Foundations and Trends® in Communications and Information Theory

Online Universal Learning from Information-theoretic Perspective

Suggested Citation: Changlong Wu, Ananth Grama and Wojciech Szpankowski (2025), "Online Universal Learning from Information-theoretic Perspective", Foundations and Trends® in Communications and Information Theory: Vol. XX, No. XX, pp 1–225. DOI: XXX.

Changlong Wu
Purdue University
wuchangl@hawaii.edu

Ananth Grama
Purdue University
ayg@cs.purdue.edu

Wojciech Szpankowski
Purdue University
szpan@purdue.edu

now
the essence of knowledge
Boston — Delft

Contents

Online Universal Learning from Information-theoretic Perspective

Changlong Wu[1], Ananth Grama[2] and Wojciech Szpankowski[3]

[1] *Purdue University; wuchangl@hawaii.edu*
[2] *Purdue University; ayg@cs.purdue.edu*
[3] *Purdue University; szpan@purdue.edu*

ABSTRACT

In this paper, we explore online universal learning from an information-theoretic perspective. In information theory, universality is a fundamental concept that refers to the design of algorithms for tasks such as compression, estimation, or coding that perform well across broad classes of sources (e.g., memoryless or Markov sources). In online learning, universality involves the construction of a predictor that can accurately forecast future observations without prior knowledge of the underlying data-generating process. We first present preliminary results in online learning and introduce the concept of minimax regret, which quantifies the excess loss incurred by a predictor compared to the best fixed predictor (or constant comparator) within a given class of experts. Next, we analyze minimax regret under both logarithmic and bounded loss functions. We then examine the expected regret for i.i.d. data, as well as for more general stochastic processes. Our analysis also extends to online learning in the presence of noisy data. Finally, we propose several efficient learning algorithms.

ISSN 1567-2190; DOI XXXXXXXX

1

Introduction

One of the central themes in information theory is the concept of *universality*, which refers to the design of algorithms—such as compressors, estimators, or encoders—that perform well across a broad class of models, without prior knowledge of the specific model generating the data. A representative example of this idea is *universal source coding*, initiated by Davisson's seminal work [1], where the goal is to design *universal compressors* that achieve code lengths close to optimal for the unknown underlying source. The difference between the achieved length and the optimal is known as the *redundancy*. This fundamental concept has inspired a rich body of research [2]–[8].

A closely related concept has been widely studied in the machine learning (ML) community, known as *online learning*. In this setting, the goal is to design a *universal predictor* that can make accurate predictions for future observations without relying on any assumptions about the data-generation process. The objective, as before, is to achieve performance comparable to that of the best *expert* in a given class, with the performance gap measured by the notion of *regret*. Indeed, the concept of regret is closely related to redundancy, as discussed in the seminal book by Cesa-Bianchi and Lugosi [9].

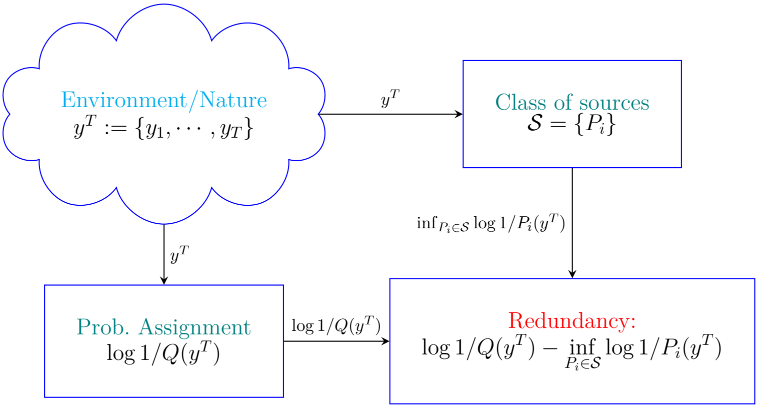

Environment/Nature
$y^T := \{y_1, \cdots, y_T\}$
y^T
Class of sources
$\mathcal{S} = \{P_i\}$
$\inf_{P_i \in \mathcal{S}} \log 1/P_i(y^T)$
y^T
Prob. Assignment
$\log 1/Q(y^T)$
$\log 1/Q(y^T)$
Redundancy:
$\log 1/Q(y^T) - \inf_{P_i \in \mathcal{S}} \log 1/P_i(y^T)$

Figure 1.1: A general universal source coding paradigm

Our current exposition is centered on the unifying principle of *universality in online learning*. We view the online learning problem through an information-theoretic lens, connecting classical ideas from universal coding with modern notions of regret minimization. In doing so, we build a unified theoretical framework that models a wide range of learning settings, including adversarial, stochastic, and hybrid scenarios within a common regret-based formalism.

1.1 Information Theory versus Online Learning

Perhaps the closest counterpart to online learning in information theory is the problem of *universal source coding* (see also Figure 1.1). In this setting, the goal is to find or learn the best (i.e., shortest) description of a sequence generated by a source from a class of sources $\mathcal{S}$. Since the seminal paper by Davisson [1], the quality of universal compression has been measured by various forms of *minimax redundancy*, defined as the excess of the actual compression length over the optimal one, either on average or in the worst case.

Formally, for a *given* source P (i.e., a probability distribution) and a (label) sequence $y^T := (y_1, \dots, y_T)$, the *pointwise redundancy* $R_T(P; y^T)$

and the *average redundancy* $\bar{R}_T(P)$ are defined as

$$R_T(P; y^T) = L(y^T) + \log \frac{1}{P(y^T)},$$
$$\bar{R}_T(P) = \mathbb{E}_{y^T \sim P}[R_T(P; y^T)],$$

where $L(y^T)$ is the coding length assigned to y^T by a compression algorithm, often taken to be

$$L(y^T) = -\log Q(y^T),$$

for some (universal) distribution Q that approximates P. Often, we do not know P exactly, but only that it belongs to a class of sources $\mathcal{S}$. Following [1], the average and worst-case *minimax redundancies* are defined as (cf. [2], [7], [8]):

$$\bar{r}_T(\mathcal{S}) = \min_Q \sup_{P \in \mathcal{S}} \mathbb{E}_{y^T \sim P} \left[-\log Q(y^T) + \log \sup_{P \in \mathcal{S}} P(y^T) \right],$$
$$r_T^*(\mathcal{S}) = \min_Q \max_{y^T} \left[-\log Q(y^T) + \log \sup_{P \in \mathcal{S}} P(y^T) \right].$$

The central question is how these different redundancies relate for various classes of sources $\mathcal{S}$. In [2], it is shown that if the maximum likelihood distribution belongs to the convex hull of $\mathcal{S}$, then

$$r_T^*(\mathcal{S}) - \bar{r}_T(\mathcal{S}) = O(c_T(\mathcal{S})),$$

where

$$c_T(\mathcal{S}) = \sum_{y^T} P(y^T) \log \left(\frac{\sup_{P \in \mathcal{S}} P(y^T)}{P(y^T)} \right).$$

It is also shown that

$$r_T^*(\mathcal{S}) = \bar{r}_T(\mathcal{S}) + O(1),$$

provided the maximum likelihood distribution lies in the convex hull of $\mathcal{S}$—for example, when $\mathcal{S}$ consists of finite-memory (e.g., Markov) sources. Furthermore, it is known (cf. [2]–[8]) that for a broad class of sources, the redundancy grows as

$$\frac{m-1}{2} \log T \quad \text{when the alphabet size } m \text{ is fixed},$$

and as

$$\frac{m-1}{2}\log(T/m) \quad \text{when } m = o(T),$$

(see also [3], [5], [10]).

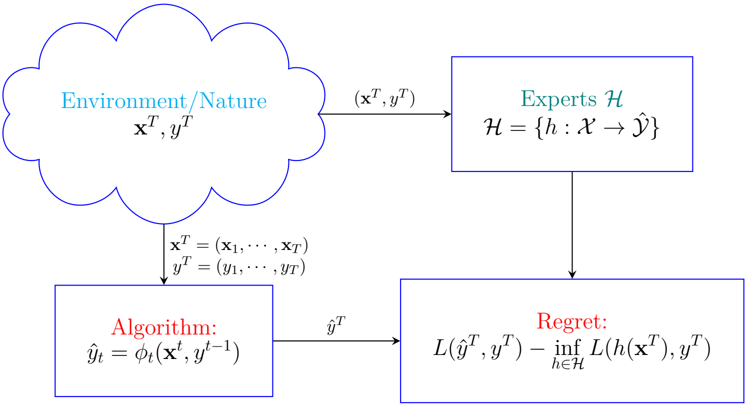

Environment/Nature
$\mathbf{x}^T, y^T$
$(\mathbf{x}^T, y^T)$
Experts $\mathcal{H}$
$\mathcal{H} = \{h : \mathcal{X} \to \hat{\mathcal{Y}}\}$
$\mathbf{x}^T = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$
$y^T = (y_1, \cdots, y_T)$
Algorithm:
$\hat{y}_t = \phi_t(\mathbf{x}^t, y^{t-1})$
$\hat{y}^T$
Regret:
$L(\hat{y}^T, y^T) - \inf_{h\in\mathcal{H}} L(h(\mathbf{x}^T), y^T)$

Figure 1.2: Illustration to online learning

Online Learning. We now focus on the *online learning* paradigm, illustrated in Figure 1.2, which can be formulated as a game between nature (or the environment) and a learner (or predictor). Broadly, the learner's objective is to use past observations to predict the next outcome in nature's labeling sequence. At each time step $t \in \mathbb{N}$, the learner receives a d-dimensional input $\mathbf{x}_t \in \mathcal{X}$. Based on the current and past observations, it outputs a prediction $\hat{y}_t = \Phi_t(y^{t-1}, \mathbf{x}^t)$, where Φ_t denotes the learner's strategy at round t—a function of the past labels $y^{t-1} = (y_1, \ldots, y_{t-1})$ and the current input sequence $\mathbf{x}^t = (\mathbf{x}_1, \ldots, \mathbf{x}_t)$. After the prediction is made, nature reveals the true label $y_t \in \mathcal{Y}$, and the learner incurs a loss determined by a predefined loss function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$, where $\hat{\mathcal{Y}}$ and $\mathcal{Y}$ denote the prediction and label spaces, respectively.

