigma_{i,j} \Big(\ell\big(h,(x_{i,j},y_{i,j})\big) - \ell\big(h,(x_{i,j}^{+},y_{i,j}^{+})\big) \Big) \right\} \right) \,\Big|\, \mathcal{C} \right] \\
&\leqslant \big|\mathcal{H}_{\min,\mathcal{C}}\big| \max_{h\in\mathcal{H}_{\min,\mathcal{C}}} \mathop{\mathbb{E}}_{\sigma(n)} \left[\exp\left(\lambda \left\{ \sum_{i=1}^{k} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \sigma_{i,j} \Big(\ell\big(h,(x_{i,j},y_{i,j})\big) - \ell\big(h,(x_{i,j}^{+},y_{i,j}^{+})\big) \Big) \right\} \right) \,\Big|\, \mathcal{C} \right] \\
&\leqslant \big(2kT_1+1\big)^d \max_{h\in\mathcal{H}} \prod_{i=1}^{k} \prod_{j=1}^{n_i} \mathop{\mathbb{E}}_{\sigma_{i,j}} \left[\exp\left(\lambda \left\{ \frac{w_i}{n_i} \sigma_{i,j} \Big(\ell\big(h,(x_{i,j},y_{i,j})\big) - \ell\big(h,(x_{i,j}^{+},y_{i,j}^{+})\big) \Big) \right\} \right) \,\Big|\, \mathcal{C} \right] \\
&\leqslant \big(2kT_1+1\big)^d \exp\left(2\lambda^2 \sum_{i=1}^{k} \frac{(w_i)^2}{n_i} \right).
\end{aligned} \tag{50}$$

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

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

Taking (46), (47) and (50) together reveals that

$$E(\lambda) \leqslant (2kT_1+1)^d \exp\left(2\lambda^2 \sum_{i=1}^{k} \frac{(w_i)^2}{n_i} \right). \tag{51}$$

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

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

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

$$\begin{aligned}
&\mathbb{P}\left(\max_{h\in\mathcal{H}} \left| \sum_{i=1}^{k} w_i \frac{1}{n_i} \sum_{i=1}^{n_i} \ell\big(h,(x_{i,j},y_{i,j})\big) - L(h,w) \right| \geqslant \varepsilon' \right) \\
&\qquad \leqslant \min_{0\leqslant\lambda\leqslant\min_i \frac{n_i}{w_i}} \frac{E(\lambda)+\overline{E}(\lambda)}{\exp(\lambda\varepsilon')} \\
&\qquad \leqslant \min_{0\leqslant\lambda\leqslant\min_i \frac{n_i}{w_i}} 2\cdot(2kT_1+1)^d \exp\left(2\lambda^2 \sum_{i=1}^{k} \frac{(w_i)^2}{n_i} - \lambda\varepsilon' \right).
\end{aligned} \tag{52}$$

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 $\|x - x_0\|_\infty \leqslant \varepsilon_2$. We shall choose $\Delta_{\varepsilon_2}(k)$ properly so that
$$|\Delta_{\varepsilon_2}(k)| \leqslant (1/\varepsilon_2)^k.$$
- Define the following set
$$\mathcal{B} = \left\{ n = \{n_i\}_{i=1}^k, w = \{w_i\}_{i=1}^k \,\middle|\, \frac{n_i}{w_i} \geqslant \frac{T_1}{2}, 0 \leqslant n_i \leqslant T_1, \forall i \in [k], w \in \Delta_{\varepsilon_1/(8k)}(k) \right\},$$
which clearly satisfies
$$|\mathcal{B}| \leqslant T_1^k \cdot \left(\frac{8k}{\varepsilon_1}\right)^k.$$

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

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

where the second inequality holds since $\sum_{i=1}^k \frac{w_i^2}{n_i} \leqslant \frac{2}{T_1} \sum_{i=1}^k w_i = \frac{2}{T_1}$ (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 $n^t = [n_i^t]_{1 \leqslant i \leqslant k}$ denotes the number of samples for all k distributions in $\mathcal{S}^t$, and let $w^t = [w_i^t]_{1 \leqslant i \leqslant k}$ represent the weight iterates in the t-th round. It is easily seen from lines 6 and 9 of Algorithm 1 that $n_i^t \leqslant T_1$ and $n_i^t / w_i^t \geqslant n_i^t / (2\widehat{w}_i^t) \geqslant T_1/2$. For analysis purposes, it suffices to take $\mathcal{S}^t = \widetilde{\mathcal{S}}(n^t)$.

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

$$\begin{aligned}
&\mathbb{P}\left(\exists t \in [T], \max_{h \in \mathcal{H}} \left| \sum_{i=1}^k w_i^t \frac{1}{n_i^t} \sum_{i=1}^{n_i^t} \ell\big(h, (x_{i,j}, y_{i,j})\big) - L(h, w^t) \right| \geqslant \varepsilon' + \frac{\varepsilon_1}{4} \right) \\
&\qquad \leqslant \mathbb{P}\left(\exists t \in [T], \max_{h \in \mathcal{H}} \left| \sum_{i=1}^k \widetilde{w}_i^t \frac{1}{n_i^t} \sum_{i=1}^{n_i^t} \ell\big(h, (x_{i,j}, y_{i,j})\big) - L(h, \widetilde{w}^t) \right| \geqslant \varepsilon' \right) \\
&\qquad \leqslant 2 \cdot (8kT_1/\varepsilon_1)^k (2kT_1 + 1)^d \cdot \exp\left(-\frac{T_1(\varepsilon')^2}{16} \right), \qquad (53)
\end{aligned}$$

where the second inequality arises from the fact that $\frac{1}{n_i}\sum_{i=1}^{n_i}\ell\big(h,(x_{i,j},y_{i,j})\big)\in[-1,1]$ and $L(h,\widetilde{w}^t)\in[-1,1]$. 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 (53), we can obtain

$$
\begin{aligned}
&\mathbb{P}\left(\exists t\in[T],\max_{h\in\mathcal{H}}\left|\sum_{i=1}^{k}w_i^t\cdot\frac{1}{n_i^t}\sum_{i=1}^{n_i^t}\ell\big(h,(x_{i,j},y_{i,j})\big)-L(h,w^t)\right|\geqslant\frac{\varepsilon_1}{2}\right)\\
&\leqslant 2\cdot(8kT_1/\varepsilon_1)^k(2kT_1+1)^d\cdot\exp\left(-\frac{T_1\varepsilon_1^2}{16}\right)\\
&\leqslant 2\cdot(8kT_1/\varepsilon_1)^k(2kT_1+1)^d\cdot\exp\Big(-10\big(k\log(k/\varepsilon_1)+d\log(kd/\varepsilon_1)+\log(1/\delta)\big)\Big)\\
&\leqslant 2\cdot(8kT_1/\varepsilon_1)^k(2kT_1+1)^d\cdot(k/\varepsilon_1)^{-10k}\cdot(kd/\varepsilon_1)^{-10d}\cdot\delta\\
&\leqslant\delta/4.
\end{aligned}
\tag{54}
$$

Step 4: putting all this together. Recalling that

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

one can see from (54) that, with probability exceeding $1-\delta/4$,

$$
\left|\widehat{L}^t(h,w^t)-L(h,w^t)\right|\leqslant\frac{\varepsilon_1}{2}
\tag{55}
$$

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

$$
h^t=\arg\min_{h\in\mathcal{H}}\widehat{L}^t(h,w^t),
\tag{56}
$$

we can immediately deduce that

$$
L(h^t,w^t)\leqslant\widehat{L}(h^t,w^t)+\frac{\varepsilon_1}{2}=\min_{h\in\mathcal{H}}\widehat{L}(h,w^t)+\frac{\varepsilon_1}{2}\leqslant\min_{h\in\mathcal{H}}L(h,w^t)+\varepsilon_1.
\tag{57}
$$

This concludes the proof of Lemma 1.

B.2 Proof of Lemma 2

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

$$
\mathsf{OPT}:=\min_{h\in\mathcal{H}}\max_{1\leqslant i\leqslant k}L(h,\mathrm{e}_i^{\mathsf{basis}})
$$

to be the optimal objective value. Additionally, set

$$
v^t:=L(h^t,w^t)-\mathsf{OPT}.
\tag{58}
$$

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

$$
v^t\leqslant\min_{h\in\mathcal{H}}L(h,w^t)-\mathsf{OPT}+\varepsilon_1=\min_{h\in\mathcal{H}}L(h,w^t)-\min_{h\in\mathcal{H}}\max_{i}L(h,\mathrm{e}_i^{\mathsf{basis}})+\varepsilon_1\leqslant\varepsilon_1,\qquad\forall 1\leqslant t\leqslant T.
\tag{59}
$$

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 $1-(k+1)\delta'$,

$$
\left|\sum_{t=1}^{T}\left\langle w^t,\widehat{r}^t\right\rangle-\sum_{t=1}^{T}L(h^t,w^t)\right|\leqslant 2\sqrt{T\log(1/\delta')},
\tag{60a}
$$

$$\left|\sum_{t=1}^{T}\widehat{r}_i^t - \sum_{t=1}^{T}L(h^t, \mathrm{e}_i^{\mathsf{basis}})\right| \leqslant 2\sqrt{T\log(1/\delta')}. \tag{60b}$$

These motivate us to look at $\sum_{t=1}^{T}\left\langle w^t, \widehat{r}^t\right\rangle$ (resp. $\sum_{t=1}^{T}\widehat{r}_i^t$) as a surrogate for $\sum_{t=1}^{T}L(h^t, w^t)$ (resp. $\sum_{t=1}^{T}L(h^t, \mathrm{e}_i^{\mathsf{basis}})$).

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

$$\begin{aligned}
\log\left(\frac{\sum_{i=1}^{k}W_i^{t+1}}{\sum_{i=1}^{k}W_i^t}\right) &\overset{\text{(i)}}{=} \log\left(\sum_{i=1}^{k}w_i^t\exp(\eta\widehat{r}_i^t)\right) \overset{\text{(ii)}}{\leqslant} \log\left(\sum_{i=1}^{k}w_i^t\big(1+\eta\widehat{r}_i^t+\eta^2(\widehat{r}_i^t)^2\big)\right)\\
&\leqslant \log\left(1+\eta\sum_{i=1}^{k}w_i^t\widehat{r}_i^t+\eta^2\sum_{i=1}^{k}w_i^t(\widehat{r}_i^t)^2)\right) \leqslant \eta\sum_{i=1}^{k}w_i^t\widehat{r}_i^t+\eta^2.
\end{aligned}\tag{61}$$

Here, (i) is valid since $w_i^t = \frac{W_i^t}{\sum_j W_j^t}$ and $W_i^{t+1} = W_i^t\exp(\eta\widehat{r}_i^t)$ (cf. lines 5 and 15 of Algorithm 1); (ii) arises from the elementary inequality $e^x \leqslant 1+x+x^2$ for $x\in[0,1]$ as well as the facts that $\eta\leqslant 1$ and $|\widehat{r}_i^t|\leqslant 1$. Summing the inequality (61) over all t and rearranging terms, we are left with

$$\begin{aligned}
\eta\sum_{t=1}^{T}\left\langle w^t, \widehat{r}^t\right\rangle &\geqslant \sum_{t=1}^{T}\left\{\log\left(\frac{\sum_{i=1}^{k}W_i^{t+1}}{\sum_{i=1}^{k}W_i^t}\right)-\eta^2\right\}\\
&= \log\left(\sum_{i=1}^{k}W_i^{T+1}\right)-\log\left(\sum_{i=1}^{k}W_i^1\right)-T\eta^2\\
&\geqslant \max_{1\leqslant i\leqslant k}\log(W_i^{T+1})-\log(k)-T\eta^2\\
&\geqslant \eta\max_{1\leqslant i\leqslant k}\sum_{t=1}^{T}\widehat{r}_i^t-\log(k)-T\eta^2,
\end{aligned}\tag{62}$$

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

$$\sum_{t=1}^{T}\left\langle w^t, \widehat{r}^t\right\rangle \geqslant \max_i\sum_{t=1}^{T}\widehat{r}_i^t - \left(\frac{\log(k)}{\eta}+\eta T\right). \tag{63}$$

Combine the above inequality with (60) to show that, with probability at least $1-(k+1)\delta'$,

$$\sum_{t=1}^{T}L(h^t, w^t) \geqslant \max_{1\leqslant i\leqslant k}\sum_{t=1}^{T}L(h^t, \mathrm{e}_i^{\mathsf{basis}}) - \left(\frac{\log(k)}{\eta}+\eta T+4\sqrt{T\log(1/\delta')}\right). \tag{64}$$

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

$$\begin{aligned}
\max_{1\leqslant i\leqslant k}\sum_{t=1}^{T}L(h^t, \mathrm{e}_i^{\mathsf{basis}}) &\leqslant T\mathsf{OPT}+\sum_{t=1}^{T}v^t+\left(\frac{\log(k)}{\eta}+\eta T+4\sqrt{T\log(1/\delta')}\right)\\
&\leqslant T\mathsf{OPT}+T\varepsilon_1+\left(\frac{\log(k)}{\eta}+\eta T+4\sqrt{T\log(1/\delta')}\right)\\
&\leqslant T\mathsf{OPT}+T\varepsilon,
\end{aligned}\tag{65}$$

where the penultimate line results from (59). Given that h^{final} is taken to be uniformly distributed over $\{h^t\}_{1\leqslant t\leqslant T}$, we arrive at

$$\max_{1\leqslant i\leqslant k}L(h^{\mathsf{final}}, \mathrm{e}_i^{\mathsf{basis}}) = \max_{1\leqslant i\leqslant k}\frac{1}{T}\sum_{t=1}^{T}L(h^t, \mathrm{e}_i^{\mathsf{basis}}) \leqslant \mathsf{OPT}+\varepsilon \tag{66}$$

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

Remark 4. *Note that the proof of this lemma works as long as $\widehat{r}_i^t\in[0,1]$ is an unbiased estimate of $L(h^t, \mathrm{e}_i^{\mathsf{basis}})$ for each $i\in[k]$, regardless of how many samples are used to construct $\widehat{r}_i^t$.*

B.3 Proof of Lemma 3

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

$$\bar{j} = \left\lfloor \log_2\left(\frac{k\log^2(2)}{50(\log_2(1/\eta)+1)^2\log_2^2(k)}\right)\right\rfloor - 2. \tag{67}$$

Let us define

$$\mathcal{W}_j := \big\{i \in [k] \mid \max_{1\leqslant t\leqslant T} w_i^t \in (2^{-j}, 2^{-(j-1)}]\big\}, \qquad 1 \leqslant j \leqslant \bar{j} \tag{68a}$$

$$\overline{\mathcal{W}} := [k]/\cup_j \mathcal{W}_j. \tag{68b}$$

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 $1 - 8T^4k\delta'$,*

$$|\mathcal{W}_j| \leqslant 8\cdot 10^7 \cdot \Big(\big(\log_2(1/\eta)+1\big)^2\log_2^2(k)\left(\log(k)+\log(1/\delta')\right)^3\left(\log_2(T)+1\right)\Big)\cdot 2^j \tag{69}$$

holds all $1 \leqslant j \leqslant \bar{j}$, with $\bar{j}$ defined in (67).

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

$$|\mathcal{W}_j| \leqslant \widetilde{O}(2^j), \qquad 1 \leqslant j \leqslant \bar{j}.$$

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

$$\begin{aligned}
\|\overline{w}^T\|_1 = \sum_{i=1}^k \max_{1\leqslant t\leqslant T} w_i^t &\leqslant k\cdot 2^{-(\bar{j}-1)} + \sum_{j=1}^{\bar{j}} |\mathcal{W}_j| 2^{-(j-1)} \\
&\leqslant k\cdot\frac{800\big(\log_2(1/\eta)+1\big)^2\log_2^2(k)}{k\log^2(2)} + \sum_{j=1}^{\bar{j}}|\mathcal{W}_j|2^{-(j-1)} \\
&\leqslant 2\cdot 10^8\cdot\Big(\big(\log_2(1/\eta)+1\big)^2\log_2^2(k)\big(\log(Tk)+\log(1/\delta)\big)^3\big(\log_2(T)+1\big)\Big),
\end{aligned}$$

where the first inequality is valid since $\max_{1\leqslant t\leqslant T} w_i^t \leqslant 2^{-(j-1)}$ holds for any $i \in \mathcal{W}_j$. This immediately concludes the proof of Lemma 3, as long as Lemma 13 can be established.

Noteworthily, proving Lemma 13 is the most challenging part of our analysis, and we dedicate the next section (Appendix C) to the proof of Lemma 13.

C 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" introduced in Section 4.2. For convenience, we restate the definition below.

Definition 4 (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) $\sum_{i\in\mathcal{I}} w_i^{t_2} \geqslant p\exp(x)$,

(iii) $\sum_{i\in\mathcal{I}} w_i^{t} \geqslant q$ *for any* $t_1 \leqslant t \leqslant t_2$.

We shall refer to t_1 as the starting point and t_2 as the end 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 either $s_1 < e_1 \leqslant s_2 < e_2$ or $s_2 < e_2 \leqslant s_1 < e_1$.

In addition, throughout this section, we shall take $\delta' = \delta/(32T^4k^2)$, and focus on any j obeying

$$1 \leqslant j \leqslant \log_2(k) - 2. \tag{70}$$

C.1 Main steps of the proof

In this subsection, we present the main steps of the proof. Before continuing, we find it helpful to underscore a high-level idea: 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. The key steps to construct such disjoint segments of interest are as follows:

1. Construct a suitable segment for each $i \in \mathcal{W}_j$ (see Lemma 14);
2. Identify sufficiently many disjoint blocks such that the segments within each block have nonempty intersection (see Lemma 15 and Figure 3);
3. From the above disjoint blocks, identify sufficiently many disjoint subsets such that the distributions associated with each subset can be linked with a common (sub)-segment, and that each of these (sub)-segments experiences sufficient changes between its starting and end points (see Lemma 16, Figure 4 and Figure 6).

In the sequel, we shall present the details of our proof, which consist of multiple steps.

C.1.1 Step 1: showing existence of a segment for each distribution in $\mathcal{W}_j$

Recall that $\mathcal{W}_j$ contains those distributions whose corresponding weight iterates obey $\max_{1\leqslant t\leqslant T} w_i^t \in (2^{-j}, 2^{-j+1}]$ (cf. (68a)). As it turns out, for any $i \in \mathcal{W}_j$, one can find an $\big(\frac{1}{2^{j+1}}, \frac{1}{2^{j+2}}, \log(2)\big)$-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 $1 \leqslant s_i < e_i \leqslant T$, such that*

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

In other words, there exists a $\big(\frac{1}{2^{j+1}}, \frac{1}{2^{j+2}}, \log(2)\big)$-segment (s_i, e_i) with the index set as $\{i\}$ (see Definition 4).

Proof. From the definition (68a) 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

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

which is properly defined since $w_i^1 = 1/k \leqslant 2^{-(j+2)}$ (see (70)). With this choice in mind, we have

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

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

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

where the last inequality results from our choice of η. This in turn allows us to show that

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

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

C.1.2 Step 2: constructing disjoint segments with good coverage

While Lemma 14 justifies the existence of suitable segments $\{(s_i, e_i)\}$ associated with each distribution in $\mathcal{W}_j$, we need to divide (a nontrivial subset of) them into certain disjoint blocks, where the segments in each block have at least one common inner points. This is accomplished in the following lemma.

Lemma 15. *Recall the definition of $\mathcal{W}_j$ in (68a). 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) $\mathcal{W}_j^p \subseteq \mathcal{W}_j$, $\mathcal{W}_j^p \cap \mathcal{W}_j^{p'} = \varnothing$, $\forall p \neq p'$;

(ii) $\sum_{p=1}^P |\mathcal{W}_j^p| \geqslant \frac{|\mathcal{W}_j|}{3(\log_2(T)+1)}$;

(iii) *Let* $\widetilde{s}_p = \min_{i\in\mathcal{W}_j^p} s_i$ *and* $\widetilde{e}_p = \max_{i\in\mathcal{W}_j^p} e_i$ *for each* $1 \leqslant p \leqslant P$. *One has* $1 \leqslant \widetilde{s}_1 < \widetilde{e}_1 \leqslant \widetilde{s}_2 < \widetilde{e}_2 \leqslant \cdots \leqslant \widetilde{s}_P < \widetilde{e}_P \leqslant T$ *and* $\max_{i\in\mathcal{W}_j^p} s_i \leqslant \min_{i\in\mathcal{W}_j^p} e_i$ *for each* $1 \leqslant p \leqslant P$.

Proof. See Appendix C.2. ☐

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 the indices contained in $\mathcal{W}_j$, and (b) the segments in each subset $\mathcal{W}_j^p$ share at least one common inner point.

Thus far, each of the segments constructed above is associated with a single distribution in $\mathcal{W}_j$. Clearly, it is likely that many of these segments might have non-trivial overlap; in other words, many of them might have shared sub-segments. What we intend to do next is to further group the indices in $\{\mathcal{W}_j^p\}$ into disjoint subgroups, and identify a common (sub)-segment for each of these subgroups. What remains unclear, however, is whether each of these (sub)-segments experiences sufficient weight changes between its starting and end points. We address these in the following lemma.

Lemma 16. *Recall the definition of $\mathcal{W}_j$ in (68a). 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 $\{\mathcal{V}_j^n\}_{n=1}^N$ of $\mathcal{W}_j$ satisfying the following properties:*

(i) $\mathcal{V}_j^n \subseteq \mathcal{W}_j$, $\mathcal{V}_j^n \cap \mathcal{V}_j^{n'} = \varnothing$, $\forall n \neq n'$;

(ii) $\sum_{n=1}^N |\mathcal{V}_j^n| \geqslant \frac{|\mathcal{W}_j|}{24\log_2(k)(\log_2(T)+1)}$;

(iii) *There exist* $1 \leqslant \widehat{s}_1 < \widehat{e}_1 \leqslant \widehat{s}_2 < \widehat{e}_2 \leqslant \cdots \leqslant \widehat{s}_N < \widehat{e}_N \leqslant T$, *and* $\{g_n\}_{n=1}^N \in [1,\infty)^N$, *such that for each* $1 \leqslant n \leqslant N$, $(\widehat{s}_n, \widehat{e}_n)$ *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* $1 \leqslant n \leqslant N$:

   (a) $\frac{g_n|\mathcal{V}_j^n|}{2^{j+2}} \leqslant \sum_{i\in\mathcal{V}_j^n} w_i^{\widehat{s}_n} \leqslant \frac{g_n|\mathcal{V}_j^n|}{2^{j+1}}$;

   (b) $\frac{g_n|\mathcal{V}_j^n|}{2^j} \cdot \exp\left(\frac{\log(2)}{2\log_2(k)}\right) \leqslant \sum_{i\in\mathcal{V}_j^n} w_i^{\widehat{e}_n}$;

   (c) $\sum_{i\in\mathcal{V}_j^n} w_i^t \geqslant \frac{|\mathcal{V}_j^n|}{2^{j+2}}$ *for any* t *obeying* $\widehat{s}_n \leqslant t \leqslant \widehat{e}_n$.

Proof. See Appendix C.3. ☐

In brief, the subsets $\{\mathcal{V}_j^n\}$ identified in Lemma 16 enjoy the following useful properties: (a) they are disjoint; (b) these subsets taken collectively also cover a reasonably large fraction of the elements in $\mathcal{W}_j$; (c) each subset $\mathcal{V}_j^n$ is linked with a suitable segment, whose associated weights have increased sufficiently from its starting point to its end point.

C.1.3 Step 3: bounding the length of segments

In this step, we turn attention to the length of segments, unveiling the interplay between the segment length and certain sub-optimality gaps.

Recall the definition (58) of v^t as follows

$$v^t := L(h^t, w^t) - \mathsf{OPT} \qquad \text{with } \mathsf{OPT} := \min_{h \in \mathcal{H}} \max_{1 \leqslant i \leqslant k} L(h, \mathrm{e}_i^{\mathsf{basis}}), \tag{73}$$

with $\mathrm{e}_i^{\mathsf{basis}}$ the i-th standard basis vector. The following lemma assists in bounding the length of the segments defined in Definition 4.

Lemma 17. *Let $j_{\max} = \lfloor \log_2(1/\eta) \rfloor + 1$. Assume the conditions in Lemma 3 hold. Suppose (t_1, t_2) is a (p, q, x)-segment satisfying $p \geqslant 2q > 0$. Then one has*

$$t_2 - t_1 \geqslant \frac{x}{2\eta}. \tag{74}$$

Moreover, if

$$\frac{qx^2}{50\big(\log_2(1/\eta) + 1\big)^2} \geqslant \frac{1}{k} \tag{75}$$

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

(a) the length of the segment satisfies

$$t_2 - t_1 \geqslant \frac{qx^2}{200\big(\log_2(1/\eta) + 1\big)^2 \eta^2}. \tag{76}$$

(b) the quantities $\{v^t\}$ obey

$$4 \sum_{\tau = t_1}^{t_2 - 1} (-v^\tau + \varepsilon_1) \geqslant \frac{qx^2}{100\big(\log_2(1/\eta) + 1\big)^2 \eta}. \tag{77}$$

Proof. See Appendix C.4. $\square$

In words, Lemma 17 shows that (i) in general the length of a segment scales at least linearly in $1/\eta$; (ii) if the parameter q of a segment (i.e., some lower bound on the weights of interest within this segment) is sufficiently large, then either the length of this segment scales at least *quadratically* in $1/\eta$, or the sum of certain sub-optimality gaps needs to be large enough.

C.1.4 Step 4: putting all this together

With the above lemmas in place, we are positioned to establish Lemma 13. In what follows, we denote by $\{\mathcal{V}_j^n\}_{n=1}^N$ and $\{(\widehat{s}_n, \widehat{e}_n)\}_{n=1}^N$ the construction in Lemma 16.

To begin with, it is observed that: for any $1 \leqslant j \leqslant \overline{j}$ (with $\overline{j}$ defined in (67)), one has

$$2^{-(j+2)} |\mathcal{V}_j^n| \cdot \frac{\log^2(2)}{50(\log_2(1/\eta) + 1)^2 \log_2^2(k)} \geqslant 2^{-(\overline{j}+2)} \cdot \frac{\log^2(2)}{50(\log_2(1/\eta) + 1)^2 \log_2^2(k)} \geqslant \frac{1}{k}. \tag{78}$$

Recall that $v^\tau \leqslant \varepsilon_1$ (see (59)). Combining this fact with Lemma 17 (by setting $q = 2^{-(j+2)} |\mathcal{V}_j^n|$ and $x = \frac{\log(2)}{\log_2(k)}$) reveals that, for each $1 \leqslant n \leqslant N$,

$$T\eta + \left\{ 4T\varepsilon_1 + 4 \sum_{t=1}^{T} (-v^t) \right\} \geqslant \sum_{n=1}^{N} (\widehat{e}_n - \widehat{s}_n)\eta + 4 \sum_{n=1}^{N} \sum_{\tau = \widehat{s}_n}^{\widehat{e}_n - 1} (-v^\tau + \varepsilon_1) \tag{79}$$

$$\geqslant \frac{2^{-(j+2)} \sum_{n=1}^{N} |\mathcal{V}_j^n| \log^2(2)}{800 \log_2^2(k) \big(\log_2(1/\eta) + 1 \big)^2 \eta}, \tag{80}$$

where the first inequality results from the disjoint nature of the segments $\{[\widehat{s}_n, \widehat{e}_n]\}_{1 \leqslant n \leqslant N}$, and the second inequality comes from Lemma 17. Moreover, it follows from (64) that

$$\sum_{t=1}^{T} (-v^t) \leqslant \frac{\log(k)}{\eta} + \eta T + 4\sqrt{T \log(1/\delta')}, \tag{81}$$

which taken together with (80) and the choice $\varepsilon_1 = \eta$ gives

$$\begin{aligned} \sum_{n=1}^{N} |\mathcal{V}_j^n| \leqslant \frac{3200\eta \big(\log_2(1/\eta) + 1 \big)^2 \log_2^2(k) \cdot 2^{j+2}}{\log^2(2)} \cdot \left(\frac{\log(k)}{\eta} + \eta T + 4\sqrt{T \log(1/\delta')} \right) \\ + \frac{4000 T \big(\log_2(1/\eta) + 1 \big)^2 \log_2^2(k) \cdot 2^{j+2} \eta^2}{\log^2(2)}. \end{aligned} \tag{82}$$

To finish up, it follows from Property (ii) of Lemma 16 that

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

for any $1 \leqslant j \leqslant \overline{j}$. The proof is completed by recalling that $\delta' = \frac{\delta}{4(T+k+1)}$.

C.2 Proof of Lemma 15

For any integer $1 \leqslant x \leqslant \log_2(T) + 1$, define

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

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

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

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

$$|\mathcal{W}_j(x^\star)| \geqslant \frac{|\mathcal{W}_j|}{\log_2(T) + 1}. \tag{83}$$

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

To proceed, let us set

$$\kappa_1 := \min_{i \in \mathcal{W}_j(x^\star)} e_i, \qquad \mathcal{U}_j^1 := \big\{ i \in \mathcal{W}_j(x^\star) \mid s_i \leqslant \kappa_1 \big\}, \tag{84a}$$

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

$$\kappa_{o+1} := \min_{i \in \mathcal{W}_j(x^\star) / \cup_{o'=1}^{o} \mathcal{U}_j^{o'}} e_i, \tag{84b}$$

$$\mathcal{U}_j^{o+1} := \big\{ i \in \mathcal{W}_j(x^\star) / \cup_{o'=1}^{o} \mathcal{U}_j^{o'} \mid s_i \leqslant \kappa_{o+1} \big\}. \tag{84c}$$

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

$$s_{i_2} \leqslant \kappa_o \leqslant e_{i_1}, \quad \forall i_1, i_2 \in \mathcal{U}_j^o \qquad \Longleftrightarrow \qquad \max_{i \in \mathcal{U}_j^o} s_i \leqslant \min_{i \in \mathcal{U}_j^o} e_i. \tag{85}$$

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

$$\sum_{o\geqslant 1}|\mathcal{U}_j^{3o-2}|\geqslant \frac{1}{3}\sum_{o\geqslant 1}|\mathcal{U}_j^{o}| = \frac{1}{3}|\mathcal{W}_j(x^\star)|. \tag{86}$$

With the above construction in place, we would like to verify that $\{\mathcal{U}_j^{3o-2}\}_{o\geqslant 1}$ forms the desired group of subsets. First of all, Condition (i) holds directly from the definition of $\{\mathcal{U}_j^{o}\}_{o\geqslant 1}$. When it comes to Condition (ii), it follows from (86) and (83) that

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

Regarding Condition (iii), it suffices to verify that

$$\max_{i\in\mathcal{U}_j^{3o-2}} e_i \leqslant \min_{i\in\mathcal{U}_j^{3o+1}} s_i \tag{87}$$

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

$$\kappa_{o+1}=e_i\geqslant s_i+2^{x^\star-1}\geqslant\kappa_o+2^{x^\star-1}. \tag{88}$$

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

$$s_i\geqslant\kappa_{3o}\geqslant\kappa_{3o-1}+2^{x^\star-1}\geqslant\kappa_{3o-2}+2^{x^\star}.$$

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

$$e_l\leqslant s_l+2^{x^\star}\leqslant\kappa_{3o-2}+2^{x^\star}.$$

Putting all this together yields

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

The proof is thus complete.