The central objective of online learning is to design algorithms that minimize the *regret*. Formally, given a learner Φ_t for each $t > 0$, and a sequence of data $(y_t, \mathbf{x}_t)_{t=1}^T$, the *pointwise regret* is defined as

$$R(\Phi^T, y^T, \mathcal{H} \mid \mathbf{x}^T) = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t),$$

where $\hat{y}_t = \Phi_t(y^{t-1}, \mathbf{x}^t)$ is the learner's prediction at time t, and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ is a class of experts (or hypotheses).

Similar to the average/worst-case redundancy dichotomy in universal source coding, there are various ways to define the notion of *minimax regret*, depending on how the data are generated.

Fixed Design: This point of view studies the minimal regret for the worst realization of the label sequence y^T when the feature sequence $\mathbf{x}^T$ is known in advance. Let Φ_t, for $t > 0$, denote the predictor's strategy. Then, the *fixed design minimax regret* is defined as

$$\mathsf{reg}_T^*(\mathcal{H} \mid \mathbf{x}^T) = \inf_{\Phi^T} \sup_{y^T} R(\Phi^T, y^T, \mathcal{H} \mid \mathbf{x}^T). \tag{1.1}$$

Furthermore, the fixed design *maximal* minimax regret is given by

$$\mathsf{reg}_T^*(\mathcal{H}) := \sup_{\mathbf{x}^T} \inf_{\Phi^T} \sup_{y^T} R(\Phi^T, y^T, \mathcal{H} \mid \mathbf{x}^T). \tag{1.2}$$

It is not hard to show that the *fixed design minimax regret* coincides with the (worst-case) *minimax redundancy*, when the loss function ℓ is the *logarithmic loss* (cf. Chapter 5).

Sequential Design: In this formulation, the optimization over regret is performed sequentially at each time step t, without prior knowledge of the entire feature sequence $\mathbf{x}^T$. The *sequential minimax regret* is defined as

$$\mathsf{reg}_T(\mathcal{H}) := \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \cdots \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} R(\hat{y}^T, y^T, \mathcal{H} \mid \mathbf{x}^T). \tag{1.3}$$

In fact, the sequential minimax regret can be equivalently expressed as (cf. Chapter 4)

$$\mathsf{reg}_T(\mathcal{H}) = \inf_{\Phi^T} \sup_{\mathbf{x}^T, y^T} R(\Phi^T, y^T, \mathcal{H} \mid \mathbf{x}^T).$$

Stochastic Design: Another important paradigm, analogous to the *average-case* minimax redundancy, arises when the data are generated by a *stochastic* (but unknown) source. In this setting, Nature selects an unknown *distribution* $\boldsymbol{\nu}^T$ over $\mathcal{X}^T$—that is, a random process—and samples a sequence $\mathbf{x}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T) \sim \boldsymbol{\nu}^T$. At each time step $t \leq T$, Nature reveals $\mathbf{x}_t$ to the predictor, who then makes a prediction $\hat{y}_t = \Phi_t(\mathbf{x}^t, y^{t-1})$, potentially using the history of inputs $\mathbf{x}^t = (\mathbf{x}_1, \ldots, \mathbf{x}_t)$ and past labels $y^{t-1} = (y_1, \ldots, y_{t-1})$. After the prediction is made, Nature reveals the true label y_t, and the predictor incurs a loss $\ell(\hat{y}_t, y_t)$. The *expected worst-case regret* is then defined as

$$\tilde{\mathsf{reg}}_T(\mathcal{H}, \mathsf{P}) = \inf_{\Phi^T} \sup_{\boldsymbol{\nu}^T \in \mathcal{P}} \mathbb{E}_{\mathbf{x}^T \sim \boldsymbol{\nu}^T} \left[\sup_{y^T} R(\Phi^T, y^T, \mathcal{H} \mid \mathbf{x}^T) \right], \tag{1.4}$$

where $\mathcal{H}$ is a class of predictors, P is a family of distributions over $\mathcal{X}^T$.

In fact, the *expected worst-case regret* subsumes both the fixed design and sequential minimax regret by selecting an appropriate distribution class P. This unifying viewpoint will serve as the central theme of the monograph. The main question we aim to address is how the complexity and structure of *both* the expert class $\mathcal{H}$ and the distribution class P influence the performance of online learning systems.

Outline. In Chapter 2, we summarize several useful tools that will be used throughout the monograph. Chapter 3 introduces preliminary results and foundational concepts in machine learning, such as Skolemization, Fano's inequality, and Le Cam's two-point method. In Chapter 4, we formally define both minimax regret and average regret, and present two important technical results that recur throughout the monograph: the Switch Lemma and the Shtarkov sum.

The next six chapters present novel contributions to online learning. Chapter 5 begins with the analysis of minimax regret under logarithmic loss. Chapter 6 extends the analysis to the case of Lipschitz loss functions. Chapter 7 is devoted to the expected regret when features are generated by an i.i.d. random process, while Chapter 8 studies the behavior of expected minimax regret under general, unknown data-generating processes. In Chapter 9, we shift focus to algorithm design, presenting efficient online learning algorithms that closely approximate the optimal minimax regret. Finally, Chapter 10 provides a detailed analysis of the minimax regret in the presence of label noise, where the learner only observes corrupted labels.

1.2 Bibliographical Notes

There is vast literature on machine learning, however, here we focus on performance of online learning, mostly from information-theoretic perspective. Universal coding can be traced back to seminal works of [1], [11], and [12]. Rissanen et al. [4] highlight the fact that that universal coding evolved into universal modeling and learning. The last 20 years have seen a resurgence of interest in redundancy rates for lossless coding: [2], [6]–[8], [13], and more [5], [14]–[16]. We shall discuss some of these references in a more detailed way in the next chapters. Some of them deal with finite alphabets while others such as [5], [14]–[16] consider unbounded alphabets or alphabets with some extra properties such as monotonicity condition.

Machine learning problems and techniques are discussed in many books, e.g., [17], [18]. We present more detailed bibliographical notes regarding machine learning the following chapters.

2

Useful Tools

In this chapter, we describe some useful results and tools that will be used throughout this exposition. We start with Bregman divergence and exp-convexity. We then summarize VC-dimension and the fat-shattering number. Skolemization is a useful tool that we also discuss in this chapter. Following the description of minimax expressions, we recall Fano inequality and discuss Le Cam two point methods. We conclude this section with some large deviation inequalities.

2.1 Bregman Divergence and Exp-concavity.

Let $\mathcal{D}(\tilde{\mathcal{Y}})$ be the set of probability distributions over some (finite) set $\tilde{\mathcal{Y}}$ of size M. A function $L : \mathcal{D}(\tilde{\mathcal{Y}}) \times \mathcal{D}(\tilde{\mathcal{Y}}) \to \mathbb{R}^{\geq 0}$ is referred to as a *divergence*. We say a divergence L is a *Bregman divergence* if there exists a strictly convex function $F : \mathcal{D}(\tilde{\mathcal{Y}}) \to \mathbb{R}$ such that for any $p, q \in \mathcal{D}(\tilde{\mathcal{Y}})$,

$$L(p,q) = F(p) - F(q) - (p-q)^{\mathsf{T}} \nabla F(q).$$

Note that both KL-divergence $\mathsf{KL}(p,q) = \sum_{\tilde{y} \in \tilde{\mathcal{Y}}} p[\tilde{y}] \log \frac{p[\tilde{y}]}{q[\tilde{y}]}$ and the L^2-divergence $L^2(p,q) = ||p-q||_2^2$ are Bregman divergences [9, Chapter 11.2].

We now present some properties of Bregman divergence.

Proposition 2.1. Let P be a random variable over $\mathcal{D}(\tilde{\mathcal{Y}})$ (i.e., a random variable with values in $\mathbb{R}^M$) and L be a Bregman divergence. Then for any $q_1, q_2 \in \mathcal{D}(\tilde{\mathcal{Y}})$

$$\mathbb{E}_{p\sim P}[L(p, q_1) - L(p, q_2)] = L(\mathbb{E}_{p\sim P}[p], q_1) - L(\mathbb{E}_{p\sim P}[p], q_2).$$

Proof. By definition of Bregman divergence, we have

$$\begin{aligned} L(p, q_1) - L(p, q_2) =& F(q_2) - F(q_1) - p^{\mathsf{T}}(\nabla F(q_1) - \nabla F(q_2)) \\ & + q_1^{\mathsf{T}} \nabla F(q_1) - q_2^{\mathsf{T}} \nabla F(q_2). \end{aligned}$$

Note that the above expression is a *linear* function w.r.t. p. Therefore, by taking expectation over $p \sim P$ and using linearity of expectation, one can verify the claimed identity holds. □

A function $\ell : \mathcal{D}(\tilde{\mathcal{Y}}) \times \tilde{\mathcal{Y}} \to \mathbb{R}^{\geq 0}$ is referred to as a *loss* function. For instance, the *log-loss* is defined as

$$\ell^{\mathsf{log}}(p, \tilde{y}) = \mathsf{KL}(e_{\tilde{y}}, p) = -\log p[\tilde{y}],$$

and the *Brier loss* is defined as

$$\ell^{\mathsf{B}}(p, \tilde{y}) = \|e_{\tilde{y}} - p\|_2^2,$$

where $e_{\tilde{y}}$ is the probability distribution assigning mass 1 to $\tilde{y}$. We say a loss ℓ is *α-exp-concave* if for any $\tilde{y} \in \tilde{\mathcal{Y}}$, the function $p \mapsto e^{-\alpha\ell(p,\tilde{y})}$ is concave with respect to p for some $\alpha \in \mathbb{R}^{\geq 0}$.

Proposition 2.2. The log-loss is 1-exp-concave and the Brier loss is 1/4-exp-concave.