C.3 Proof of Lemma 16

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 $\{(\widetilde{s}_p,\widetilde{e}_p)\}_{p=1}^P$ be the construction in Lemma 15.

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

$$t_{\mathrm{mid}}^p \coloneqq \min_{i\in\mathcal{W}_j^p} e_i,$$

which, in view of Lemma 15, is a common inner point of the segments in $\mathcal{W}_j^p$. We can derive, for each $i\in\mathcal{W}_j^p$,

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

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

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

Without loss of generality, we assume that

$$\sum_{i\in\mathcal{W}_j^p}\mathbb{1}\left\{\log\left(\frac{w_i^{e_i}}{w_i^{t_{\mathrm{mid}}^p}}\right)\geqslant\frac{x}{2}\right\}\geqslant\frac{|\mathcal{W}_j^p|}{2}. \tag{89}$$

This means that the set define below

$$\widehat{\mathcal{W}}_j^p := \left\{i\in\mathcal{W}_j^p \mid \log\left(w_i^{e_i}/w_i^{t_{\mathrm{mid}}^p}\right)\geqslant\frac{x}{2}\right\} \tag{90}$$

satisfies

$$|\widehat{\mathcal{W}}_j^p|\geqslant\frac{|\mathcal{W}_j^p|}{2}. \tag{91}$$

In what follows, we take[8]

$$Q(p):=|\widehat{\mathcal{W}}_j^p|,\qquad \widetilde{l}(p):=\max\left\{l\geqslant 0\mid 2^l\leqslant Q(p)\right\}\qquad\text{and}\qquad \widetilde{Q}(p):=2^{\widetilde{l}(p)}.$$

Without loss of generality, we assume

$$\widehat{\mathcal{W}}_j^p=\left\{1,2,\ldots,Q(p)\right\}\qquad\text{and}\qquad e_1\leqslant e_2\leqslant\cdots\leqslant e_{Q(p)}. \tag{92}$$

In the sequel, we shall often abbreviate $Q(p)$, $\widetilde{l}(p)$ and $\widetilde{Q}(p)$ as Q, $\widetilde{l}$ and $\widetilde{Q}$, respectively, as long as it is clear from the context.

Step b): constructing $\mathcal{K}_l$ and $\widetilde{\mathcal{W}}_j^p(l)$. Let us take $e_0=t_{\mathrm{mid}}^p$, 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:

$$\mathcal{K}_1=\left\{[e_0,e_{2^{\widetilde{l}-1}}]\right\}; \tag{93a}$$
$$\mathcal{K}_2=\left\{[e_0,e_{2^{\widetilde{l}-2}}],[e_0,e_{2^{\widetilde{l}-2}}]\oplus 2^{\widetilde{l}-1}\right\}; \tag{93b}$$
$$\mathcal{K}_3=\left\{[e_0,e_{2^{\widetilde{l}-3}}],[e_0,e_{2^{\widetilde{l}-3}}]\oplus 2^{\widetilde{l}-2},[e_0,e_{2^{\widetilde{l}-3}}]\oplus 2\cdot 2^{\widetilde{l}-2},[e_0,e_{2^{\widetilde{l}-3}}]\oplus 3\cdot 2^{\widetilde{l}-2}\right\}; \tag{93c}$$
$$\cdots$$
$$\mathcal{K}_l=\left\{[e_0,e_{2^{\widetilde{l}-l}}],[e_0,e_{2^{\widetilde{l}-l}}]\oplus 2^{\widetilde{l}-l+1},[e_0,e_{2^{\widetilde{l}-l}}]\oplus 2\cdot 2^{\widetilde{l}-l+1},\ldots,[e_0,e_{2^{\widetilde{l}-l}}]\oplus(2^{l-1}-1)2^{\widetilde{l}-l+1}\right\}; \tag{93d}$$
$$\cdots$$
$$\mathcal{K}_{\widetilde{l}}=\left\{[e_{2i},e_{2i+1}]\mid i=0,1,2,\ldots,2^{\widetilde{l}-1}-1\right\}; \tag{93e}$$
$$\mathcal{K}_{\widetilde{l}+1}=\left\{[e_{2i+1},e_{2i+2}]\mid i=0,1,2,\ldots,2^{\widetilde{l}-1}-1\right\}. \tag{93f}$$

For each $i\in[\widetilde{Q}-1]$ with binary form $\{i_l\}_{l=1}^{\widetilde{l}}$ and $0\leqslant l\leqslant\widetilde{l}$, we define $\mathsf{trunc}(i,l)$ to be the number with binary form $\{i_1,i_2,\ldots,i_l,0,0,\ldots,0\}$. For example, $\mathsf{trunc}(i,0)=0$ and $\mathsf{trunc}(i,\widetilde{l})=i$.

From the definition (90) of $\widehat{\mathcal{W}}_j^p$, we know that for each $i\in[\widetilde{Q}-1]$,

$$\frac{x}{2}\leqslant\log\left(\frac{w_i^{e_i}}{w_i^{e_0}}\right)=\sum_{l=1}^{\widetilde{l}}\log\left(\frac{w_i^{e_{\mathsf{trunc}(i,l)}}}{w_i^{e_{\mathsf{trunc}(i,l-1)}}}\right)=\sum_{l=1}^{\widetilde{l}}\log\left(\frac{w_i^{e_{\mathsf{trunc}(i,l)}}}{w_i^{e_{\mathsf{trunc}(i,l-1)}}}\right)\mathbb{1}\left\{e_{\mathsf{trunc}(i,l)}\neq e_{\mathsf{trunc}(i,l-1)}\right\}, \tag{94}$$

[8] We assume $\widetilde{Q}\geqslant 2$ without loss of generality. In the case $\widetilde{Q}=1$, 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

$$\max_{1 \leqslant l \leqslant \widetilde{l}} \log\left(\frac{w_i^{e_{\mathsf{trunc}(i,l)}}}{w_i^{e_{\mathsf{trunc}(i,l-1)}}}\right) \geqslant \frac{x}{2\widetilde{l}}. \tag{95}$$

By defining

$$\widetilde{\mathcal{W}}_j^p(l) := \left\{ i \in \widehat{\mathcal{W}}_j^p : \arg\max_{1 \leqslant l' \leqslant \widetilde{l}} \log\left(\frac{w_i^{e_{\mathsf{trunc}(i,l')}}}{w_i^{e_{\mathsf{trunc}(i,l'-1)}}}\right) = l \right\}$$

for each[9] $1 \leqslant l \leqslant \widetilde{l}$, we can demonstrate that

$$\sum_{l=1}^{\widetilde{l}} \left|\widetilde{\mathcal{W}}_j^p(l)\right| \geqslant \widetilde{Q} - 1, \tag{96}$$

thus implying the existence of some $1 \leqslant l^\star \leqslant \widetilde{l}$ obeying

$$\left|\widetilde{W}_j^p(l^\star)\right| \geqslant \frac{\widetilde{Q}-1}{\widetilde{l}} \geqslant \frac{\widetilde{Q}}{2\widetilde{l}}. \tag{97}$$

Step c): constructing $\widetilde{\mathcal{W}}_j^p(l,o)$, $\widehat{s}(p,o)$ and $\widehat{e}(p,o)$. By definition, for any i, if $\mathsf{trunc}(i,l^\star) \neq \mathsf{trunc}(i,l^\star-1)$, then one has

$$[e_{\mathsf{trunc}(i,l^\star-1)}, e_{\mathsf{trunc}(i,l^\star)}] \in \mathcal{K}_{l^\star},$$

where the set $\mathcal{K}_l$ has been defined in (93). In addition, from the construction of $\widetilde{\mathcal{W}}_j^p(l^\star)$ (see (97)), we know that $\mathsf{trunc}(i,l^\star) \neq \mathsf{trunc}(i,l^\star-1)$ for any $i \in \widetilde{\mathcal{W}}_j^p(l^\star)$. For each $1 \leqslant o \leqslant 2^{l^\star-1}$, define

$$\widetilde{\mathcal{W}}_j^p(l^\star,o) := \left\{ i \in \widetilde{\mathcal{W}}_j^p(l^\star) \mid [e_{\mathsf{trunc}(i,l^\star-1)}, e_{\mathsf{trunc}(i,l^\star)}] = [e_0, e_{2^{\widetilde{l}-l^\star}}] \oplus (o-1)2^{\widetilde{l}-l^\star+1} \right\}, \tag{98}$$

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

In addition, for any $1 \leqslant p \leqslant P$ and $1 \leqslant o \leqslant 2^{l^\star(p)-1}$, we set

$$\widehat{s}(p,o) = e_{(o-1)2^{\widetilde{l}(p)-l^\star(p)+1}}, \tag{99a}$$

$$\widehat{e}(p,o) = e_{2^{\widetilde{l}(p)-l^\star(p)}+(o-1)2^{\widetilde{l}(p)-l^\star(p)+1}}. \tag{99b}$$

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

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

$$\left\{ \left\{ \widetilde{\mathcal{W}}_j^p\big(l^\star(p),o\big) \right\}_{o=1}^{2^{l^\star(p)-1}} \right\}_{p=1}^{P} \qquad \text{with intervals} \qquad \left\{ \left\{ \widehat{s}(p,o), \widehat{e}(p,o) \right\}_{o=1}^{2^{l^\star(p)-1}} \right\}_{p=1}^{P}$$

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

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

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

$$\widetilde{\mathcal{W}}_j^p\big(l^\star(p),o\big) \subseteq \widetilde{\mathcal{W}}_j^p\big(l^\star(p)\big) \subseteq \widehat{\mathcal{W}}_j^p \subseteq \mathcal{W}_j^p \subseteq \mathcal{W}_j.$$

In addition, if

$$\widetilde{W}_j^{p_1}\big(l^\star(p_1),o_1\big) \cap \widetilde{W}_j^{p_2}\big(l^\star(p_2),o_2\big) \neq \varnothing,$$

then one has $\mathcal{W}_j^{p_2} \cap \mathcal{W}_j^{p_2} \neq \varnothing$, and as a result, $p_1 = p_2$ (otherwise it violates the condition that $\mathcal{W}_j^{p_2} \cap \mathcal{W}_j^{p_2} = \varnothing$ for $p_1 \neq p_2$). It also follows from the definition in (98) that $o_1 = o_2$. Therefore, for any (p_1,o_1) that does not equal (p_2,o_2), we have $\widetilde{W}_j^{p_1}\big(l^\star(p_1),o_1\big) \cap \widetilde{W}_j^{p_2}\big(l^\star(p_2),o_2\big) = \varnothing$.

[9]Without loss of generality, we assume the arg max function is a single-valued function.

Property (ii). By construction, we have

$$\sum_{o=1}^{2^{l^\star(p)-1}} \left|\widetilde{\mathcal{W}}_j^p\big(l^\star(p),o\big)\right| = \left|\widetilde{\mathcal{W}}_j^p\big(l^\star(p)\big)\right| \geqslant \frac{|\widehat{\mathcal{W}}_j^p|}{4\log_2\left(|\widehat{\mathcal{W}}_j^p|\right)} \geqslant \frac{|\mathcal{W}_j^p|}{8\log_2\left(|\widehat{\mathcal{W}}_j^p|\right)}, \tag{100}$$

where we have made use of (97) and (91). Summing over p and applying Lemma 15 yield

$$\sum_{p=1}^{P}\sum_{o=1}^{2^{l^\star(p)-1}} \left|\widetilde{\mathcal{W}}_j^p\big(l^\star(p),o\big)\right| \geqslant \sum_{p=1}^{P}\frac{|\mathcal{W}_j^p|}{8\log_2(k)} \geqslant \frac{|\mathcal{W}_j|}{24\log_2(k)\big(\log_2(T)+1\big)}. \tag{101}$$

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

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

where the last inequality holds since, by construction, $w_i^{\widehat{s}(p,o)} \geqslant 2^{-(j+2)}$ (see Lemma 14). Then Property (iii)(a) is satisfied since

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

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

$$s_i \leqslant e_{\mathsf{trunc}(i,l-1)} \leqslant e_i \qquad \text{for any } 1\leqslant l\leqslant \widetilde{l}(p),$$

which is valid since $\max_{i\in\mathcal{W}_j^p} s_i \leqslant \min_{i\in\mathcal{W}_j^p} e_i$ (see Lemma 15) and (92). It then holds that

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

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

Moreover, by construction, we know that for any $i\in\widetilde{\mathcal{W}}_j^p\big(l^\star(p),o\big)$,

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

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

$$\sum_{i\in\widetilde{\mathcal{W}}_j^p(l^\star(p),o)} w_i^{\widehat{s}(p,o)} \geqslant 2^{-(j+2)}\cdot\left|\widetilde{\mathcal{W}}_j^p\big(l^\star(p),o\big)\right|$$

$$\log\left(\frac{\sum_{i\in\widetilde{\mathcal{W}}_j^p(l^\star(p),o)} w_i^{\widehat{e}(p,o)}}{\sum_{i\in\widetilde{\mathcal{W}}_j^p(l^\star(p),o)} w_i^{\widehat{s}(p,o)}}\right) \geqslant \log\left(\frac{\sum_{i\in\widetilde{\mathcal{W}}_j^p(l^\star(p),o)} w_i^{\widehat{s}(p,o)}\cdot\exp\left(\frac{x}{2\widetilde{l}(p)}\right)}{\sum_{i\in\widetilde{\mathcal{W}}_j^p(l^\star(p),o)} w_i^{\widehat{s}(p,o)}}\right) = \frac{x}{2\widetilde{l}(p)} \geqslant \frac{\log(2)}{2\log_2(k)}.$$

Property (iii)(c). Note that for any t obeying $\widehat{s}(p,o)\leqslant t\leqslant\widehat{e}(p,o)$, and any $i\in\widetilde{\mathcal{W}}_j^p\big(l^\star(p),o\big)$, it holds that $s_i\leqslant\widehat{s}(p,o)\leqslant t\leqslant\widehat{e}(p,o)\leqslant e_i$. Recall that $w_i^t\geqslant 2^{-(j+2)}$ for any $t\in[s_i,e_i]$ (see Lemma 14). As a result,

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

Proper ordering. To finish up, it remains to verify that the intersection of $[\widehat{s}(p_1, o_1), \widehat{e}(p_1, o_1)]$ and $[\widehat{s}(p_2, o_2), \widehat{e}(p_2, o_2)]$ 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

$$\widetilde{s}_{p_1} \leqslant \widehat{s}(p_1, o_1) < \widehat{e}(p_1, o_1) \leqslant \widetilde{e}_{p_1} \leqslant \widetilde{s}_{p_2} \leqslant \widehat{s}(p_2, o_2) < \widehat{e}(p_2, o_2),$$

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

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

which comes from the construction (99).

We have thus completed the proof of this lemma.

C.4 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 (74). Recall that $[t_1, t_2]$ is assumed to be a (p, q, x)-segment. From the definition of the segment (see Definition 4), there exists $i \in [k]$ such that

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

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

$$\eta \sum_{\tau=t_1}^{t_2-1} \widehat{r}_i^{\tau} - \log(Z^{t_2}/Z^{t_1}) \geqslant x. \tag{102}$$

Moreover, recognizing that

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

and $\widehat{r}_i^{\tau} \leqslant 1$ for any $1 \leqslant \tau \leqslant T$, we can invoke (102) to show that

$$x \leqslant 2(t_2 - t_1)\eta, \tag{103}$$

from which the claimed inequality (74) 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 $v_h \in \mathbb{R}^k$:

$$v_h = [v_{h,i}]_{i\in[k]} \qquad \text{with} \quad v_{h,i} = L(h, \mathrm{e}_i^{\mathsf{basis}}) - \mathsf{OPT}. \tag{104}$$

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

$$\langle v_{h^\tau}, w^t \rangle \geqslant \min_{h\in\mathcal{H}} \langle v_h, w^t \rangle \geqslant \langle v_{h^t}, w^t \rangle - \varepsilon_1 = v^t - \varepsilon_1, \tag{105}$$

where we recall the definition of v^t in (73). 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

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

where $\xi_i^\tau = \widehat{r}_i^\tau - L(h^\tau, \mathrm{e}_i^{\mathsf{basis}})$ 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

$$\left\langle \frac{1}{\eta}\log\left(\frac{W^{t_2}}{W^{t_1}}\right) - \Delta_{t_1,t_2},\, w^t \right\rangle - (t_2 - t_1)\mathsf{OPT} = \sum_{\tau=t_1}^{t_2-1} \langle v_{h^\tau}, w^t\rangle \geqslant (t_2 - t_1)(v^t - \varepsilon_1), \tag{106}$$

where the last inequality results from (105).

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 (106) that

$$\left\langle \log\left(\frac{w^{t_2}}{w^t}\right) - \eta\Delta_{t,t_2}, w^t \right\rangle + \log\left(\frac{Z^{t_2}}{Z^t}\right) - \eta(t_2 - t)\mathsf{OPT} \geqslant \eta(t_2 - t)(v^t - \varepsilon_1). \tag{107}$$

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

$$\begin{aligned} \mathsf{KL}(w^t \,\|\, w^{t_2}) &:= \left\langle w^t, \log\left(\frac{w^t}{w^{t_2}}\right)\right\rangle \\ &\leqslant \log(Z^{t_2}/Z^t) - \eta(t_2 - t)\mathsf{OPT} - \eta\langle w^t, \Delta_{t,t_2}\rangle + \eta(t_2 - t)(\varepsilon_1 - v^t). \end{aligned} \tag{108}$$

In what follows, we shall cope with the right-hand side of (108).

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

$$\begin{aligned} \log(Z^{t_2}/Z^t) &= \sum_{\tau=t}^{t_2-1} \log(Z^{\tau+1}/Z^\tau) = \sum_{\tau=t}^{t_2-1} \log\left(\sum_{i\in[k]} \frac{W_i^\tau \exp(\eta\widehat{r}_i^\tau)}{\sum_{j\in[k]} W_i^\tau}\right) \\ &\overset{\text{(i)}}{=} \sum_{\tau=t}^{t_2-1} \log\left(\sum_{i=1}^k w_i^\tau \exp(\eta\widehat{r}_i^\tau)\right) \overset{\text{(ii)}}{\leqslant} \sum_{\tau=t}^{t_2-1} \log\left(\sum_{i=1}^k w_i^\tau + \sum_{i=1}^k w_i^\tau(\eta\widehat{r}_i^\tau) + 2\sum_{i=1}^k w_i^\tau \eta^2 (\widehat{r}_i^\tau)^2\right) \\ &\overset{\text{(iii)}}{\leqslant} \sum_{\tau=t}^{t_2-1} \log\left(1 + \eta\sum_{i=1}^k w_i^\tau \widehat{r}_i^\tau + 2\eta^2\right) \leqslant \sum_{\tau=t}^{t_2-1} \left(\eta\sum_{i=1}^k w_i^\tau \widehat{r}_i^\tau + 2\eta^2\right) \\ &\overset{\text{(iv)}}{=} \eta\sum_{\tau=t}^{t_2-1} v^\tau + \eta(t_2 - t)\mathsf{OPT} + \eta\sum_{\tau=t}^{t_2-1}\sum_{i=1}^k \left\langle w_i^\tau, \widehat{r}_i^\tau - v_{h^\tau,i} - \mathsf{OPT}\right\rangle + 2(t_2 - t)\eta^2 \\ &\overset{\text{(v)}}{\leqslant} \eta(t_2 - t)\varepsilon_1 + \eta(t_2 - t)\mathsf{OPT} + \eta\sum_{\tau=t}^{t_2-1}\sum_{i=1}^k \left\langle w_i^\tau, \widehat{r}_i^\tau - v_{h^\tau,i} - \mathsf{OPT}\right\rangle + 2(t_2 - t)\eta^2. \end{aligned} \tag{109}$$

Here, (i) comes from line 5 of Algorithm 1, (ii) follows from the elementary inequality $\exp(x) \leqslant 1 + x + 2x^2$ for any $x \leqslant 1$, (iii) is valid since $\sum_i w_i^\tau = 1$ and $|\widehat{r}_i^\tau| \leqslant 1$, (iv) holds due to the fact that $v^t = \langle w^t, v_{h^t}\rangle$, and (v) arises from the fact that $v^\tau \leqslant \varepsilon_1$ (see (59)).

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

$$\eta \sum_{\tau=t}^{t_2-1} \sum_{i=1}^{k} \big\langle w_i^\tau, \widehat{r}_i^\tau - v_{h^\tau,i} - \mathsf{OPT} \big\rangle = \eta \sum_{\tau=t}^{t_2-1} \sum_{i=1}^{k} w_i^\tau \xi_i^\tau, \tag{110a}$$

$$\eta \big\langle w^t, \Delta_{t,t_2} \big\rangle = \eta \sum_{\tau=t}^{t_2-1} \sum_{i=1}^{k} w_i^t \xi_i^\tau. \tag{110b}$$

Let $\mathcal{F}^\tau$ 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 $O\big(\frac{1}{k\overline{w}_i^\tau}\big)$, according to the update rule (see line 14 in Algorithm 1);
- $\{\xi_i^\tau\}_{i\in[k]}$ are independent conditioned on $\mathcal{F}^\tau$.

Let us develop bounds on the two quantities in (110) below.

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

$$|q^\tau| \leqslant 1, \qquad \mathbb{E}[q^\tau | \mathcal{F}^\tau] = 0 \qquad \text{and} \qquad \mathsf{Var}[q^\tau|\mathcal{F}^\tau] \leqslant \sum_{i=1}^{k} \frac{(w_i^t)^2}{k\overline{w}_i^\tau} \leqslant \sum_{i=1}^{k} \frac{w_i^t}{k} = \frac{1}{k}. \tag{111}$$

  By virtue of Freedman's inequality (cf. Lemma 7), with probability at least $1-\delta'$ one has

$$\left| \sum_{\tau=t}^{t_2-1} \sum_{i=1}^{k} w_i^t \xi_i^\tau \right| \leqslant 2\sqrt{\frac{t_2-t}{k}\log(2/\delta')} + 2\log(2/\delta'); \tag{112}$$

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

$$|\widehat{q}^\tau| \leqslant 1, \qquad \mathbb{E}[\widehat{q}^\tau | \mathcal{F}^\tau] = 0 \qquad \text{and} \qquad \mathsf{Var}[\widehat{q}^\tau|\mathcal{F}^\tau] \leqslant \sum_{i=1}^{k} \frac{(w_i^\tau)^2}{k\overline{w}_i^\tau} \leqslant \sum_{i=1}^{k} \frac{w_i^\tau}{k} = \frac{1}{k}.$$

  Invoke Freedman's inequality (cf. Lemma 7) once again to show that, with probability exceeding $1-\delta'$,

$$\left| \sum_{\tau=t}^{t_2-1} \sum_{i=1}^{k} w_i^\tau \xi_i^\tau \right| \leqslant 2\sqrt{\frac{t_2-t}{k}\log(2/\delta')} + 2\log(2/\delta'). \tag{113}$$

Step 4: bounding the KL divergence between w^t and w^{t_2}. Combining (108), (109), (112) and (113), and applying the union bound over (t, t_2), we can demonstrate that with probability at least $1 - 6T^4 k\delta'$,

$$\begin{aligned}\mathsf{KL}(w^t \,\|\, w^{t_2}) &\leqslant 2(t_2-t)\eta\varepsilon_1 - (t_2-t)\eta v^t \\ &\quad + 4\eta\sqrt{\frac{(t_2-t)\log(2/\delta')}{k}} + 2(t_2-t)\eta^2 + 4\eta\log(2/\delta')\end{aligned} \tag{114}$$

holds for any $1 \leqslant t < t_2 \leqslant T$. The analysis below then operates under the condition that (114) holds for any $1 \leqslant t < t_2 \leqslant T$.

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

$$\tau_{\widehat{j}} := \min\{\tau \mid t_1 \leqslant \tau \leqslant t_2 - 1, \, -v^\tau \leqslant 2^{-(\widehat{j}-1)}\}, \qquad 1 \leqslant \widehat{j} \leqslant j_{\max}; \tag{115a}$$

$$\tau_{j_{\max+1}} := t_2. \tag{115b}$$

By definition, we know that $\tau_1 = t_1$ and $\tau_{j_2} \geqslant \tau_{j_1}$ for $j_2 \geqslant j_1$. Let $\mathcal{I}$ be the index set associated with 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 $1 \leqslant \widetilde{j} \leqslant j_{\max}$ such that

$$\log\left(\frac{y_{\widetilde{j}+1}}{y_{\widetilde{j}}}\right) \geqslant \frac{x}{\log_2(1/\eta)+1}. \tag{116}$$

Proof of (116). Suppose that none of $1 \leqslant \widetilde{j} \leqslant j_{\max}$ satisfies (116). Then for any $1 \leqslant \widehat{j} \leqslant j_{\max}$, it holds that $\log\left(\frac{y_{\widehat{j}+1}}{y_{\widehat{j}}}\right) < \frac{x}{\log_2(1/\eta)+1}$, which implies that $y_{\widehat{j}} > y_{\widehat{j}+1}\exp\left(-\frac{x}{\log_2(1/\eta)+1}\right)$. As a result, we have

$$\begin{aligned} y_1 &> y_{j_{\max}+1}\cdot\exp\left(-j_{\max}\cdot\frac{x}{\log_2(1/\eta)+1}\right) = \left(\sum_{i\in\mathcal{I}} w_i^{t_2}\right)\cdot\exp\left(-j_{\max}\cdot\frac{x}{\log_2(1/\eta)+1}\right) \\ &\geqslant p\exp(x)\cdot\exp(-x) = p, \end{aligned}$$

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

Now, assume that $\widetilde{j}$ satisfies (116). From the definition of the (p,q,x)-segment, we have $y_{\widetilde{j}} \geqslant q$. It follows from (114) that

$$\begin{aligned} \mathsf{KL}(w^{\tau_{\widetilde{j}}} \,\|\, w^{\tau_{\widetilde{j}+1}}) &\leqslant 2(\tau_{\widetilde{j}+1}-\tau_{\widetilde{j}})\eta\varepsilon_1 + (\tau_{\widetilde{j}+1}-\tau_{\widetilde{j}})\eta 2^{-(\widetilde{j}-1)} \\ &\quad + 4\eta\sqrt{\frac{(\tau_{\widetilde{j}+1}-\tau_{\widetilde{j}})\log(2/\delta')}{k}} + 2(\tau_{\widetilde{j}+1}-\tau_{\widetilde{j}})\eta^2 + 4\eta\log(2/\delta'). \end{aligned} \tag{117}$$

Since $\log\left(\frac{y_{\widetilde{j}+1}}{y_{\widetilde{j}}}\right) \geqslant \frac{x}{\log_2(1/\eta)+1}$ and $y_{\widetilde{j}} \geqslant q$, we can invoke Lemma 10 and Lemma 11 to show that

$$\mathsf{KL}(w^{\tau_{\widetilde{j}}} \,\|\, w^{\tau_{\widetilde{j}+1}}) \geqslant \mathsf{KL}\left(\mathsf{Ber}\big(y_{\widetilde{j}}\big) \,\|\, \mathsf{Ber}\big(y_{\widetilde{j}+1}\big)\right) \geqslant \frac{qx^2}{4\big(\log_2(1/\eta)+1\big)^2},$$

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

$$\begin{aligned} \frac{qx^2}{4\big(\log_2(1/\eta)+1\big)^2} &\leqslant 2(\tau_{\widetilde{j}+1}-\tau_{\widetilde{j}})\eta\varepsilon_1 + (\tau_{\widetilde{j}+1}-\tau_{\widetilde{j}})\eta 2^{-(\widetilde{j}-1)} \\ &\quad + 4\eta\sqrt{\frac{(\tau_{\widetilde{j}+1}-\tau_{\widetilde{j}})\log(2/\delta')}{k}} + 2(\tau_{\widetilde{j}+1}-\tau_{\widetilde{j}})\eta^2 + 4\eta\log(2/\delta'), \end{aligned} \tag{118}$$

which in turn results in

$$\begin{aligned} \tau_{\widetilde{j}+1}-\tau_{\widetilde{j}} &\geqslant \min\left\{\frac{qx^2}{100\big(\log_2(1/\eta)+1\big)^2}\min\left\{\frac{1}{\eta\varepsilon_1}, \frac{2^{\widetilde{j}-1}}{\eta}, \frac{1}{\eta^2}\right\}, \frac{kq^2x^4}{10000\eta^2\log(1/\delta')\big(\log_2(1/\eta)+1\big)^4}\right\} \\ &= \frac{qx^2}{100\big(\log_2(1/\eta)+1\big)^2}\min\left\{\frac{2^{\widetilde{j}-1}}{\eta}, \frac{1}{\eta^2}\right\} \qquad (119) \\ &= \frac{qx^2}{100\big(\log_2(1/\eta)+1\big)^2}\frac{2^{\widetilde{j}-1}}{\eta}. \qquad (120) \end{aligned}$$

Here, to see why (119) and (120) hold, it suffices to note that

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

a property that arises from the fact that $2^{\widetilde{j}-1} \leqslant 1/\eta$ and the assumption that $\frac{qx^2}{50(\log_2(1/\eta)+1)^2} \geqslant \frac{1}{k}$.