Proof. The 1-exp-concavity of the log-loss can be verified directly. To prove the 1/4-exp-concavity of the Brier loss, we use the characterization from [19, Lemma 4.2], which states that a function f is α-exp-concave if and only if:

$$\alpha \nabla f(p) \nabla f(p)^{\mathsf{T}} \preceq \nabla^2 f(p).$$

Let $q \in \mathcal{D}(\tilde{\mathcal{Y}})$, and define $f(p) = \|p - q\|_2^2$. Then $\nabla f(p) = 2(p - q)$ and $\nabla^2 f(p) = 2I$, where I is the identity matrix. For any $u \in \mathbb{R}^M$, we have:

$$\frac{1}{4}\langle u, 2(p - q)\rangle^2 \leq \|u\|_2^2 \|p - q\|_2^2 \leq 2\|u\|_2^2 = 2u^{\mathsf{T}} I u,$$

where the first inequality follows from Cauchy–Schwarz, and the second inequality follows from:

$$\|p-q\|_2^2 = \sum_{\tilde{y}\in\tilde{\mathcal{Y}}} (p[\tilde{y}]-q[\tilde{y}])^2 \le \sum_{\tilde{y}\in\tilde{\mathcal{Y}}} \max\{p[\tilde{y}], q[\tilde{y}]\}^2 \le \sum_{\tilde{y}\in\tilde{\mathcal{Y}}} p[\tilde{y}]^2+q[\tilde{y}]^2 \le 2,$$

since $p, q \in \mathcal{D}(\tilde{\mathcal{Y}})$. This completes the proof. □

2.2 VC and Fat-Shattering Dimensions

For completeness, we recall the definition of the VC-dimension. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a class of functions mapping a set $\mathcal{X}$ into $\{0,1\}$. The *VC-dimension* of $\mathcal{H}$, denoted by $\mathsf{VC}(\mathcal{H})$, is defined as the largest integer $d \ge 1$ for which there exists a sequence $\mathbf{x}^d = (\mathbf{x}_1, \dots, \mathbf{x}_d) \in \mathcal{X}^d$ such that, for every $\mathbf{y}^d = (y_1, \dots, y_d) \in \{0,1\}^d$, there exists an $h \in \mathcal{H}$ satisfying $h(\mathbf{x}_t) = y_t$ for all $t \in [d]$. An equivalent and often more generalizable formulation is based on the notion of *shattering*. We say that $\mathcal{H}$ *shatters* $\mathbf{x}^d \in \mathcal{X}^d$ if, for every subset $I \subset [d]$, there exists $h \in \mathcal{H}$ such that $h(\mathbf{x}_t) = 1$ for all $t \in I$ and $h(\mathbf{x}_t) = 0$ for all $t \notin I$. The VC-dimension of $\mathcal{H}$ is then the largest d for which some $\mathbf{x}^d$ is shattered by $\mathcal{H}$.

The following combinatorial result characterizes the growth of finite VC-dimension classes:

Lemma 2.1 (Sauer's Lemma). Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a class of binary-valued functions with VC-dimension d. Then, for any sequence $\mathbf{x}^T = (\mathbf{x}_1, \dots, \mathbf{x}_T) \in \mathcal{X}^T$, we have:

$$|\{(h(\mathbf{x}_1), \dots, h(\mathbf{x}_T)) : h \in \mathcal{H}\}| \le \sum_{i=0}^{d} \binom{T}{i}.$$

Proof. We proceed by induction on T and d. For $T = 0$ or $d = 0$, the result holds trivially. Fix $\mathbf{x}^T \in \mathcal{X}^T$. For each $h \in \mathcal{H}$, let $h|_{T-1} = (h(\mathbf{x}_1), \dots, h(\mathbf{x}_{T-1}))$, and define:

$$A_0 = \{h|_{T-1} : h(\mathbf{x}_T) = 0,\, h \in \mathcal{H}\},$$
$$A_1 = \{h|_{T-1} : h(\mathbf{x}_T) = 1,\, h \in \mathcal{H}\}.$$

Denote $\mathcal{H}(\mathbf{x}^T) := \{(h(\mathbf{x}_1), \dots, h(\mathbf{x}_T)) : h \in \mathcal{H}\}$, we have:

$$|\mathcal{H}(\mathbf{x}^T)| = |A_0| + |A_1| = |A_0 \cup A_1| + |A_0 \cap A_1|.$$

Note that $A_0 \cup A_1$ is the set of all labelings of the first $T-1$ inputs realized by $\mathcal{H}$, so $\mathsf{VC}(A_0 \cup A_1) \le d$. Moreover, $A_0 \cap A_1$ corresponds to functions in $\mathcal{H}$ that agree on the first $T-1$ inputs and differ only on $\mathbf{x}_T$. If $A_0 \cap A_1$ were to shatter a set of size d among the first $T-1$ inputs, then $\mathcal{H}$ would shatter $d+1$ points (including $\mathbf{x}_T$), contradicting $\mathsf{VC}(\mathcal{H}) \le d$. Hence, $\mathsf{VC}(A_0 \cap A_1) \le d-1$. By the inductive hypothesis, we conclude

$$|\mathcal{H}(\mathbf{x}^T)| \le \sum_{i=0}^{d} \binom{T-1}{i} + \sum_{i=0}^{d-1} \binom{T-1}{i} = \sum_{i=0}^{d} \binom{T}{i},$$

where the final equality follows by the identity $\binom{T}{i} = \binom{T-1}{i} + \binom{T-1}{i-1}$. □

The definition of the VC-dimension can be generalized to real-valued functions as well. This leads to the notion of the *fat-shattering dimension*, which can be viewed as a scale-sensitive version of the VC-dimension. For any class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, we say that $\mathcal{H}$ *α-fat shatters* $\mathbf{x}^d \in \mathcal{X}^d$ if there exists a vector $\mathbf{s}^d \in [0,1]^d$ such that for every subset $I \subset [d]$, there exists a function $h \in \mathcal{H}$ satisfying, for all $t \in [d]$:

- If $t \in I$, then $h(\mathbf{x}_t) \ge s_t + \alpha$; and
- If $t \notin I$, then $h(\mathbf{x}_t) < s_t - \alpha$.

The *fat-shattering dimension* of $\mathcal{H}$ at scale α is defined as the largest integer $d := d(\alpha)$ such that there exists $\mathbf{x}^d \in \mathcal{X}^d$ that is α-fat shattered by $\mathcal{H}$ (see Figure 2.1). The special case of the fat-shattering dimension with $\alpha = 0$ is also referred to as the *pseudo-dimension.* Observe that the 0-fat-shattering dimension, with $s_i = 1/2$, reduces to the VC-dimension for binary-valued functions.

2.3 Minimax Inequalities and Skolemization

In this exposition, we often deal with expressions of the following form $\inf_x \sup_y f(x,y)$ for some function $f(x,y)$. On many occasions we need to know its relation to $\sup_y \inf_x f(x,y)$. So let us deal first with the so called *minimax inequality* which claims that

$$\inf_{x\in X} \sup_{y\in Y} f(x,y) \ge \sup_{y\in Y} \inf_{x\in X} f(x,y) \tag{2.1}$$

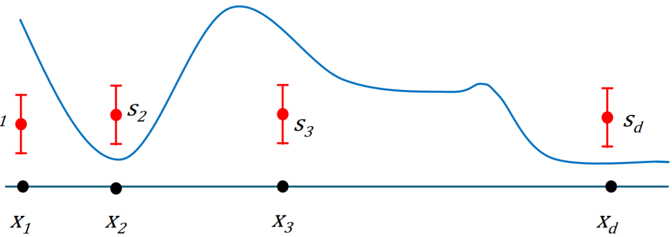


Figure 2.1: Illustration of the fat-shattering dimension

for any compact X and Y. Indeed, observe that for every $x \in X$ and $y \in Y$ we have

$$\sup_{y' \in Y} f(x, y') \geq f(x, y).$$

Taking inf over X we arrive at

$$\inf_{x \in X} \sup_{y' \in Y} f(x, y') \geq \inf_{x \in X} f(x, y)$$

which holds for all $y \in Y$, thus sup over Y can safely be appended giving

$$\inf_{x \in X} \sup_{y' \in Y} f(x, y') \geq \sup_{y \in Y} \inf_{x \in X} f(x, y)$$

proving (2.1).

We also notice that for any functions f and q

$$\sup \left(f(x) + g(x) \right) \leq \sup_x f(x) + \sup_x g(x). \tag{2.2}$$

Finally, we deal with an important tool known as skolemization which is usually used in logic but we adapt it to our context. Thoralf Albert Skolem was a Norwegian mathematician who worked in mathematical logic and showed how to eliminate existential quantifiers from logical formulas.

Lemma 2.2 (Skolemization). Let A, B be two sets, and $F : A \times B \to \mathbb{R}$ be an arbitrary function, then

$$\sup_{b \in B} \inf_{a \in A} F(a, b) = \inf_{g \in \mathcal{G}} \sup_{b \in B} F(g(b), b),$$

where $\mathcal{G} := A^B$ is the class of all functions from $B \to A$.

Proof. Define $\hat{g}(b) := \arg\inf_{a \in A} F(a,b)$ we have

$$\sup_b \inf_a F(a,b) = \sup_b F(\hat{g}(b), b) \geq \inf_g \sup_b F(g(b), b).$$

Moreover, let $g^* := \arg\min_{g \in \mathcal{G}}(\sup_b F(g(b), b))$ we have

$$\inf_g \sup_b F(g(b), b) = \sup_b F(g^*(b), b) \geq \sup_b \inf_a F(a,b).$$

Therefore, all inequalities become equality and the result follows. □

Minimax Theorem. We finish this section with a formulation of von Neumann minimax theorem which is used later to prove the so called *minimax switching trick* that allows us to express some minimax formulations via the average over some distributions.

We start with the *Minimax Theorem* that we present without a proof, see [9, Theorem 7.1].

Theorem 2.3 (Minimax Theorem). Let $f : A \times B \to \mathbb{R}$ be a bounded real-valued function, where both A and B are convex sets and A is compact. If $f(\cdot, b)$ is convex and continuous on A for any $b \in B$, and $f(a, \cdot)$ is concave on B for any $a \in A$, then

$$\inf_{a \in A} \sup_{b \in B} f(a,b) = \sup_{b \in B} \inf_{a \in A} f(a,b).$$

This theorem is stronger than von Neumann's minimax theorem, which specifically considers the case when f is a bi-linear function. It differs slightly from Sion's minimax theorem, which requires only semi-continuity and quasi-convexity (-concavity).

We complete this section with a useful trick, called the minimax switching trick that we shall use throughout this exposition when dealing with minimax regrets. Here, we write $\Delta(B)$ to denote a set of distributions over B.

Theorem 2.4 (Minimax Switching Trick). Let A be a convex set, B be a set such that $\Delta(B)$ is compact, and let $f : A \times B \to \mathbb{R}$ be a bounded function such that $f(\cdot, b)$ is convex for all $b \in B$. Then:

$$\inf_{a \in A} \sup_{b \in B} f(a,b) = \sup_{\mu \in \Delta(B)} \inf_{a \in A} \mathbb{E}_{b \sim \mu}[f(a,b)].$$

Proof. Note that:

$$\inf_{a\in A}\sup_{b\in B} f(a,b) = \inf_{a\in A}\sup_{\mu\in\Delta(B)} \mathbb{E}_{b\sim\mu}[f(a,b)].$$

Denote $F(a,\mu) = \mathbb{E}_{b\sim\mu}[f(a,b)]$. We have $F(\cdot,\mu)$ is convex over A, and $F(a,\cdot)$ is linear (therefore concave) over $\Delta(B)$. By the Minimax Theorem 2.3 we conclude:

$$\inf_{a\in A}\sup_{\mu\in\Delta(B)} F(a,\mu) = \sup_{\mu\in\Delta(B)}\inf_{a\in A} F(a,\mu)$$

and this completes the proof. □

2.4 Fano Inequality and Le Cam Two Point method

We often use terminology and methods of estimation in this paper. Fano inequality and Le Cam two-point methods are the most useful tools in this area that help us simplify our proofs.

Fano Inequality. Assume that X is a random variable with finite outcome $\mathcal{X}$. Furthermore, $\hat{X} = g(Y)$ is the predicted value of X after seeing Y, which is also defined over $\mathcal{X}$. We write $H(X)$ and $H(X|Y)$ for the entropy of X and conditional entropy of X under Y, respectively. Throughout we assume that g is a deterministic function. Then *Fano inequality* states

$$p_e := P(\hat{X} \neq X) \geq \frac{H(X|Y) - 1}{\log|\mathcal{X}|} \tag{2.3}$$

or more strongly

$$H(Ber(p_e)) + p_e \log(|\mathcal{X}| - 1) \geq H(X|Y) \tag{2.4}$$

where $Ber(p_e)$ is the Bernoulli random variable E with $P(E=1) = p_e$. To see this, let $E = 1$ if $\hat{X} \neq X$ and 0 otherwise, that is, $P(E=1) = p_e$. We now write $H(E,X|Y)$ in two different ways:

$$\begin{aligned} H(E,X|Y) &= H(X|Y) + H(E|X,Y) \\ H(E,X|Y) &= H(E|Y) + H(X|E,Y). \end{aligned}$$

Since $H(E|X,Y) = 0$, after comparing both sides of the above, and noting that for any X and Y, $H(X) \geq H(X|Y)$ we conclude that

$$H(X|Y) \leq H(Ber(p_e)) + H(X|E,Y)$$

and

$$H(X|E,Y) = H(X|E=0,Y)P(E=0) + H(X|E=1,Y)P(E=1).$$

But $H(X|E=0,Y) = 0$ and $H(X|E=1,Y) \leq \log(|\mathcal{X}|-1)$ since when $E=1$, the random variable X can take only $|\mathcal{X}|-1$ values. This completes the derivation of (2.4).

Using the fact $H(X|Y) = H(X) - I(X;Y)$ and assuming X is *uniform* over $\mathcal{X}$, we obtain the following useful form from (2.3)

$$p_e \geq 1 - \frac{I(X;Y)+1}{\log|\mathcal{X}|} \geq 1 - \frac{\inf_Q \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \mathsf{KL}(P_{Y|X=x}\|Q) + 1}{\log|\mathcal{X}|}. \tag{2.5}$$

Here, we used the following fact (for X uniform over $\mathcal{X}$):

$$\begin{aligned} I(X;Y) := \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \mathsf{KL}(P_{Y|X=x}\|P_Y) \\ &= \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \mathsf{KL}(P_{Y|X=x}\|Q) - \mathsf{KL}(P_Y\|Q) \\ &\leq \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \mathsf{KL}(P_{Y|X=x}\|Q) \end{aligned}$$

where Q is *any* distribution over $\mathcal{X}$ and KL is the Kullback–Leibler divergence.

Le Cam Two Point Method. We are again in the realm of estimation and let X be drawn from distribution P_θ parametrized by $\theta \in \Theta$. For a given loss function $\ell(\theta,\hat{\theta})$ with $\hat{\theta}$ being an estimation of θ, the average minimax risk can be defined as follows:

$$R_n(\Theta) = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_X \ell(\theta,\hat{\theta}(X)) \geq \inf_{\hat{\theta}} \sup_{\theta\in\{\theta_0,\theta_1\}} \mathbb{E}_X \ell(\theta,\hat{\theta}(X))$$

where $\theta_0, \theta_1 \in \Theta$ are any two sources from Θ.

We now state the Le Cam method in its generality, but we derive it only for the indicator function loss.

Lemma 2.5 (Le Cam Method)**.** Let the positive loss function $\ell(\theta_0, \theta_1)$ satisfy α-triangle inequality

$$\ell(\theta_0, \theta_1) \le \alpha[\ell(\theta_0, \theta) + \ell(\theta_1, \theta)]$$

for all $\theta, \theta_0, \theta_1 \in \Theta$ and some $\alpha > 0$. Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_X \ell(\theta, \hat{\theta}(X)) \ge \sup_{\theta_0, \theta_1 \in \Theta} \frac{\ell(\theta_0, \theta_1)}{2\alpha}(1 - \mathsf{TV}(P_{\theta_0}, P_{\theta_1}) \tag{2.6}$$

where TV is the total variation distance.

The proof can be found in [20], so here we only sketch the derivation for the case when the loss function is just the indicator function, that is, $\ell(\theta_0, \theta_1) = 1\{\theta_0 \neq \theta_1\}$. For simplicity, we denote θ_0, θ_1 as $0, 1$, respectively, and write P_0 and P_1 for P_{θ_0} and P_{θ_1}. Then, we observe

$$\inf_{\hat{\theta}} \sup_{\theta \in \{0,1\}} P_\theta(\hat{\theta} \neq \theta) \ge \frac{1}{2} \inf_{\hat{\theta}} \left(P_0(\hat{\theta} \neq 0) + P_1(\hat{\theta} \neq 1) \right).$$

But it is easy to observe that

$$\inf_{\hat{\theta}} \left(P_0(\hat{\theta} \neq 0) + P_1(\hat{\theta} \neq 1) \right) = 1 - \mathsf{TV}(P_0, P_1).$$

Thus combining the last two assertions we obtain the following expression for the Le Cam method.

Lemma 2.6. The following holds

$$\inf_{\hat{\theta}} \sup_{\theta \in \{0,1\}} P_\theta(\hat{\theta} \neq \theta) \ge \frac{1}{2} - \frac{1}{2}\mathsf{TV}(P_0, P_1) \ge \frac{1}{2} - \frac{1}{2\sqrt{2 \log e}}\sqrt{\mathsf{KL}(P_0 \| P_1)} \tag{2.7}$$

where the last inequality follows from Pinsker's inequality.

For any distributions P_0, P_1, the Pinsker's inequality states (see [20]):

$$\mathsf{TV}(P_0, P_1) \le \sqrt{\mathsf{KL}(P_0 \| P_1)/(2 \log e)}. \tag{2.8}$$

2.5 Martingale Concentration Inequalities

We present some standard concentration results for Martingales, which will be useful for deriving high probability guarantees. We refer to [21, Chapter 13.1] for the proofs.

Lemma 2.7 (Azuma's Inequality)**.** Let $X_1, \cdots, X_T$ be an arbitrary random process adapted to some filtration $\{\mathcal{F}_t\}_{t \le T}$ such that $|X_t| \le M$ for all $t \le T$. Let $Y_t = \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$ be the conditional expected random variable of X_t. Then for all $\delta > 0$, we have:

$$\Pr\left[\sum_{t=1}^{T} Y_t < \sum_{t=1}^{T} X_t + M\sqrt{(T/2)\log(1/\delta)}\right] \ge 1 - \delta,$$

and

$$\Pr\left[\sum_{t=1}^{T} Y_t > \sum_{t=1}^{T} X_t - M\sqrt{(T/2)\log(1/\delta)}\right] \ge 1 - \delta.$$

The following lemma provides a tighter concentration when $X_t \ge 0$, which can be viewed as a Martingale version of the multiplicative Chernoff bound.

Lemma 2.8 ([21, Theorem 13.5])**.** Let $X_1, \cdots, X_T$ be an arbitrary random process adapted to some filtration $\{\mathcal{F}_t\}_{t \le T}$ such that $0 \le X_t \le M$ for all $t \le T$. Let $Y_t = \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$ be the conditional expected random variable of X_t. Then for all $\delta > 0$ we have:

$$\Pr\left[\sum_{t=1}^{T} Y_t < 2\sum_{t=1}^{T} X_t + 2M\log(1/\delta)\right] \ge 1 - \delta,$$

and

$$\Pr\left[\sum_{t=1}^{T} Y_t > \frac{1}{2}\sum_{t=1}^{T} X_t - (M/2)\log(1/\delta)\right] \ge 1 - \delta.$$

Proof. Applying Zhang [21, Thm 13.5] with $\xi_t = X_t/M$ and $\lambda = 1$ in the theorem. □

Remark 2.1. It should be noted that the assumption $X_t \ge 0$ is *required* for Lemma 2.8 to hold. To see this, we group X^T as $X_1X_2, X_3X_4, \cdots$ such that X_{2t-1} is uniform over $\{-1, 1\}$ and $X_{2t} = -X_{2t-1}$ for all $t \in [T]$. It is easy to verify that $X_1 + \cdots + X_T = 0$ almost surely. But $Y_{2t-1} = 0$ and $Y_{2t} = -X_{2t-1}$, hence, we have $Y_1 + \cdots + Y_T$ is sum of $T/2$ independent uniform distributions over $\{-1, 1\}$. Therefore, by the central limit theorem $Y_1 + \cdots + Y_T \ge \Omega(\sqrt{T})$ with constant probability.