With (120) in mind, we are ready to finish the proof by dividing into two cases.

- *Case 1:* $\widetilde{j} = j_{\max}$. It follows from (120) that

$$t_2 - t_1 \geqslant \tau_{\widetilde{j}+1}-\tau_{\widetilde{j}} \geqslant \frac{qx^2}{100\big(\log_2(1/\eta)+1\big)^2}\frac{2^{j_{\max}-1}}{\eta} \geqslant \frac{qx^2}{200\big(\log_2(1/\eta)+1\big)^2\eta^2}.$$

- *Case 2:* $1 \leqslant \widetilde{j} \leqslant j_{\max} - 1$. It comes from the definition (115) that

$$\sum_{\tau=t_1}^{t_2-1} \mathbb{1}\big\{ -v^{\tau} \geqslant 2^{-\widehat{j}} \big\} \geqslant \sum_{\tau=t_1}^{t_2-1} \mathbb{1}\big\{ -v^{\tau} > 2^{-\widehat{j}} \big\} \geqslant \tau_{\widehat{j}+1} - t_1 \geqslant \tau_{\widehat{j}+1} - \tau_{\widehat{j}} \qquad \text{for any } 1 \leqslant \widehat{j} \leqslant j_{\max} - 1.$$

When $1 \leqslant \widetilde{j} \leqslant j_{\max} - 1$, the above display taken collectively with (120) gives

$$\sum_{\tau=t_1}^{t_2-1} \mathbb{1}\big\{ -v^{\tau} \geqslant 2^{-\widetilde{j}} \big\} \geqslant \tau_{\widetilde{j}+1} - \tau_{\widetilde{j}} \geqslant \frac{qx^2 \cdot 2^{\widetilde{j}-1}}{100\big(\log_2(1/\eta) + 1\big)^2 \eta}, \tag{121}$$

and as a result,

$$\sum_{\tau=t_1}^{t_2-1} \sum_{\widehat{j}=1}^{j_{\max}-1} \mathbb{1}\big\{ \max\{-v^{\tau}, 0\} \geqslant 2^{-\widehat{j}} \big\} 2^{-(\widehat{j}-1)} \geqslant \sum_{\tau=t_1}^{t_2-1} \mathbb{1}\big\{ -v^{\tau} \geqslant 2^{-\widetilde{j}} \big\} 2^{-(\widetilde{j}-1)} \geqslant \frac{qx^2}{100\big(\log_2(1/\eta) + 1\big)^2 \eta}. \tag{122}$$

By observing that

$$\sum_{\widehat{j}=1}^{\infty} \mathbb{1}\big\{ x \geqslant 2^{-\widehat{j}} \big\} \cdot 2^{-(\widehat{j}-1)} \leqslant 4x$$

holds for any $x \geqslant 0$, we can combine this fact with (122) to derive

$$4 \sum_{\tau=t_1}^{t_2-1} \max\{-v^{\tau}, 0\} \geqslant \frac{qx^2}{100\big(\log_2(1/\eta)) + 1\big)^2 \eta}. \tag{123}$$

Furthermore, recalling that $v^{\tau} \leqslant \varepsilon_1$ (cf. (59)), one can deduce that

$$\sum_{\tau=t_1}^{t_2-1} \max\{-v^{\tau}, 0\} = \sum_{\tau=t_1}^{t_2-1} (-v^{\tau}) - \sum_{\tau=t_1}^{t_2-1} \min\{-v^{\tau}, 0\} \leqslant \sum_{\tau=t_1}^{t_2-1} (-v^{\tau}) + (t_2 - t_1)\varepsilon_1,$$

which combined with (123) yields

$$4(t_2 - t_1)\varepsilon_1 + 4 \sum_{\tau=t_1}^{t_2-1} (-v^{\tau}) \geqslant \frac{qx^2}{100\big(\log_2(1/\eta) + 1\big)^2 \eta}.$$

The proof is thus complete.

C.5 Additional illustrative figures for segment construction

In this section, we provide several illustrative figures to help the readers understand the concept of "segments" and certain key properties, which play an important role in the segment construction described in Section 4.2.2 and the proof of Lemma 16.

- Figure 1 illustrates an example in which any two segments either coincide or are disjoint. In this case, our heuristic arguments in Section 4.2.2 about "shared intervals" and "disjoint intervals" become applicable.
- Figure 2 gives an example where two non-identical segments might overlap. Due to the presence of non-disjoint segments, our heuristic arguments in Section 4.2.2 about "shared intervals" and "disjoint intervals" are not readily applicable.
- In Figure 3, we provide an example of the partition of blocks (as in the proof of Lemma 15), whereas in Figure 4, we illustrate how to align one side of the segments using a common inner point (as in the proof of Lemma 16).
- In Figure 5 and Figure 6, we illustrate how to construct disjoint segments from a group of segments with common starting points in the case with $k = 8$. In this particular example, we have in total 5 groups of disjoint segments, marked with different colors.

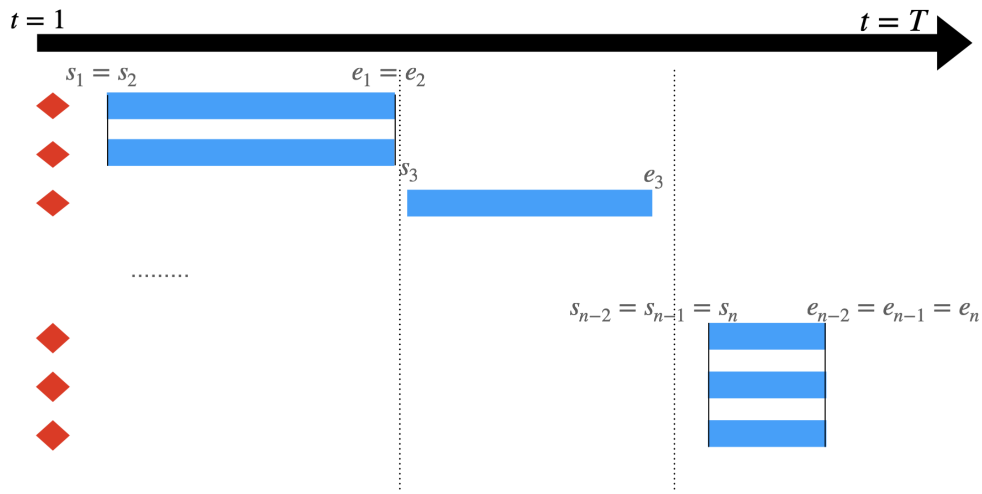

Figure 1: An example where any two segments either coincide or are disjoint.

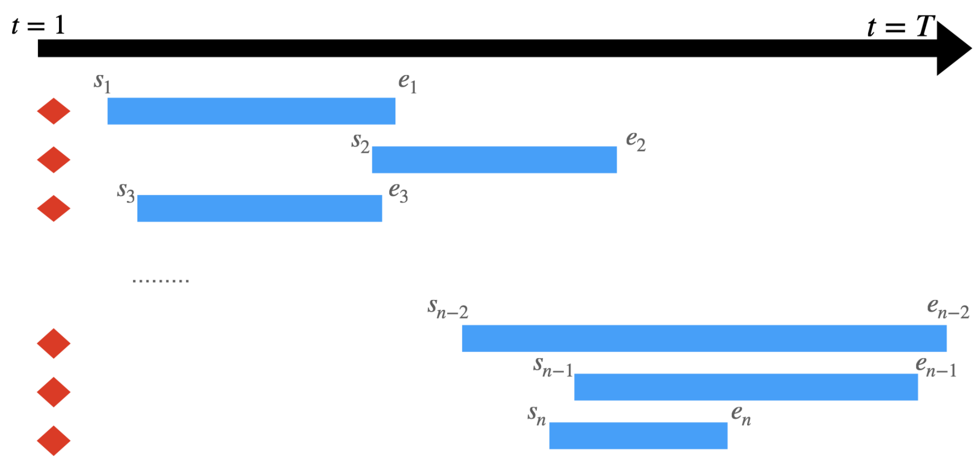

Figure 2: An example where two non-identical segments might overlap.

D Proofs of the lower bound in Theorem 2

D.1 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 a single element. Without loss of generality, we write $\ell(h, (x, y)) = \ell(h, x)$.

Our construction. We now introduce our construction, with the key components described below.

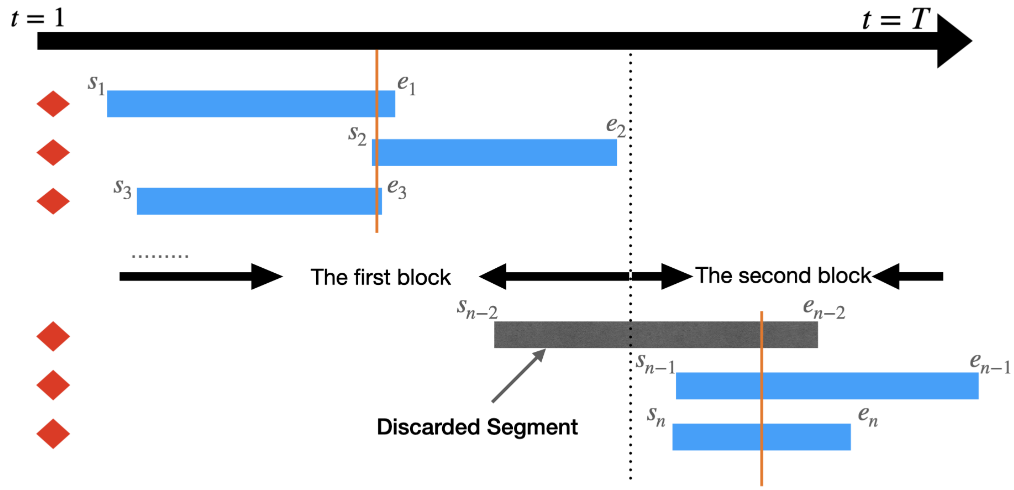

Figure 3: Partition of blocks as in the proof of Lemma 16.

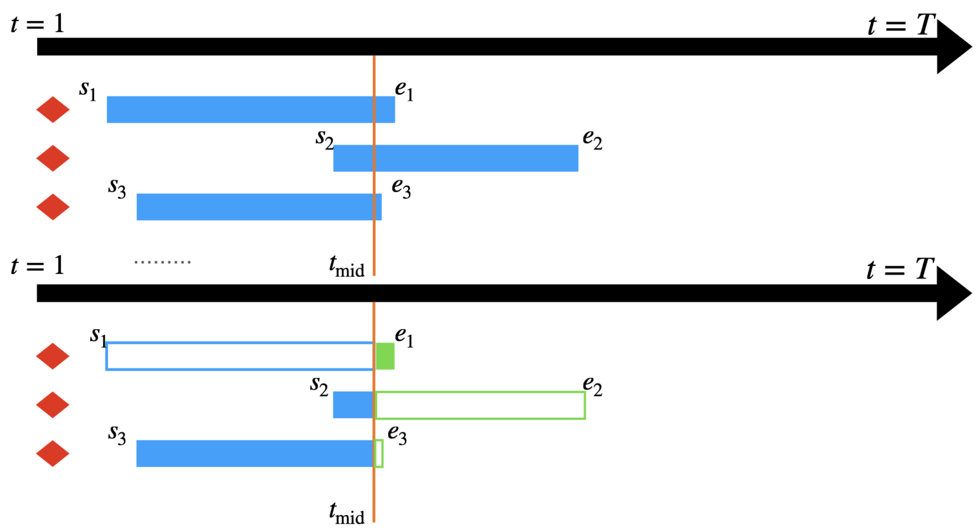

Figure 4: An illustration of how we use a common inner point (i.e., t_{mid} to align one side of the segments (as in the proof of Lemma 16). The unfilled part of the segments means that the weight changes from $w_i^{s_i}$ to $w_i^{t_{\text{mid}}}$ (i.e., $\log\left(w_i^{t_{\text{mid}}}/w_i^{s_i}\right)$) are not significant enough.

- *Hypothesis class and loss function.* There are N hypotheses in $\mathcal{H}$, where each hypothesis is assigned k unique dimensions (recall that $\mathcal{X}$ consists of (KN)-dimensional vectors). For each $h \in \mathcal{H}$, we let $\mathcal{I}_h = \{j_{h,i}\}_{i=1}^k$ denote the k dimensions assigned to h, so that $\mathcal{I}_h \cap \mathcal{I}_{h'} = \varnothing$ for $h \neq h'$. We then define $h(x)$ and $\ell(h, x)$ as follows:

$$h(x) = \ell(h, x) = \begin{cases} 0, & \text{if } x_i = 0 \text{ for all } i \in \mathcal{I}_h; \\ x_{i'} \text{ with } i' = \arg\min_{i \in \mathcal{I}_h, x_i \neq 0} x_i, & \text{else.} \end{cases}$$

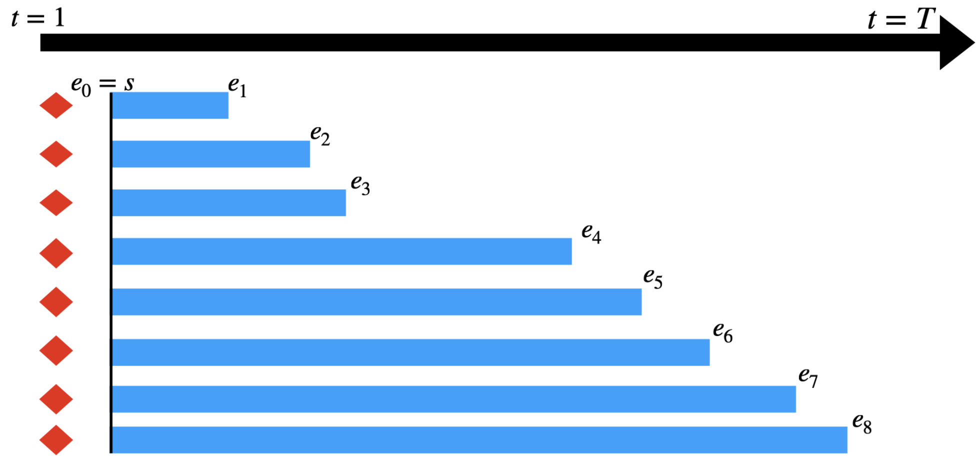

Figure 5: A group of segments with common starting points.