The following lemmas provide tighter bounds for general (not necessarily non-negative) processes by leveraging higher-order moment information.

Lemma 2.9 ([21, Theorem 13.2])**.** Let $X_1, \cdots, X_T$ be a random process adapted to some filtration $\{\mathcal{F}_t\}_{t \le T}$, and $\mathbb{E}_t$ be the conditional expectation on $\mathcal{F}_{t-1}$. Then, for any $\alpha, \delta > 0$ we have:

$$\Pr\left[-\sum_{t-1}^{T} \log \mathbb{E}_t[e^{-\alpha X_t}] \le \alpha \sum_{t=1}^{T} X_t + \log(1/\delta)\right] \ge 1 - \delta.$$

Lemma 2.10 (Freedman's inequality)**.** Let $X_1, \ldots, X_T$ be a martingale difference sequence adapted to a filtration $\{\mathcal{F}_t\}_{t \le T}$, that is, $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = 0$ for all $t \le T$. Assume $|X_t| \le M$ almost surely. Define the partial sums

$$S_t = \sum_{j=1}^{t} X_j \quad \text{for } t = 1, \ldots, T,$$

and the sum of conditional variances

$$V_T^2 = \sum_{t=1}^{T} \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}].$$

Then for any $u, v > 0$,

$$\Pr\left[\max_{1 \le t \le T} S_t > u \text{ and } V_T^2 \le v\right] \le \exp\left(-\frac{u^2}{2(v + Mu/3)}\right). \tag{2.9}$$

and therefore,

$$\Pr\left[\max_{1 \le t \le T} S_t > \sqrt{2vu} + (\sqrt{2}/3)Mu \text{ and } V_T^2 \le v\right] \le e^{-u}.$$

Proof. See [9, Lemma A.8] or the original proof of Freedman [22]. □

Note that Lemma 2.10 is also sometimes referred to as Bernstein's inequality for martingales (see also Lemma 2.12 below).

We finish this discussion by quoting Hoeffding's Lemma which is used in many places of this exposition.

Lemma 2.11 (Hoeffding's Lemma). Let X be a random variable with $a \leq X \leq b$. Then for any $s \in \mathbb{R}$, we have

$$\ln \mathbb{E}[e^{sX}] \leq s\mathbb{E}[X] + \frac{s^2(b-a)^2}{8}.$$

Proof. Note that $\ln \mathbb{E}[e^{sX}] = s\mathbb{E}[X] + \ln \mathbb{E}[e^{s(X-\mathbb{E}[X])}]$, so we only need to consider the case where $\mathbb{E}[X] = 0$. Observe that for all $a \leq x \leq b$, we have

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa},$$

by Jensen's inequality and the convexity of e^{sx} over x. Taking expectation over $x \sim X$ on both sides and using $\mathbb{E}[X] = 0$, the right-hand side can be expressed as a function of s. The lemma follows by Taylor expansion of this function up to the second order. □

Another version of Hoeffding's Lemma is the Bernstein's inequality that we formulate next.

Lemma 2.12. Let X be a random variable taking values in $[0, 1]$. Then for any $s \in \mathbb{R}$

$$\ln \mathbb{E}[e^{sX}] \leq (e^s - 1)\mathbb{E}[X].$$

We conclude this chapter with a classical inequality from probability theory, known as Khinchine's inequality [23].

Lemma 2.13 (Khinchine's inequality). Let $a_1, \ldots, a_T \in \mathbb{R}$, and let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_T)$ be a random vector uniformly distributed over $\{-1, +1\}^T$. Then

$$\frac{1}{\sqrt{2}}\sqrt{\sum_{t=1}^{T} a_t^2} \;\leq\; \mathbb{E}_{\boldsymbol{\epsilon}}\left|\sum_{t=1}^{T} a_t\epsilon_t\right| \;\leq\; \sqrt{\sum_{t=1}^{T} a_t^2}.$$

Proof. By Jensen's inequality, we have

$$\mathbb{E}\left|\sum_{t=1}^{T} a_t\epsilon_t\right| \leq \left(\mathbb{E}\left[\left(\sum_{t=1}^{T} a_t\epsilon_t\right)^2\right]\right)^{1/2} = \sqrt{\sum_{t=1}^{T} a_t^2},$$

since ϵ_t are independent Rademacher variables. The upper bound follows.

We give a short proof of the lower bound with a suboptimal constant $1/\sqrt{3}$ from [9, Lemma A.9]. Let $X := \sum_{t=1}^{T} a_t \epsilon_t$. By Hölder's inequality, for any bounded random variable X,

$$\mathbb{E}[X^2] = \mathbb{E}[|X|^{4/3}|X|^{2/3}] \leq (\mathbb{E}[X^4])^{1/3}(\mathbb{E}[|X|])^{2/3}.$$

Letting $X = \sum_{t=1}^{T} a_t \epsilon_t$, we obtain

$$\mathbb{E}[|X|] \geq \frac{\left(\sum_{t=1}^{T} a_t^2\right)^{3/2}}{\left(\sum_{t=1}^{T} a_t^4 + 3\sum_{i\neq j} a_i^2 a_j^2\right)^{1/2}} \overset{\text{(a)}}{\geq} \frac{1}{\sqrt{3}}\sqrt{\sum_{t=1}^{T} a_t^2},$$

where (a) follows from $\sum_t a_t^4 + 3\sum_{i\neq j} a_i^2 a_j^2 \leq 3\left(\sum_t a_t^2\right)^2$. □

Remark 2.2. The upper bound in Lemma 2.13 holds even when each coefficient a_t is a function of past signs, i.e., $a_t = \phi_t(\epsilon_1, \ldots, \epsilon_{t-1})$ for some measurable function ϕ_t. In this case,

$$\mathbb{E}_\epsilon \left|\sum_{t=1}^{T} a_t \epsilon_t\right| \leq \sqrt{\mathbb{E}_\epsilon\left[\sum_{t=1}^{T} a_t^2\right]}.$$

However, the lower bound does *not* generally hold in this adaptive setting; see [23, Section 6].

2.6 Bibliographical Notes

Most of the tools discussed in this chapter can be found in books such as [20], [17] and [24]. The minimax switching lemma is from [25], [26]. The large deviations results can be found in [21]. Khinchine inequality and its generalization can be found in [23].

3

Preliminary Results

In this Chapter we introduce basic concepts of online learning, and present some general simple predictors such as Consistent Predictor and Halving Predictor. Then, for the agnostic case, we discuss a popular algorithm known as the *Exponential Weighted Average Algorithm* (EWA) that we use often in our exposition.

3.1 Basic Concepts in Online learning

Let $\mathcal{X}$ denote the feature (or instance) space and $\mathcal{Y}$ the label space. A concept is a function $c : \mathcal{X} \to \mathcal{Y}$ representing the true labeling rule. The set of all possible such functions under consideration is called the concept class, denoted $\mathcal{C} \subset \mathcal{Y}^{\mathcal{X}}$. A hypothesis is a function $h : \mathcal{X} \to \mathcal{Y}$ that a learning algorithm may output, and the set of candidate hypotheses it can choose from is the hypothesis class, denoted $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$. We say the learning problem is realizable if $\mathcal{C} \subset \mathcal{H}$; that is, the true concept lies within the hypothesis class. Conversely, the problem is agnostic when no assumptions are made about the concept class, i.e., $\mathcal{C} := \mathcal{Y}^{\mathcal{X}}$. There can be intermediate cases between the realizable and agnostic settings. However, in this paper, we focus primarily on the realizable vs. agnostic dichotomy, and thus do not explicitly refer to the concept class in subsequent discussions. Instead, we concentrate on the properties and implications of different hypothesis classes. Finally, we sometimes allow the output of the learner Φ to lie outside the hypothesis class $\mathcal{H}$—a setting known as *improper learning*. If the output of Φ is always within $\mathcal{H}$, we refer to it as *proper learning*.

We now define our basic game of *online learning*.
For $t = 1, 2, \cdots, T$:

1. Nature/Environment presents an instance $\mathbf{x}_t \in \mathcal{X}$
2. Learner predicts a label $\hat{y}_t \in \mathcal{Y}$
3. Nature reveals true label $y_t \in \mathcal{Y}$
4. Learner suffers loss $\ell(\hat{y}_t, y_t)$, for certain function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

The objective is to find a learning rule Φ that minimizes the risk

$$\mathsf{risk}_T(\Phi) := \sum_{t=1}^{T} \ell(\hat{y}_t, y_t). \tag{3.1}$$

How can $\mathsf{risk}_T(\Phi)$ grow? We first prove an old result by Thomas M. Cover from 1965 showing that without any assumption the risk can grow as T. Indeed, take $\mathcal{Y} := \{0, 1\}$ and let $\ell(\hat{y}, y) := 1\{\hat{y} \neq y\}$. Then, $\mathsf{risk}_T(\Phi)$ reduces to the number of mistakes made by Φ to predict the y_t's. Let Φ be any learning rule. Consider the following simple strategy for Nature:

- At each time step t, after the learner makes the prediction $\hat{y}_t$, Nature *adversarially* chooses $y_t \in \mathcal{Y}$ such that $y_t \neq \hat{y}_t$.

The number of mistakes made by the learner equals T, i.e., the learner errs at every step.

Corollary 3.1 (Cover, 1965)**.** Any learning rule Φ cannot achieve a mistake bound better than T.

This sounds uninteresting, so what is the catch? Observe that we did *not* use any prior knowledge about the learning target.

3.2 Incorporating Prior Knowledge: Realizable Case

Let $\mathcal{H} := \{h_1, \cdots, h_K\} \subset \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, and assume that Nature's strategy is *realizable*, i.e., there exists an $h \in \mathcal{H}$ such that for all $t \leq T$ we have $h(\mathbf{x}_t) = y_t$.

We now introduce the first, not very efficient, learning algorithm called the *consistent predictor*.