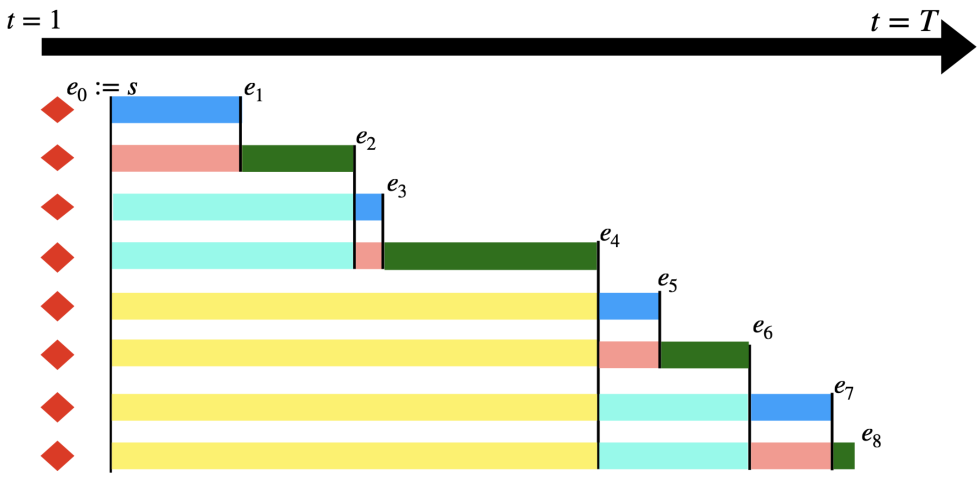

Figure 6: Construction of groups of disjoint segments, where each of the original segments is divided into at most $\log_2(k) + 1$ sub-segments marked with different colors.

Next, divide $\mathcal{H}$ arbitrarily into $(k + 1)$ disjoint subsets as

$$\mathcal{H} = (\cup_{i=1}^{k} \mathcal{H}_i) \cup \{h^\star\},$$

where each $\mathcal{H}_i$ contains N_0 hypotheses. In our construction, we intend to make $h^\star$ the unique optimal policy, and will design properly so that the hypothesis $h \in \mathcal{H}_i$ performs poorly on the i-th distribution $\mathcal{D}_i$ $(i \in [k])$.

- *Data distributions.* We design the k data distributions $\{\mathcal{D}_i\}_{i=1}^k$. For any given $i \in [k]$ and any $x \in \mathcal{X}$, we let

$$\mathbb{P}_{\mathcal{D}_i}\{x\} = \prod_{l=1}^{kN} \mathbb{P}_{\mathcal{D}_{i,l}}\{x_l\}$$

  be a product distribution, where

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_{i,l}}\{x_l\} = \mathbb{1}\{x_l = 0\}, && l \notin \{j_{h,i} \mid h \in \mathcal{H}\},\\
&\mathbb{P}_{\mathcal{D}_{i,l}}\{x_l\} = \frac{1}{2}\mathbb{1}\{x_l = 1\} + \frac{1}{2}\mathbb{1}\{x_l = -1\}, && l \in \{j_{h,i} \mid h \notin \mathcal{H}_i\},\\
&\mathbb{P}_{\mathcal{D}_{i,l}}\{x_l\} = \left(\frac{1}{2} + 4\varepsilon\right)\mathbb{1}\{x_l = 1\} + \left(\frac{1}{2} - 4\varepsilon\right)\mathbb{1}\{x_l = -1\}, && l \in \{j_{h,i} \mid h \in \mathcal{H}_i\}.
\end{aligned}$$

The above construction enjoys the following properties:

(i) $\ell(h, x) \in [-1, 1]$ for any $h \in \mathcal{H}$ and $x \in \mathcal{X}$;

(ii) $\mathbb{E}_{x\sim\mathcal{D}_i}[\ell(h, x)] = \mathbb{E}_{x\sim\mathcal{D}_{i,j_{h,i}}}[x_{j_{h,i}}] = 8\varepsilon \cdot \mathbb{1}\{h \in \mathcal{H}_i\}$ for any $i \in [k]$ and $h \in \mathcal{H}$;

(iii) the only ε-optimal hypothesis is $h^\star$, because for any $h \neq h^\star$, there exists some i such that $h \in \mathcal{H}_i$;

(iv) $h(x) \in \{-1, 1\}$ for $x \in \cup_{i=1}^k \mathrm{supp}(\mathcal{D}_i)$,[10] and $|\mathcal{H}| = N = kN_0 + 1 = 2^d$, which imply that

$$\mathsf{VC\text{-}dim}(\mathcal{H}) \leqslant \log_2(N) \leqslant d$$

over $\cup_{i=1}^k \mathrm{supp}(\mathcal{D}_i)$;

(v) $\ell(h, x)$ could be regarded as a function of $h(x)$ because $\ell(h, x) = h(x)$.

Sample complexity lower bound. Before proceeding, let us introduce the notation $\mathsf{Query}(\mathcal{D}_i)$ such that: for each call to $\mathsf{Query}(\mathcal{D}_i)$, we can obtain independent observations $\{x_{j_{h,i}}\}_{h\in\mathcal{H}}$ where $x_{j_{h,i}} \sim \mathcal{D}_{i,j_{h,i}}$ for each $h \in \mathcal{H}$. Now for $i \in [k]$, denote by M_i the number of calls to $\mathsf{Query}(\mathcal{D}_i)$. Our aim is to show that: in order to distinguish $h^\star$ from $\mathcal{H}_i$, the quantity 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 $h^\star$ with probability at least 3/4. Let $\mathbb{P}_{\mathcal{G}}\{\cdot\}$ and $\mathbb{E}_{\mathcal{G}}[\cdot]$ denote respectively the probability and expectation when Algorithm $\mathcal{G}$ is executed, and let h_{out} be the output hypothesis. It then holds that

$$\mathbb{P}_{\mathcal{G}}\{h_{\mathrm{out}} = h^\star\} \geqslant \frac{3}{4}.$$

Let $\Pi_{\mathcal{H}}$ be the set of permutations over $\mathcal{H}$, and let $\mathsf{Unif}(\Pi_{\mathcal{H}})$ be the uniform distribution over $\Pi(\mathcal{H})$. With slight abuse of notation, 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 $\sigma \sim \mathsf{Unif}(\Pi_{\mathcal{H}})$. Recognizing that $\mathcal{G}$ returns the optimal hypothesis with probability at least 3/4 for all problem instances, we can see that

$$\mathbb{P}_{\mathcal{G}'}\{h_{\mathrm{out}} = h^\star\} \geqslant \frac{3}{4}.$$

The lemma below then assists in bounding the probability of returning a sub-optimal hypothesis.

Lemma 18. *Consider $\widetilde{i} \in [k]$. If $\mathbb{P}_{\mathcal{G}'}\{h_{\mathrm{out}} = h^\star, M_i \leqslant m\} \geqslant 1/2$ for some $m \geqslant 0$, then for any $h \in \mathcal{H}_{\widetilde{i}}$, one has*

$$\mathbb{P}_{\mathcal{G}'}\big\{h_{\mathrm{out}} = h, M_{\widetilde{i}} \leqslant m\big\} \geqslant \frac{1}{2}\mathbb{P}_{\mathcal{G}'}\big\{h_{\mathrm{out}} = h^\star, M_{\widetilde{i}} \leqslant m\big\} \exp\big(-80\sqrt{m}\varepsilon - 40m\varepsilon^2\big),$$

and moreover, m necessarily exceeds $m \geqslant \frac{\log(N_0/4)}{30000\varepsilon^2}$.

[10] We denote by $\mathrm{supp}(\mathcal{D})$ the support of the distribution $\mathcal{D}$.

Proof. See Appendix D.2. ☐

In view of Lemma 18, we have

$$\mathbb{P}_{\mathcal{G}'}\left\{h_{\rm out}=h^\star, M_{\tilde{i}}<\frac{\log(N_0/4)}{30000\varepsilon^2}\right\}<\frac{1}{2}. \tag{124}$$

Observing that $\mathbb{P}_{\mathcal{G}'}\{h_{\rm out}=h^\star\}\geqslant 3/4$, we can derive

$$\begin{aligned}\mathbb{P}_{\mathcal{G}'}\left\{h_{\rm out}=h^\star, M_{\tilde{i}}\geqslant\frac{\log(N_0/4)}{30000\varepsilon^2}\right\}&=\mathbb{P}_{\mathcal{G}'}\left\{h_{\rm out}=h^\star\right\}-\mathbb{P}_{\mathcal{G}'}\left\{h_{\rm out}=h^\star, M_{\tilde{i}}<\frac{\log(N_0/4)}{30000\varepsilon^2}\right\}\\&>\frac{3}{4}-\frac{1}{2}=\frac{1}{4},\end{aligned}$$

which implies that

$$\begin{aligned}\mathbb{E}_{\mathcal{G}'}[M_{\tilde{i}}]&\geqslant\frac{\log(N_0/4)}{30000\varepsilon^2}\cdot\mathbb{P}_{\mathcal{G}'}\left\{M_{\tilde{i}}\geqslant\frac{\log(N_0/4)}{30000\varepsilon^2}\right\}\geqslant\frac{\log(N_0/4)}{30000\varepsilon^2}\cdot\mathbb{P}_{\mathcal{G}'}\left\{h_{\rm out}=h^\star, M_{\tilde{i}}\geqslant\frac{\log(N_0/4)}{30000\varepsilon^2}\right\}\\&\geqslant\frac{\log(N_0/4)}{120000\varepsilon^2}\geqslant\frac{d-\log_2(8k)}{120000\varepsilon^2}\geqslant\frac{d}{240000\varepsilon^2}.\end{aligned}$$

Summing over $i\in[k]$ gives

$$\mathbb{E}_{\mathcal{G}'}\left[\sum_{i=1}^{k}M_{\tilde{i}}\right]\geqslant\frac{dk}{240000\varepsilon^2}, \tag{125}$$

thereby concluding the proof.

D.2 Proof of Lemma 18

Consider any $h\in\mathcal{H}_{\tilde{i}}$ (and hence $h\neq h^\star$). For $v=[v_p]_{1\leqslant p\leqslant m}\in\{-1,1\}^m$, we let

$$n^+(v)=\sum_{p=1}^{m}\mathbb{1}\{v_p=1\}$$

denote the number of 1's in the coordinates of v. Let $\mathcal{V}$ be a subset of $\{-1,1\}^{2m}$ defined as

$$\mathcal{V}:=\left\{v^1,v^2\in\{-1,1\}^m\mid n^+(v^1)-n^+(v^2)\leqslant 4\sqrt{m}+2m\varepsilon\right\}.$$

Let $\mathbb{P}_{\mathcal{C}}\{\cdot\}$ denote the probability distribution of $\left\{\{x^l_{j_{h,\tilde{i}}}(\tilde{i})\}_{l=1}^m,\{x^l_{j_{h^\star,\tilde{i}}}(\tilde{i})\}_{l=1}^m\right\}$, and $\mathbb{P}_{\mathcal{C}'}\{\cdot\}$ the probability distribution of $\left\{\{x^l_{j_{h^\star,\tilde{i}}}(\tilde{i})\}_{l=1}^m,\{x^l_{j_{h,\tilde{i}}}(\tilde{i})\}_{l=1}^m\right\}$. Hoeffding's inequality then tells us that

$$\mathbb{P}_{\mathcal{C}}\{\mathcal{V}\}\geqslant\frac{3}{4}.$$

Also, observe that

$$\begin{aligned}&\mathbb{P}_{\mathcal{G}'}\left\{h_{\rm out}=h^\star, M_{\tilde{i}}\leqslant m,\left\{\{x^l_{j_{h,\tilde{i}}}(\tilde{i})\}_{l=1}^m,\{x^l_{j_{h^\star,\tilde{i}}}(\tilde{i})\}_{l=1}^m\right\}\in\mathcal{V}\right\}\\&\quad\geqslant\mathbb{P}_{\mathcal{G}'}\left\{h_{\rm out}=h^\star, M_{\tilde{i}}\leqslant m\right\}-\left(1-\mathbb{P}_{\mathcal{C}}\left\{\mathcal{V}\right\}\right)\geqslant\frac{3}{4}-\left(1-\frac{3}{4}\right)=\frac{1}{2},\end{aligned} \tag{126}$$

namely,

$$\sum_{v=\{v^1,v^2\}\in\mathcal{V}}\mathbb{P}_{\mathcal{G}'}\left\{h_{\rm out}=h^\star, M_{\tilde{i}}\leqslant m,\{x^l_{j_{h,\tilde{i}}}(\tilde{i})\}_{l=1}^m=v^1,\{x^l_{j_{h^\star,\tilde{i}}}(\tilde{i})\}_{l=1}^m=v^2\right\}\geqslant\frac{1}{2}.$$

In addition, for any $v = \{v^1, v^2\} \in \mathcal{V}$, it is readily seen that

$$
\begin{aligned}
\mathbb{P}_{\mathcal{C}'}\{v\} &= \mathbb{P}_{\mathcal{C}}\{v\} \cdot (1-8\varepsilon)^{n^+(v^1)-n^+(v^2)} (1+8\varepsilon)^{n^+(v^2)-n^+(v^1)} \\
&= \mathbb{P}_{\mathcal{C}}\{v\} \left(\frac{1-8\varepsilon}{1+8\varepsilon}\right)^{n^+(v^1)-n^+(v^2)} \\
&\geqslant \mathbb{P}_{\mathcal{C}}\{v\} \exp\Big(-20\big(n^+(v^1)-n^+(v^2)\big)\varepsilon\Big) \\
&\geqslant \mathbb{P}_{\mathcal{C}}\{v\} \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2),
\end{aligned}
$$

where the last line follows from the definition of $\mathcal{V}$. As a result, we can demonstrate that, for any $v = \{v^1, v^2\} \in \mathcal{V}$,

$$
\begin{aligned}
&\mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h, M_{\widetilde{i}} \leqslant m, \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^1, \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^2\Big\} \\
&= \mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h, M_{\widetilde{i}} \leqslant m \mid \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^1, \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^2\Big\} \mathbb{P}_{\mathcal{C}'}\{v\} \\
&\geqslant \mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h, M_{\widetilde{i}} \leqslant m \mid \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^1, \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^2\Big\} \mathbb{P}_{\mathcal{C}}\{v\} \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2) \\
&= \mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h^\star, M_{\widetilde{i}} \leqslant m \mid \{x^\ell_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{\ell=1}^m = v^1, \{x^\ell_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{\ell=1}^m = v^2\Big\} \mathbb{P}_{\mathcal{C}}\{v\} \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2) \qquad (127) \\
&= \mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h^\star, M_{\widetilde{i}} \leqslant m, \{x^\ell_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{\ell=1}^m = v^1, \{x^\ell_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{\ell=1}^m = v^2\Big\} \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2).
\end{aligned}
$$

To see why (127) holds, observe that (here, for any $1 \leqslant l \leqslant m$, let v^1_l (resp. v^2_l) be the l-th coordinate of v^1 (resp. v^2)):

$$
\begin{aligned}
&\mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h, M_{\widetilde{i}} \leqslant m \mid \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^1, \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^2\Big\} \\
&= \sum_{m'=1}^{m} \mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h, M_{\widetilde{i}} = m' \mid \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^1, \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^2\Big\} \\
&= \sum_{m'=1}^{m} \mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h, M_{\widetilde{i}} = m' \mid \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m'} = \{v^1_l\}_{l=1}^{m'}, \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m'} = \{v^2_l\}_{l=1}^{m'}\Big\} \qquad (128) \\
&= \sum_{m'=1}^{m} \mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h^\star, M_{\widetilde{i}} = m' \mid \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m'} = \{v^1_l\}_{l=1}^{m'}, \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m'} = \{v^2_l\}_{l=1}^{m'}\Big\} \qquad (129) \\
&= \sum_{m'=1}^{m} \mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h^\star, M_{\widetilde{i}} = m' \mid \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^1, \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^2\Big\} \qquad (130) \\
&= \mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h^\star, M_{\widetilde{i}} \leqslant m \mid \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^1, \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^2\Big\}.
\end{aligned}
$$

where (129) results from Lemma 19 in Appendix D.3, and (128) and (130) hold since for $h' = h, h^\star$, the event $\{h_{\mathrm{out}} = h', M_{\widetilde{i}} = m'\}$ is independent of $\{x^l(\widetilde{i})\}_{l \geqslant m'+1}$. Taking the sum over $v = \{v^1, v^2\} \in \mathcal{V}$, we obtain

$$
\begin{aligned}
&\mathbb{P}_{\mathcal{G}'}\{h_{\mathrm{out}} = h, M_{\widetilde{i}} \leqslant m\} \\
&\geqslant \sum_{v \in \mathcal{V}} \mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h, M_{\widetilde{i}} \leqslant m, \{x^\ell_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{\ell=1}^m = v^1, \{x^\ell_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{\ell=1}^m = v^2\Big\} \\
&\geqslant \sum_{v \in \mathcal{V}} \mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h^\star, M_{\widetilde{i}} \leqslant m, \{x^\ell_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{\ell=1}^m = v^1, \{x^\ell_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{\ell=1}^m = v^2\Big\} \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2) \\
&= \mathbb{P}_{\mathcal{G}'}\Big\{h_{\mathrm{out}} = h^\star, M_{\widetilde{i}} \leqslant m, \Big\{\{x^\ell_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{\ell=1}^m, \{x^\ell_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{\ell=1}^m\Big\} \in \mathcal{V}\Big\} \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2) \\
&\geqslant \frac{1}{2}\mathbb{P}_{\mathcal{G}'}\{h_{\mathrm{out}} = h^\star, M_{\widetilde{i}} \leqslant m\} \exp(-80\sqrt{m}\varepsilon - 40m\varepsilon^2),
\end{aligned}
$$

where the last line arises from (126).

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

$$1 \geqslant \mathbb{P}_{\mathcal{G}'}\big\{h_{\text{out}} \in \mathcal{H}_{\widetilde{i}}, M_{\widetilde{i}} \leqslant m\big\} \geqslant \frac{N_0}{2}\mathbb{P}_{\mathcal{G}'}\big\{h_{\text{out}} = h^\star, M_{\widetilde{i}} \leqslant m\big\}\exp(-40m\varepsilon^2 - 80\sqrt{m}\varepsilon), \tag{131}$$

given that each $\mathcal{H}_{\widetilde{i}}$ contains N_0 hypotheses. This in turn reveals that

$$\frac{1}{2} \leqslant \mathbb{P}_{\mathcal{G}'}\left\{h_{\text{out}} = h^\star, M_{\widetilde{i}} \leqslant m\right\} \leqslant \frac{2}{N_0}\exp(40m\varepsilon^2 + 80\sqrt{m}\varepsilon). \tag{132}$$

Consequently, we arrive at

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

which implies that

$$m \geqslant \min\left\{\frac{\log(N_0/4)}{80\varepsilon^2}, \frac{\log^2(N_0/4)}{30000\varepsilon^2}\right\} \geqslant \frac{\log(N_0/4)}{30000\varepsilon^2}.$$

D.3 Statement and proof of Lemma 19

Lemma 19. *For any $i \in [k]$ and $l \geqslant 1$, let $x^l(i)$ denote the l-th sample from $\mathcal{D}_i$. For any $\widetilde{i} \in [k]$, $h \in \mathcal{H}_{\widetilde{i}}$, $m > 0$, and $v^1, v^2 \in \{-1,1\}^{2m}$, one has*

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

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

Consider a given sequence $\{m_i\}_{i=1}^k$. Let $X(i) = \{X^l(i)\}_{l=1}^{m_i} \in \{-1,0,1\}^{kNm_i}$ for $i \in [k]$, and let $x(i) = \{x^l(i)\}_{l=1}^{m_i}$ be the datapoints of the first m_i calls to $\mathsf{Query}(D_i)$. With slight abuse of notation, we take $\sigma(x(i)) = \{\sigma(x^l(i))\}_{l=1}^{m_i}$ for each $i \in [k]$. It then follows from Lemma 20 in Appendix D.4 that

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

which implies that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{G},\sigma(\mathcal{H})}\left\{h_{\text{out}} = h^\star, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k, \sigma(x(i)) = X(i), \forall i \in [k]\right\} \\
&= \mathbb{P}_{\mathcal{G},\overline{\sigma}\sigma(\mathcal{H})}\left\{h_{\text{out}} = h, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k, \overline{\sigma}\sigma(x(i)) = X(i), \forall i \in [k]\right\} \cdot \frac{\mathbb{P}\left\{\sigma(x(i)) = X(i), \forall i \in [k]\right\}}{\mathbb{P}\left\{\overline{\sigma}\sigma(x(i)) = X(i), \forall i \in [k]\right\}} \\
&= \mathbb{P}_{\mathcal{G},\overline{\sigma}\sigma(\mathcal{H})}\left\{h_{\text{out}} = h, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k, \overline{\sigma}\sigma(x(i)) = X(i), \forall i \in [k]\right\} \cdot \frac{\mathbb{P}\left\{\sigma(x(\widetilde{i})) = X(\widetilde{i})\right\}}{\mathbb{P}\left\{\overline{\sigma}\sigma(x(\widetilde{i})) = X(\widetilde{i})\right\}} \\
&= \mathbb{P}_{\mathcal{G},\overline{\sigma}\sigma(\mathcal{H})}\left\{h_{\text{out}} = h, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k, \overline{\sigma}\sigma(x(i))) = X(i), \forall i \in [k]\right\} \\
&\qquad \cdot \frac{\mathbb{P}\left\{\{x^l_{j_{h,i}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = \{X^l_{j_{\sigma(h),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}}, \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = \{X^l_{j_{\sigma(h^\star),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}}\right\}}{\mathbb{P}\left\{\{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = \{X^l_{j_{\sigma(h),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}}, \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = \{X^l_{j_{\sigma(h^\star),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}}\right\}}.
\end{aligned} \tag{133}$$

Rearrange the equation to arrive at

$$\begin{aligned}
\mathbb{P}_{\mathcal{G},\sigma(\mathcal{H})}\Bigg\{&h_{\text{out}} = h^\star, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k, \sigma(x(i)) = X(i), \forall i \in [k] \\
&\Bigg| \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = \{X^l_{j_{\sigma(h),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}}, \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = \{X^l_{j_{\sigma(h^\star),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}}\Bigg\}
\end{aligned}$$

$$= \mathbb{P}_{\mathcal{G},\overline{\sigma}\sigma(\mathcal{H})}\Big\{ h_{\text{out}} = h, \{M_i\}_{i=1}^k = \{m_i\}_{i=1}^k, \overline{\sigma}\sigma(x(i)) = X(i), \forall i \in [k]$$

$$\Big| \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = \{X^l_{j_{\sigma(h),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}}, \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = \{X^l_{j_{\sigma(h^\star),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} \Big\}.$$

Taking the sum over all possible choices of $\{X(i)\}_{i \neq \widetilde{i}}$, $\Big\{\{X^l_{j_{h',i'}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}}\Big\}_{i' \in [k], h' \notin \{h,h^\star\}}$, $\Big\{\{X^l_{j_{h',i'}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}}\Big\}_{h' \in \{h,h^\star\}, i' \neq \widetilde{i}}$ and $\{m_i\}_{i \neq \widetilde{i}}$, we reach

$$\mathbb{P}_{\mathcal{G},\sigma(\mathcal{H})}\Big\{ h_{\text{out}} = h^\star, M_{\widetilde{i}} = m_{\widetilde{i}}$$

$$\Big| \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = \{X^l_{j_{\sigma(h),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}}, \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = \{X^l_{j_{\sigma(h^\star),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} \Big\}$$

$$= \mathbb{P}_{\mathcal{G},\overline{\sigma}\sigma(\mathcal{H})}\Big\{ h_{\text{out}} = h, M_{\widetilde{i}} = m_{\widetilde{i}}$$

$$\Big| \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = \{X^l_{j_{\sigma(h),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}}, \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = \{X^l_{j_{\sigma(h^\star),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} \Big\}$$

for any $X(\widetilde{i}) \in \{-1,0,1\}^{kNm_{\widetilde{i}}}$.

Fix $m_{\widetilde{i}} = m$, and choose $\{X^l_{j_{\sigma(h),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = v_1$, $\{X^l_{j_{\sigma(h^\star),\widetilde{i}}}(\widetilde{i})\}_{l=1}^{m_{\widetilde{i}}} = v_2$. We then have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{G}',\mathcal{H}}\Big\{ h_{\text{out}} = h^\star, M_i = m \mid \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^1, \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^2 \Big\} \\
&= \frac{1}{|\Pi_{\mathcal{H}}|}\sum_{\sigma \in \Pi_{\mathcal{H}}} \mathbb{P}_{\mathcal{G},\sigma(\mathcal{H})}\Big\{ h_{\text{out}} = h^\star, M_i = m \mid \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^1, \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^2 \Big\} \\
&= \frac{1}{|\Pi_{\mathcal{H}}|}\sum_{\sigma \in \Pi_{\mathcal{H}}} \mathbb{P}_{\mathcal{G},\overline{\sigma}\sigma(\mathcal{H})}\Big\{ h_{\text{out}} = \overline{\sigma}^{-1}(h^\star), M_i = m \mid \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^2, \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^1 \Big\} \\
&= \frac{1}{|\Pi_{\mathcal{H}}|}\sum_{\sigma \in \Pi_{\mathcal{H}}} \mathbb{P}_{\mathcal{G},\sigma(\mathcal{H})}\Big\{ h_{\text{out}} = h, M_i = m \mid \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^2, \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^1 \Big\} \\
&= \mathbb{P}_{\mathcal{G}',\mathcal{H}}\Big\{ h_{\text{out}} = h, M_i = m \mid \{x^l_{j_{h,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^2, \{x^l_{j_{h^\star,\widetilde{i}}}(\widetilde{i})\}_{l=1}^m = v^1 \Big\}.
\end{aligned}$$

This completes the proof. □

D.4 Statement and proof of Lemma 20

Lemma 20. *Consider any $\{m_i\}_{i \in [k]}$, $\sigma \in \Pi_{\mathcal{H}}$ 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 in Lemma 19. It then holds that*

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

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

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

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

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

E Proofs of auxiliary lemmas for Rademacher classes

E.1 Proof of Lemma 6

In this subsection, we shall follow the notation adopted in the proof of Lemma 1 (e.g., the dataset $\widetilde{\mathcal{S}}$, the data subset $\widetilde{\mathcal{S}}(n)$ and its independent copy $\widetilde{\mathcal{S}}^+(n)$, the Rademacher random variables $\{\sigma_i^j\}$).

Consider any given $n = \{n_i\}_{i=1}^k$ obeying $n_i \geqslant 12\log(2k)$ for all $i \in [k]$, and any given $w \in \Delta(k)$. Let $\kappa = \min_i \frac{n_i}{w_i}$. Recall that $(x_{i,j}, y_{i,j})$ is the j-th sample from $\mathcal{D}_i$, and $\{(x_{i,j}^+, y_{i,j}^+\}$ are independent copies. Define

$$
F(n,w) \coloneqq \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}(n)}\left[\max_{h\in\mathcal{H}}\left(\sum_{i=1}^k \frac{w_i}{n_i}\sum_{j=1}^{n_i}\ell\big(h,(x_{i,j},y_{i,j})\big) - \sum_{i=1}^k w_i L(h, \mathrm{e}_i^{\mathsf{basis}})\right)\right],
$$

which can be upper bounded by

$$
\begin{aligned}
F(n,w) &= \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}(n)}\left[\max_{h\in\mathcal{H}}\left(\sum_{i=1}^k \frac{w_i}{n_i}\sum_{j=1}^{n_i}\left(\ell\big(h,(x_{i,j},y_{i,j})\big) - \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}^+(n)}\left[\ell\big(h,(x_{i,j}^+,y_{i,j}^+)\big)\right]\right)\right)\right] \\
&\leqslant \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}(n),\widetilde{\mathcal{S}}^+(n)}\left[\max_{h\in\mathcal{H}}\left(\sum_{i=1}^k \frac{w_i}{n_i}\sum_{j=1}^{n_i}\left(\ell\big(h,(x_{i,j},y_{i,j})\big) - \ell\big(h,(x_{i,j}^+,y_{i,j}^+)\big)\right)\right)\right] \\
&= \mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}(n),\widetilde{\mathcal{S}}^+(n),\{\{\sigma_i^j\}_{j=1}^{n_i}\}_{i=1}^k}\left[\max_{h\in\mathcal{H}}\left(\sum_{i=1}^k \frac{w_i}{n_i}\sum_{j=1}^{n_i}\sigma_i^j\left\{\ell\big(h,(x_{i,j},y_{i,j})\big) - \ell\big(h,(x_{i,j}^+,y_{i,j}^+)\big)\right\}\right)\right] \\
&\leqslant 2\mathop{\mathbb{E}}_{\widetilde{\mathcal{S}}(n),\{\{\sigma_i^j\}_{j=1}^{n_i}\}_{i=1}^k}\left[\max_{h\in\mathcal{H}}\left(\sum_{i=1}^k \frac{1}{\kappa}\sum_{j=1}^{n_i}\sigma_i^j\ell\big(h,(x_{i,j},y_{i,j})\big)\right)\right] \\
&= 2\frac{\sum_{i=1}^k n_i}{\kappa}\widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k}. \qquad (137)
\end{aligned}
$$

Here, the first inequality arises from Jensen's inequality, whereas the penultimate line applies Lemma 12.