1. At each time step t, find any consistent hypothesis $\hat{h}_t \in \mathcal{H}$ (which must exist due to realizability) such that:
$$\sum_{i=1}^{t-1} 1\{\hat{h}_t(\mathbf{x}_i) \neq y_i\} = 0.$$
2. Make the prediction: $\hat{y}_t = \hat{h}_t(\mathbf{x}_t)$.

The question is how many mistakes will we make? Notice that each mistake will eliminate at least one function from $\mathcal{H}$, so the total number of mistakes is upper bounded by $|\mathcal{H}|$. Actually, we can prove that it is also a lower bound.

Lower Bound on Consistent Predictor. Consider the following hypothesis class:

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\cdots$
h_0	0	0	0	0	$\cdots$
h_1	1	0	0	0	$\cdots$
h_2	0	1	0	0	$\cdots$
h_3	0	0	1	0	$\cdots$
h_4	0	0	0	1	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Assume that h_0 is the ground truth predictor. At each time step t, both h_t and h_0 are consistent with the prior data. Consider now a consistent predictor that always selects h_t to make predictions at step t, which will incur at least $|\mathcal{H}|$ mistakes. Thus, in the worst-case scenario, a consistent predictor cannot achieve a mistake bound better than $|\mathcal{H}|$.

Can we do better? It turns out we can use a smarter way to eliminate inconsistent hypotheses via the so-called *halving predictor*.

1. Maintain a running hypothesis class $\mathcal{H}^{(t)}$ with $\mathcal{H}^{(0)} := \mathcal{H}$
2. At each time step t, after receiving $\mathbf{x}_t$, we define for $y \in \{0, 1\}$
$$\mathcal{H}_y^{(t)} = \{h \in \mathcal{H}^{(t-1)} : h(\mathbf{x}_t) = y\}.$$
3. Predict $\hat{y}_t = \arg\max_{y \in \{0,1\}} \{|\mathcal{H}_0^{(t)}|, |\mathcal{H}_1^{(t)}|\}$
4. Let y_t be true label, update $\mathcal{H}^{(t)} = \mathcal{H}_{y_t}^{(t)}$

How many mistakes do we make? Every time a mistake happens (i.e., $\hat{y}_t \neq y_t$), we have $|\mathcal{H}^{(t)}| \leq |\mathcal{H}^{(t-1)}|/2$. Thus, total number of mistakes is upper bounded by $\log |\mathcal{H}|$ (an exponential improvement over the $|\mathcal{H}|$ lower bound!).

3.3 Incorporating Prior Knowledge: Agnostic Case

Both the consistent and halving predictors rely heavily on the assumption that the data is realizable, i.e., there exists $h \in \mathcal{H}$ that is consistent with all the data. A single mismatch between the true data and the best hypothesis in $\mathcal{H}$ will cause both predictors to catastrophically fail. So the question is whether we can develop an algorithm that is robust to potential noise? To do so we need to recognize that an absolute mistake bound is not very informative and we must instead consider guarantees relative to the minimal mistakes achievable by a hypothesis in $\mathcal{H}$. This leads to the definition of *regret*, one of the most important notations on machine learning that we shall use throughout this paper.

Let

$$\widehat{M}_T := \sum_{t=1}^{T} 1\{\hat{y}_t \neq y_t\} \tag{3.2}$$

$$M_T^* := \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} 1\{h(\mathbf{x}_t) \neq y_t\} \tag{3.3}$$

where $\widehat{M}_T$ measures the number of mistakes made by a predictor Φ, and M_T^* measures the minimal number of mistakes achievable by any hypothesis in $\mathcal{H}$. Then we define the α-agnostic *regret* for $\alpha > 0$ as

$$\mathsf{reg}_T^{(\alpha)}(\Phi, \mathcal{H}) := \widehat{M}_T - \alpha M_T^*. \tag{3.4}$$

Let $\mathcal{H} := \{h_1, \cdots, h_K\}$ be any finite hypothesis class of size K. We now introduce Algorithm 3.1 known as the *exponential weights predictor*. This algorithm is one of the most commonly used in the field.

Algorithm 3.1 Exponential Weighted Average (EWA) Algorithm

```
Maintain a weight vector $\mathbf{w}^{(t)} \in \mathbb{R}^K$, initially $\mathbf{w}^{(0)} = (1, \cdots, 1)$.
for $t = 1, \cdots, T$ do
    After receiving $\mathbf{x}_t$, compute the weighted average:
        $$\hat{p}_t = \sum_{k=1}^{K} \frac{w_k^{(t-1)}}{\sum_{k=1}^{K} w_k^{(t-1)}} h_k(\mathbf{x}_t).$$
    Predict $\hat{y}_t = 1\{\hat{p}_t \geq \frac{1}{2}\}$, i.e., we predict the weighted-majority.
    After receiving $y_t$, update the weight $w_k^{(t)}$ as follows:
        $$w_k^{(t)} = \begin{cases} w_k^{(t-1)}, & \text{if } h_k(\mathbf{x}_t) = y_t, \\ (1-\eta)\, w_k^{(t-1)}, & \text{if } h_k(\mathbf{x}_t) \neq y_t, \end{cases}$$
    where $\eta \leq 1$ is a tunable parameter.
end for
```

We now prove the following result regarding the regret.

Theorem 3.2. Regardless of how Nature generates the data, the (deterministic) EWA algorithm Φ enjoys the following mistake bound:

$$\widehat{M}_T \leq 2(1+\eta) M_T^* + \frac{2\ln(|\mathcal{H}|)}{\eta} \tag{3.5}$$

hence the 2-agnostic regret is

$$\mathsf{reg}_T^{(2)}(\Phi, \mathcal{H}) \leq O(\sqrt{M_T^* \log |\mathcal{H}|}). \tag{3.6}$$

Proof. For any time step t, we define the *potential*:

$$W^{(t)} = \sum_{k=1}^{K} w_k^{(t)}, \quad \text{with } W^{(0)} = K.$$

Let

$$I_t := \{k \in [K] : h_k(\mathbf{x}_t) = y_t\}, \quad J_t := [K] \setminus I_t.$$

If the prediction at time t is incorrect, i.e., $\hat{y}_t \neq y_t$, then by the *weighted majority* prediction:

$$\sum_{k \in J_t} w_k^{(t-1)} \geq \sum_{k \in I_t} w_k^{(t-1)}.$$

Hence, for each time step t where a *mistake* occurs, the potential satisfies:

$$\begin{aligned} W^{(t)} &= \sum_{k=1}^{K} w_k^{(t)} \\ &= (1-\eta) \underbrace{\sum_{k \in J_t} w_k^{(t-1)}}_{A} + \underbrace{\sum_{k \in I_t} w_k^{(t-1)}}_{B} \\ &\overset{(a)}{\leq} \left(\frac{1-\eta}{2} + \frac{1}{2}\right) W^{(t-1)} = \left(1 - \frac{\eta}{2}\right) W^{(t-1)}, \end{aligned}$$

where step (a) uses the facts that $A+B = W^{(t-1)}$, $A \geq B$, and $1-\eta \leq 1$.

Applying this inequality over all time steps $t = 1, \ldots, T$, we obtain:

$$(1-\eta)^{M_T^*} \leq W^{(T)} \leq W^{(0)} \left(1 - \frac{\eta}{2}\right)^{\widehat{M}_T} \leq K \cdot \left(1 - \frac{\eta}{2}\right)^{\widehat{M}_T},$$

where M_T^* is the number of mistakes by the best expert, and $\widehat{M}_T$ is the number of mistakes made by the algorithm.

Taking natural logarithms of both sides and using:

$$\ln(1-\eta) \geq -\eta - \eta^2 \quad \text{and} \quad \ln\left(1 - \frac{\eta}{2}\right) \leq -\frac{\eta}{2} \quad \text{for } \eta < \frac{1}{2},$$

we complete the proof of (3.5). To obtain (3.6) we put optimal $\eta = \sqrt{\ln \mathcal{H}/M_T^*}$ to complete the proof. □

Thus far, we have obtained a regret bound for $\alpha = 2$, and the natural question is whether it can be improved to $\alpha = 1$. It turns out that for *deterministic* predictors, this is not possible. To see why, consider the hypothesis class $\mathcal{H} = \{h_0, h_1\}$, where each $h_b(\mathbf{x}) = b$ for all $\mathbf{x} \in \mathcal{X}$ and $b \in \{0, 1\}$. Following Cover's construction, any deterministic predictor can be forced to make T mistakes, while the best hypothesis in hindsight incurs at most $T/2$ mistakes. Therefore, no deterministic predictor can achieve sublinear regret when $\alpha = 1$.

Nevertheless, we can show that sublinear regret with $\alpha = 1$ is achievable by using the *randomized* Exponentially Weighted Average (EWA) algorithm, as described below.

Algorithm 3.2 Randomized Exponentially Weighted Average (REWA)

```
Initialize: Set weights $\mathbf{w}^{(0)} = (1, \dots, 1) \in \mathbb{R}^K$.
for each round $t = 1, 2, \dots, T$ do
    Define distribution $\tilde{p}_t[k] = \frac{w_k^{(t-1)}}{\sum_{j=1}^K w_j^{(t-1)}}$ for all $k \in [K]$.
    Sample $\hat{k}_t \sim \tilde{p}_t$ and predict $\hat{y}_t := h_{\hat{k}_t}(\mathbf{x}_t)$.
    Update weights:
        $w_k^{(t)} = w_k^{(t-1)} \cdot \exp\left(-\eta \cdot 1\{h_k(\mathbf{x}_t) \neq y_t\}\right), \quad \forall k \in [K],$
      where $\eta < 1$ is a tunable learning rate.
end for
```

With this predictor in mind, we can prove the following stronger results on $\alpha = 1$ regret.