Invoking the Mcdiarmid inequality (see Lemma 9) with the choice $c = 1/\kappa$, we obtain

$$
\mathbb{P}\left\{\left|\max_{h\in\mathcal{H}}\left(\sum_{i=1}^k \frac{w_i}{n_i}\sum_{j=1}^{n_i}\ell\big(h,(x_{i,j},y_{i,j})\big) - \sum_{i=1}^k w_i L(h,\mathrm{e}_i^{\mathsf{basis}})\right) - F(n,w)\right| \geqslant \varepsilon\right\} \leqslant 2\exp\left(-\frac{2\kappa^2\varepsilon^2}{\sum_{i=1}^k n_i}\right). \qquad (138)
$$

This taken together with (137) reveals that: for any $\delta' \in (0,1]$, with probability at least $1-\delta'$ we have

$$
\max_{h\in\mathcal{H}}\left(\sum_{i=1}^k \frac{w_i}{n_i}\ell(h,(x_{i,j},y_{i,j})) - \sum_{i=1}^k w_i L(h,\mathrm{e}_i^{\mathsf{basis}})\right) \leqslant 2\frac{\sum_{i=1}^k n_i}{\kappa}\widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k} + \frac{\sum_{i=1}^k n_i}{\kappa}\sqrt{\frac{\log(2/\delta')}{2\sum_{i=1}^k n_i}}. \qquad (139)
$$

Evidently, this inequality continues to hold if we replace (ℓ, L) with $(-\ell, -L)$. As a consequence, with probability at least $1-2\delta'$ one has

$$
\max_{h\in\mathcal{H}}\left|\sum_{i=1}^k \frac{w_i}{n_i}\ell\big(h,(x_{i,j},y_{i,j})\big) - \sum_{i=1}^k w_i L(h,\mathrm{e}_i^{\mathsf{basis}})\right| \leqslant 2\frac{\sum_{i=1}^k n_i}{\kappa}\widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k} + \frac{\sum_{i=1}^k n_i}{\kappa}\sqrt{\frac{\log(2/\delta')}{2\sum_{i=1}^k n_i}}. \qquad (140)
$$

Now, fix $\kappa \geqslant 0$, and define

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

where $\Delta_{\varepsilon_1/(8k)}(k)$ (i.e., an $\varepsilon_1/(8k)$-net of $\Delta(k)$) has been defined in Appendix B.1. Inequality (140) combined with the union bound tells us that: for any $\delta' > 0$, with probability at least $1-\delta'$

$$\begin{aligned}
\max_{h\in\mathcal{H}}&\left|\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{w_i}{n_i}\ell\big(h,(x_{i,j},y_{i,j})\big)-\sum_{i=1}^{k}w_iL(h,\mathrm{e}_i^{\mathsf{basis}})\right|\\
&\leqslant\frac{\sum_{i=1}^{k}n_i}{T_1/2}\widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k}+\frac{\sum_{i=1}^{k}n_i}{T_1/2}\sqrt{\frac{\log(|\widetilde{\mathcal{L}}|)+\log(2/\delta')}{2\sum_{i=1}^{k}n_i}}\\
&\leqslant 4\widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k}+4\sqrt{\frac{\log(2|\widetilde{\mathcal{L}}|)+\log(2/\delta')}{T_1}}\\
&\leqslant 4\widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k}+4\sqrt{\frac{2k\log(16kT_1/\varepsilon_1)+\log(2/\delta')}{T_1}}\\
&\leqslant 600C_{T_1}+4\sqrt{\frac{2k\log(16kT_1/\varepsilon_1)+\log(2/\delta')}{T_1}}
\end{aligned}$$

holds simultaneously for all $\{n,w\}\in\widetilde{\mathcal{L}}$; see the use of the union bound in Appendix B.1 too. Here, we have made use of Assumption 1, Lemma 4, Lemma 5 and the fact that $\sum_{i=1}^{k}n_i \geqslant T_1/2$.

Note that in Algorithm 3, we take

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

Given our choice that $n_i^{t,\mathsf{rad}}=\min\{\lceil T_1w_i^t+12\log(2k)\rceil,T_1\}$ for $i\in[k]$, we see that

$$T_1w_i^t\leqslant n_i^{t,\mathsf{rad}}-1\qquad\text{and}\qquad 12\log(2k)\leqslant n_i^{t,\mathsf{rad}}\leqslant T_1$$

for all $i\in[k]$. In addition, it is seen from our choice of $n_i^{t,\mathsf{rad}}$ and T_1 that

$$\sum_{i=1}^{k}n_i^{t,\mathsf{rad}}\leqslant\sum_{i=1}^{k}\left\lceil T_1w_i^t+12\log(2k)\right\rceil\leqslant T_1+k+12k\log(2k)\leqslant 2T_1-2.$$

Therefore, there exists some $\widetilde{w}^t\in\Delta(k)$ satisfying

$$\big\{\{n_i^{t,\mathsf{rad}}\}_{i=1}^k,\widetilde{w}^t\big\}\in\widetilde{\mathcal{L}}\qquad\text{and}\qquad\|w^t-\widetilde{w}^t\|_1\leqslant\frac{\varepsilon_1}{8k}$$

for each $1\leqslant t\leqslant T$. Taking $\delta'=\delta/4$, we obtain that with probability at least $1-\delta/4$,

$$\begin{aligned}
\max_{h\in\mathcal{H}}\left|\widehat{L}^t(h,w^t)-L(h,w^t)\right|&\leqslant\max_{h\in\mathcal{H}}\left|\widehat{L}^t(h,\widetilde{w}^t)-L(h,\widetilde{w}^t)\right|+\max_{h\in\mathcal{H}}\left|\widehat{L}^t(h,\widetilde{w}^t)-\widehat{L}^t(h,w^t)\right|\\
&\quad+\max_{h\in\mathcal{H}}\left|L(h,\widetilde{w}^t)-L(h,w^t)\right|\\
&\leqslant 600C_{T_1}+4\sqrt{\frac{2k\log(16kT_1/\varepsilon_1)+\log(2/\delta')}{T_1}}+\frac{\varepsilon_1}{8k}+\frac{\varepsilon_1}{8k}\\
&\leqslant\frac{\varepsilon_1}{2}
\end{aligned}$$

for any $1\leqslant t\leqslant T$, where the last inequality results from the definition of T_1.

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

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

which concludes the proof.

E.2 Proof of Lemma 4

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

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

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

E.3 Proof of Lemma 5

Set $n = \sum_{i=1}^k n_i$. Let $\{X_j\}_{j=1}^n$ be n i.i.d. multinomial random variables with parameter $\{w_i\}_{i=1}^k$, and take

$$
\widehat{n}_i = \sum_{j=1}^n \mathbb{1}\{X_j = i\}
$$

for each $i \in [k]$. From (34) and Definition 2, it is easily seen that

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

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 8 tells us that: for any $i \in [k]$, one has

$$
\widehat{n}_i \geqslant \frac{1}{3} n_i - 2\log(2k) \geqslant \frac{1}{6} n_i \qquad \Longrightarrow \qquad \widehat{n}_i \geqslant \left\lceil \frac{1}{6} n_i \right\rceil =: \widetilde{n}_i \tag{141}
$$

with probability exceeding $1 - 1/(2k)$. Defining $\mathcal{E}$ to be the event that $\widehat{n}_i \geqslant n_i/6$ holds for all $i \in [k]$, we can invoke the union bound to see that

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

Consequently, we can derive

$$
\mathsf{Rad}_n\big(\mathcal{D}(w)\big) \geqslant \mathbb{P}(\mathcal{E}) \cdot \mathop{\mathbb{E}}_{\{X_i\}_{i=1}^n} \left[\mathop{\mathbb{E}}_{\{z_i^j\}_{j=1}^{\widehat{n}_i}, \forall i \in [k]} \left[\mathop{\mathbb{E}}_{\{\sigma_i^j\}_{j=1}^{\widehat{n}_i}, \forall i \in [k]} \left[\frac{1}{n} \max_{h \in \mathcal{H}} \sum_{i=1}^k \sum_{j=1}^{\widehat{n}_i} \sigma_i^j \ell(h, z_i^j) \mid \mathcal{E} \right] \right] \right]
$$

$$\geqslant \frac{1}{2} \cdot \mathop{\mathbb{E}}_{\{z_i^j\}_{j=1}^{\hat{n}_i}, \forall i \in [k]} \left[\mathop{\mathbb{E}}_{\{\sigma_i^j\}_{j=1}^{\hat{n}_i}, \forall i \in [k]} \left[\frac{1}{n} \max_{h \in \mathcal{H}} \sum_{i=1}^{k} \sum_{j=1}^{\tilde{n}_i} \sigma_i^j \ell\big(h, z_i^j\big) \mid \mathcal{E} \right] \right]$$

$$= \frac{1}{2} \cdot \frac{\sum_{i=1}^{k} \tilde{n}_i}{n} \widetilde{\mathsf{Rad}}_{\{\tilde{n}_i\}_{i=1}^{k}}$$

$$\geqslant \frac{1}{12} \cdot \frac{1}{6} \widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^{k}}, \tag{142}$$

thus concluding the proof.

E.4 Necessity of Assumption 1

In this subsection, we study whether Assumption 1 can be replaced by the following weaker assumption, the latter of which only assumes that the Rademacher complexity on each $\mathcal{D}_i$ is well-bounded.

Assumption 2. *For each $n \geqslant 1$, there exists a quantity $\widetilde{C}_n > 0$ (known to the learner) such that*

$$\widetilde{C}_n \geqslant \max_{1 \leqslant i \leqslant k} \mathsf{Rad}_n(\mathcal{D}_i). \tag{143}$$

Formally, we have the following results.

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

$$\mathsf{Rad}_n\big(\mathcal{D}(w^0)\big) \geqslant \Omega\left(\frac{1}{k}\sum_{i=1}^{k} \mathsf{Rad}_{n/k}(\mathcal{D}_i)\right) \tag{144}$$

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

Proof. Without loss of generality, consider the case where $\mathcal{Y} = \{0\}$ and $\ell(h,(x,y)) = h(x) - y = h(x)$. We can then view $\mathcal{D}_i$ as a distribution over $\mathcal{X}_i$, as 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 on $\mathcal{X}_i$. In addition, we define $\mathcal{H}_i$ to be a set of hypothesis obeying $h(x) = 0$ for all $x \notin \mathcal{X}_i$ for each $i \in [k]$. For a collection of hypothesis $\{h_i\}_{i=1}^k$ such that $h_i \in \mathcal{H}_i$, we define $\mathsf{joint}(\{h_i\}_{i=1}^k)$ to be the hypothesis h such that

$$h(x) = \begin{cases} h_i(x) & \text{if } x \in \mathcal{X}_i,\ i \in [k]; \\ 0 & \text{if } x \notin \cup_i \mathcal{X}_i. \end{cases} \tag{145}$$

The hypothesis set $\mathcal{H}$ is then constructed as

$$\mathcal{H} = \big\{\mathsf{joint}(\{h_i\}_{i=1}^k) \mid h_i \in \mathcal{H}_i, \forall i \in [k]\big\}. \tag{146}$$

Recalling the definition of $\widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k}$, we see that

$$\begin{aligned}
\widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^{k}} &= \mathop{\mathbb{E}}_{\{x_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\mathop{\mathbb{E}}_{\{\sigma_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\frac{1}{\sum_{i=1}^k n_i} \max_{h \in \mathcal{H}} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \sigma_i^j h(x_i^j) \right] \right] \\
&= \mathop{\mathbb{E}}_{\{x_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\mathop{\mathbb{E}}_{\{\sigma_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^{k} \max_{h_i \in \mathcal{H}_i} \sum_{j=1}^{n_i} \sigma_i^j h_i(x_i^j) \right] \right] \\
&= \frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^{k} n_i \mathop{\mathbb{E}}_{\{z_i^j\}_{j=1}^{n_i}, \forall i \in [k]} \left[\mathop{\mathbb{E}}_{\{\sigma_i^j\}_{j=1}^{n_i}} \left[\frac{1}{n_i} \max_{h_i \in \mathcal{H}_i} \sum_{j=1}^{n_i} \sigma_i^j h_i(x_i^j) \right] \right]
\end{aligned} \tag{147}$$

$$= \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i \mathsf{Rad}_{n_i}(\mathcal{D}_i),$$

where (147) results from the definition of $\mathcal{H}$. By taking $n_i = \frac{n}{k}$ for all $i \in [k]$ and applying Lemma 5, we reach

$$\frac{1}{k} \sum_{i=1}^{k} \mathsf{Rad}_{n/k}(\mathcal{D}_i) = \widetilde{\mathsf{Rad}}_{\{n_i\}_{i=1}^k} \leqslant 72\mathsf{Rad}_n\big(\mathcal{D}(w^0)\big) \tag{148}$$

as claimed. □

By virtue of Lemma 21, if we set $\widetilde{C}_n = \widetilde{\Theta}(\sqrt{d/n})$ in Assumption 2, then the best possible upper bound on $\mathsf{Rad}_n(\mathcal{D}(w^0))$ is $\mathsf{Rad}_n(\mathcal{D}(w^0)) = \widetilde{\Theta}(\sqrt{dk/n})$, which implies that more samples are needed to learn the mixed distribution $\mathcal{D}(w^0)$ than learning each individual distribution. Moreover, under the construction in Lemma 21, if we further assume that $\min_{h_i \in \mathcal{H}_i} \mathbb{E}_{x\sim\mathcal{D}_i}[h_i(x)] = 1/2$ for all $i \in [k]$, then to find h such that

$$\max_{1\leqslant i\leqslant k} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}[h(x)] \leqslant \frac{1}{2} + \varepsilon, \tag{149}$$

we need to find, for each $i \in [k]$, a hypothesis $h_i \in \mathcal{H}_i$ such that

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}[h_i(x)] \leqslant \frac{1}{2} + \varepsilon. \tag{150}$$

Following this intuition, we can construct a counter example under Assumption 2, with a formal theorem stated as follows.

Theorem 4. *There exist a group of distributions $\{\mathcal{D}_i\}_{i=1}^k$ and a hypothesis set $\mathcal{H}$ such that Assumption 2 holds with $\widetilde{C}_n = \widetilde{O}\left(\sqrt{\frac{d}{n}}\right)$, and it takes at least $\widetilde{\Omega}\left(\frac{dk}{\varepsilon^2}\right)$ samples to find some $h \in \mathcal{H}$ obeying*

$$\max_{i\in[k]} L(h, \mathrm{e}_i^{\mathsf{basis}}) \leqslant \min_{h'\in\mathcal{H}} \max_{i\in[k]} L(h', \mathrm{e}_i^{\mathsf{basis}}) + \varepsilon.$$

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:

$$\mathsf{Rad}_n(\mathcal{D}', \mathcal{H}') := \frac{1}{n} \mathop{\mathbb{E}}_{\{x^j\}_{j=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}', \{\sigma^j\}_{j=1}^n \overset{\text{i.i.d.}}{\sim} \{\pm 1\}} \left[\max_{h'\in\mathcal{H}'} \sum_{j=1}^{n} \sigma^j h'(x^j)\right] \leqslant \widetilde{C}_n; \tag{151}$$

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

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

This construction is also straightforward. Set $N = 2^d$ and $\mathcal{X}' = \{0,1\}^N$. Let $\mathcal{D}'$ be the distribution

$$\mathbb{P}_{\mathcal{D}'}\{x\} = \prod_{n=1}^{N} \mathbb{P}_{\mathcal{D}'_n}\{x_n\}$$

where

$$\mathbb{P}_{\mathcal{D}'_{n^*}}\{x_{n^*}\} = \frac{1}{2}\mathbb{1}\{x_{n^*} = 1\} + \frac{1}{2}\mathbb{1}\{x_{n^*} = 0\} \qquad \text{for some } n^*; \tag{152}$$

$$\mathbb{P}_{\mathcal{D}'_n}\{x_n\} = \left(\frac{1}{2} + 2\varepsilon\right)\mathbb{1}\{x_{n^*} = 1\} + \left(\frac{1}{2} - 2\varepsilon\right)\mathbb{1}\{x_{n^*} = 0\} \qquad \text{for all } n \neq n^*. \tag{153}$$

We then choose $\mathcal{H}' = \{h^n\}_{n=1}^N$ 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 $\widetilde{\Omega}\left(d/\varepsilon^2\right)$ i.i.d. samples from $\mathcal{D}'$ to identify n^*. The proof is thus completed. □

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.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms.* Cambridge University Press.

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.

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*.