Theorem 3.3. Regardless of how Nature generates the data, as long as the selection is independent to the internal randomness of the predictor, we have

$$\mathbb{E}_{\hat{y}^T}\left[\sum_{t=1}^T 1\{\hat{y}_t \neq y_t\}\right] \leq M_T^* + \frac{\ln(|\mathcal{H}|)}{\eta} + \frac{\eta T}{8}$$

leading to

$$\mathsf{reg}_T^{(1)}(\Phi, \mathcal{H}) \leq O(\sqrt{T \log |\mathcal{H}|}). \tag{3.7}$$

Proof. We again define the potential $W^{(t)} = \sum_{k=1}^{K} w_k^{(t)}$. Observe that:

$$\begin{aligned}\ln \frac{W^{(t)}}{W^{(t-1)}} &= \ln \sum_{k=1}^{K} \frac{w_k^{(t-1)}}{W^{(t-1)}} e^{-\eta 1\{h_k(\mathbf{x}_t) \neq y_t\}} \\ &\overset{(a)}{\leq} -\eta \sum_{k=1}^{K} \frac{w_k^{(t-1)}}{W^{(t-1)}} 1\{h_k(\mathbf{x}_t) \neq y_t\} + \frac{\eta^2}{8} \\ &\overset{(b)}{=} -\eta \mathbb{E}_{\hat{y}_t}[1\{\hat{y}_t \neq y_t\}] + \frac{\eta^2}{8},\end{aligned}$$

where (a) follows by Hoeffding's Lemma [9, Lemma A.1] (see also Lemma 2.11) and (b) follows from the definition of $\hat{y}_t$. Summing from $t = 1$ to T, we get:

$$-\eta M_T^* \leq \ln W^{(T)} \leq -\eta \mathbb{E}_{\hat{y}^T}\left[\sum_{t=1}^{T} 1\{\hat{y}_t \neq y_t\}\right] + \frac{\eta^2 T}{8} + \ln |\mathcal{H}|.$$

The regret bound follows by rearranging the inequality. The formula for the regret follows from it by selecting optimal η. □

3.3.1 General Losses

We complete this section with the EWA algorithm for general losses. Let $\mathcal{Y} = [0, 1]$ and $\mathcal{H} \subset [0, 1]^{\mathcal{X}}$ be a finite hypothesis class of size K. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be a loss function that is convex in its first argument. Then the generalized EWA predictor works as follows:

Using the same tools as in the proof of Theorem 3.3, we can establish the following result:

Theorem 3.4. For any data $\mathbf{x}^T, y^T$ and any bounded convex loss $\ell(\hat{y}, y)$, Algorithm 3.3 enjoys the following risk bound

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) + \frac{\ln(|\mathcal{H}|)}{\eta} + \frac{\eta T}{8}.$$

Taking $\eta = \sqrt{\frac{8 \ln(|\mathcal{H}|)}{T}}$, we have

$$\mathsf{reg}_T^{(1)}(\Phi, \mathcal{H}) \leq O(\sqrt{T \log |\mathcal{H}|}). \tag{3.8}$$

Algorithm 3.3 Exponential Weighted algorithm for General Loss

```
Initialize: Set weights $\mathbf{w}^{(0)} = (1, \ldots, 1) \in \mathbb{R}^K$.
for each round $t = 1, 2, \ldots, T$ do
    Define distribution $\tilde{p}_t[k] = \frac{w_k^{(t-1)}}{\sum_{j=1}^K w_j^{(t-1)}}, \quad \forall k \in [K]$.
    Predict $\hat{y}_t := \sum_{k=1}^K \tilde{p}_t[k] \cdot h_k(\mathbf{x}_t)$.
    Update weights:
        $w_k^{(t)} = w_k^{(t-1)} \cdot \exp\left(-\eta \cdot \ell(h_k(\mathbf{x}_t), y_t)\right), \quad \forall k \in [K],$
      where $\eta < 1$ is a tunable learning rate.
end for
```

Proof. Define the potential $W^{(t)} = \sum_{k=1}^K w_k^{(t)}$. Observe that:

$$\begin{aligned}
\ln \frac{W^{(t)}}{W^{(t-1)}} &= \ln \sum_{k=1}^K \frac{w_k^{(t-1)}}{W^{(t-1)}} e^{-\eta\ell(h_k(\mathbf{x}_t), y_y)} \\
&\overset{(a)}{\leq} -\eta \sum_{k=1}^K \frac{w_k^{(t-1)}}{W^{(t-1)}} \ell(h_k(\mathbf{x}_t), y_t) + \frac{\eta^2}{8} \\
&\overset{(b)}{=} -\eta \cdot \ell\left(\sum_{k=1}^K \frac{w_k^{(t-1)} \cdot h_k(\mathbf{x}_t)}{W^{(t-1)}}, y_t\right) + \frac{\eta^2}{8} \\
&\overset{(c)}{=} -\eta \cdot \ell(\hat{y}_t, y_t) + \frac{\eta^2}{8}
\end{aligned}$$

where (a) follows by Hoeffding's Lemma [9, Lemma A.1] and $\ell(\hat{y}, y) \in [0, 1]$; (b) follows by the convexity of ℓ; and (c) follows by the definition of $\hat{y}_t$. The theorem now follows the same way as Theorem 3.3 □

We now complete this section with a stronger regret bound for the exp-concave losses (e.g, log-los and Brier loss) discussed in Chapter 2.1.

Theorem 3.5. For any data $\mathbf{x}^T, y^T$ and any α-exp-concave loss $\ell(\hat{y}, y)$ (cf. Chapter 2.1), Algorithm 3.3 with $\eta := \alpha$ enjoys the following regret bound:

$$\mathsf{reg}_T^{(1)}(\Phi, \mathcal{H}) \leq \frac{\log |\mathcal{H}|}{\alpha}.$$

Proof. Denote $W^{(t)} = \sum_{k=1}^{K} w_k^{(t)}$. We have (taking $\eta := \alpha$):

$$\begin{aligned}
\ln \frac{W^{(t)}}{W^{(t-1)}} &= \ln \sum_{k=1}^{K} \frac{w_k^{(t-1)}}{W^{(t-1)}} e^{-\alpha\ell(h_k(\mathbf{x}_t),y_t)} \\
&\le \ln e^{-\alpha\ell\left(\sum_{k=1}^{K} \frac{w_k^{(t-1)}}{W^{(t-1)}} h_k(\mathbf{x}_t),y_t\right)} \\
&= \ln e^{-\alpha\ell(\hat{y}_t,y_t)} = -\alpha\ell(\hat{y}_t, y_t),
\end{aligned}$$

where the inequality follows by Jensen and α-Exp-concavity of ℓ. The theorem now follows the same way as Theorem 3.3. □

3.4 Bibliographical Notes

The materials in this section are mostly standard, and can be found in textbooks such as [9] and [17].

4

Minimax Regret

In this section, we collect definitions and general results regarding the *minimax regret*, one of the most important measures in online learning problems. We first discuss the worst-case minimax regret, and then briefly address the expected minimax regret. For logarithmic loss (for a detailed analysis see Chapter 5), we introduce the so-called *Shtarkov sum*, which will be used throughout to analyze the minimax regret. For Lipschitz loss (for a detailed analysis see Chapter 6), we present the so-called *Bayesian representation* via the Minimax Switching Trick, which will also be used in the analysis of regret.

4.1 Definitions and General Results

Let $\mathcal{X}$ be the instance space, $\mathcal{Y}$ the label space, and $\hat{\mathcal{Y}}$ a (convex) outcome space of predictors. We define the hypothesis class as $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$, and the (possibly improper) *learning rule* as

$$\Phi : (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{X} \to \hat{\mathcal{Y}}.$$

We consider the following general online learning game:
For $t = 1, 2, \ldots, T$:

1. Nature (or the environment) presents an instance $\mathbf{x}_t \in \mathcal{X}$.

2. The learner predicts a label $\hat{y}_t \in \hat{\mathcal{Y}}$ via $\hat{y}_t := \Phi(\mathbf{x}^t, y^{t-1})$.

3. Nature reveals the true label $y_t \in \mathcal{Y}$.

4. The learner suffers a *loss* $\ell(\hat{y}_t, y_t)$, for some loss function:

$$\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}.$$

The learner's goal is to minimize, against the *worst-case* Nature, the *regret*, which is defined as the excess of the cumulative loss over the best cumulative loss achieved by any expert in the hypothesis class $\mathcal{H}$. For any given $\mathbf{x}^T \in \mathcal{X}^T$ and $y^T \in \mathcal{Y}^T$, the *point-wise regret* is defined as

$$R_T(\mathcal{H}, \Phi, \mathbf{x}^T, y^T) := \sum_{t=1}^{T} \ell(\Phi(\mathbf{x}^t, y^{t-1}), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t). \quad (4.1)$$

The *worst-case regret* for a given learning rule Φ is defined as

$$\mathsf{reg}_T(\mathcal{H}, \Phi) := \sup_{\mathbf{x}^T, y^T} R_T(\mathcal{H}, \Phi, \mathbf{x}^T, y^T).$$

The *minimax regret* for the hypothesis class $\mathcal{H}$ is defined as

$$\mathsf{reg}_T(\mathcal{H}) := \inf_{\Phi} \mathsf{reg}_T(\mathcal{H}, \Phi) = \inf_{\Phi} \sup_{\mathbf{x}^T, y^T} R_T(\mathcal{H}, \Phi, \mathbf{x}^T, y^T). \quad (4.2)$$

The next result shows that the above definition is equivalent to the *sequential regret* briefly discussed in Chapter 1.2.

Theorem 4.1. The minimax regret satisfies

$$\mathsf{reg}_T(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \cdots \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right] \quad (4.3)$$

for any loss function ℓ and class $\mathcal{H}$.

Proof. We prove only the case for $T = 1$ to demonstrate the idea. Define the function:

$$F(a, b) := \sup_{y_1} \left[\ell(a, y_1) - \inf_{h \in \mathcal{H}} \ell(h(b), y_1) \right].$$

Note that:

$$\mathsf{reg}_1(\mathcal{H}) := \inf_{\Phi} \sup_{\mathbf{x}_1} F(\Phi(\mathbf{x}_1), \mathbf{x}_1).$$

By Skolemization Lemma 2.2, we have:

$$\inf_{\Phi} \sup_{\mathbf{x}_1} F(\Phi(\mathbf{x}_1), \mathbf{x}_1) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} F(\hat{y}_1, \mathbf{x}_1).$$

Plugging back the expression of $F(a, b)$, we find

$$\mathsf{reg}_1(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \left[\ell(\hat{y}_1, y_1) - \inf_{h \in \mathcal{H}} \ell(h(\mathbf{x}_1), y_1) \right].$$

Iterating this argument we prove Theorem 4.1. □

Note that in the definition of *minimax regret*, the predictor must compete with the worst-case data sequences $\mathbf{x}^T, y^T$, which can sometimes be overly pessimistic in real-world scenarios. Depending on how the features are selected, one may also consider the following relaxed notions of minimax regret:

Expected Worst-Case Minimax Regret. In this case, Nature selects some *distribution* $\boldsymbol{\nu}^T$ over $\mathcal{X}^T$ (i.e., a random process) and samples $\mathbf{x}^T \sim \boldsymbol{\nu}^T$ where $\mathbf{x}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$. At each time step $t \leq T$, Nature reveals $\mathbf{x}_t$ to the predictor, who makes a prediction $\hat{y}_t = \Phi(\mathbf{x}^t, y^{t-1})$ potentially using the history $\mathbf{x}^t = (\mathbf{x}_1, \cdots, \mathbf{x}_t)$ and $y^{t-1} = (y_1, \cdots, y_{t-1})$ that are observed thus far. Nature then reveals the true label y_t after the prediction, and the predictor incurs a loss $\ell(\hat{y}_t, y_t)$ for some predefined *convex* loss function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to [0, \infty)$. The *expected worst-case* minimax regret is then defined as:

$$\overline{\mathsf{reg}}_T(\mathcal{H}, \mathsf{P}) = \inf_{\Phi} \sup_{\boldsymbol{\nu}^T \in \mathsf{P}} \mathbb{E}_{\mathbf{x}^T \sim \boldsymbol{\nu}^T} \left[\sup_{y^T} \sum_{t=1}^{T} \ell(\Phi(\mathbf{x}^t, y^{t-1}), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right], \tag{4.4}$$

where $\mathcal{H}$ is a class of functions $\mathcal{X} \to \hat{\mathcal{Y}}$, P is a general class of random processes over $\mathcal{X}^T$, and the predictor Φ runs over all possible (deterministic) prediction rules.

Fixed Design Minimax Regret. In some scenarios the so called *fixed design* or *transductive* regret are of interest. In this case, the feature $\mathbf{x}^T$

is known in advance. More precisely, for any given $\mathbf{x}^T$, we define the fixed design minimax regret as:

$$\mathsf{reg}^*_T(\mathcal{H} \mid \mathbf{x}^T) = \inf_{\Phi} \sup_{y^T} \left[\sum_{t=1}^{T} \ell(\Phi(\mathbf{x}^T, y^{t-1}), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right]. \tag{4.5}$$

Recall the (sequential) minimax regret is defined as:

$$\mathsf{reg}_T(\mathcal{H}) = \inf_{\Phi} \sup_{\mathbf{x}^T, y^T} \left[\sum_{t=1}^{T} \ell(\Phi(\mathbf{x}^t, y^{t-1}), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right].$$

It is easy to observe that:

$$\mathsf{reg}^*_T(\mathcal{H}) := \sup_{\mathbf{x}^T} \mathsf{reg}^*_T(\mathcal{H} \mid \mathbf{x}^T) \leq \mathsf{reg}_T(\mathcal{H}). \tag{4.6}$$

Thus $\mathsf{reg}^*_T(\mathcal{H})$ can be viewed as a universal lower bound for regret.

4.2 Minimax Regret for Bounded Lipchitz Loss: Switch Lemma

We now present a general theorem regarding the worst-case minimax regret for bounded Lipschitz loss, which is due to [27]. Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0,1]$ and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$. The minimax regret for $\mathcal{H}$ can be expressed as in Theorem 4.1

$$\mathsf{reg}_T(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \cdots \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right].$$

However, this iterative minimax operator may be hard to compute. It turns out that we can replace it by a simpler operator as shown below.

Theorem 4.2. Assume the loss ℓ is bounded and $\ell(\cdot, y)$ is convex and continuous, $\hat{\mathcal{Y}}$ is convex and $\Delta(\mathcal{X} \times \mathcal{Y})$ is compact. Then the minimax regret $\mathsf{reg}_T(\mathcal{H})$ *equals*

$$\sup_{\boldsymbol{\nu}^T \in \Delta(\mathcal{X} \times Y)^T} \mathbb{E}_{(\mathbf{x}^T, y^T) \sim \boldsymbol{\nu}^T} \left[\sum_{t=1}^{T} \inf_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_t[\ell(\hat{y}_t, y_t)] - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right], \tag{4.7}$$

wher $\Delta(\mathcal{X} \times Y)^T$ is the set of all distributions over $(\mathcal{X} \times Y)^T$ and $\mathbb{E}_t$ denotes the conditional expectation of $\boldsymbol{\nu}^T$ conditioning on $\mathbf{x}^t, y^{t-1}$.

Proof. Observe that the iterated minimax formulation can be written as:

$$\sup_{\mathbf{z}_0}\inf_{\hat{y}_1}\sup_{\mathbf{z}_1}\cdots\inf_{\hat{y}_T}\sup_{\mathbf{z}_T}\left[\sum_{t=1}^{T}\ell(\hat{y}_t,y_t)-\inf_{h\in\mathcal{H}}\sum_{t=1}^{T}\ell(h(\mathbf{x}_t),y_t)\right],$$

where $\mathbf{z}_0=\mathbf{x}_1$, $\mathbf{z}_t=(y_t,\mathbf{x}_{t+1})$ for $t<T$ and $\mathbf{z}_T=y_T$. Consider the last layer:

$$\begin{aligned}&\inf_{\hat{y}_T}\sup_{\mathbf{z}_T}\sum_{t=1}^{T}\ell(\hat{y}_t,y_t)-\underbrace{\inf_{h\in\mathcal{H}}\sum_{t=1}^{T}\ell(h(\mathbf{x}_t),y_t)}_{F(\mathbf{z}^T)}\\&=\sum_{t=1}^{T-1}\ell(\hat{y}_t,y_t)+\inf_{\hat{y}_T}\sup_{\mathbf{z}_T}\left[\ell(\hat{y}_T,\mathbf{z}_T)-F(\mathbf{z}^T)\right].\end{aligned}$$

We now bound the second term. By the Minimax Switching Theorem 2.4, we have:

$$\begin{aligned}\inf_{\hat{y}_T}\sup_{\mathbf{z}_T}\left[\ell(\hat{y}_T,\mathbf{z}_T)-F(\mathbf{z}^T)\right]&=\sup_{\mu_T\in\Delta(\mathcal{X}\times\mathcal{Y})}\inf_{\hat{y}_T}\mathbb{E}_{\mathbf{z}_T\sim\mu_T}\left[\ell(\hat{y}_T,\mathbf{z}_T)-F(\mathbf{z}^T)\right]\\&=\sup_{\mu_T}\inf_{\hat{y}_T}\left[\mathbb{E}_{\mathbf{z}_T}[\ell(\hat{y}_T,\mathbf{z}_T)]-\mathbb{E}_{\mathbf{z}_T}[F(\mathbf{z}^T)]\right]\\&=\sup_{\mu_T}\left[\inf_{\hat{y}_T}\mathbb{E}_{\mathbf{z}_T}[\ell(\hat{y}_T,\mathbf{z}_T)]-\mathbb{E}_{\mathbf{z}_T}[F(\mathbf{z}^T)]\right]\\&=\sup_{\mu_T}\mathbb{E}_{\mathbf{z}_T}\left[\inf_{\hat{y}_T}\mathbb{E}_{\mathbf{z}_T}[\ell(\hat{y}_T,\mathbf{z}_T)]-F(\mathbf{z}^T)\right].\end{aligned}$$

Note now that

$$\sup_{\mu_1}\mathbb{E}_{\mathbf{z}_1\sim\mu_1}\cdots\sup_{\mu_T}\mathbb{E}_{\mathbf{z}_T\sim\mu_T}\overset{(a)}{\equiv}\sup_{\boldsymbol{\nu}^T\in\Delta((\mathcal{X}\times\mathcal{Y})^T)}\mathbb{E}_{\mathbf{z}^T\sim\boldsymbol{\nu}^T},$$

where $\boldsymbol{\nu}^T$ is a joint distribution over $(\mathcal{X}\times\mathcal{Y})^T$ and (a) follows by Skolemization. We conclude:

$$\mathsf{reg}_T(\mathcal{H})=\sup_{\boldsymbol{\nu}^T\in\Delta((\mathcal{X}\times\mathcal{Y})^T)}\mathbb{E}_{\mathbf{z}^T\sim\boldsymbol{\nu}^T}\left[\sum_{t=1}^{T}\inf_{\hat{y}_t}\mathbb{E}_{\mathbf{z}_t}[\ell(\hat{y}_t,\mathbf{z}_t)]-F(\mathbf{z}^T)\right]$$

and this completes the proof. □

4.3 Minimax Regret for Logarithmic Loss: Shtarkov Sum

We now consider a special yet important loss: the logarithmic loss, which is non-Lipschitz and unbounded. As a result, the guarantees from the previous section do not apply directly. Let $\mathcal{Y}$ denote the label space, and let $\Delta(\mathcal{Y})$ be the set of all probability distributions over $\mathcal{Y}$. The logarithmic loss for any $p \in \Delta(\mathcal{Y})$ and $y \in \mathcal{Y}$ is defined as:

$$\ell^{\mathsf{log}}(p, y) = -\log p[y]. \tag{4.8}$$

We start by considering the fixed design minimax regret introduced in (4.5) which for logarithmic loss becomes

$$\mathsf{reg}_T^*(\mathcal{H} \mid \mathbf{x}^T) = \inf_{\Phi} \sup_{y^T} \left[\sum_{t=1}^{T} \ell^{\mathsf{log}}(\hat{p}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\mathsf{log}}(h(\mathbf{x}_t), y_t) \right], \tag{4.9}$$

where $\hat{p}_t = \Phi(\mathbf{x}^T, y^{t-1}) \in \Delta(\mathcal{Y})$. We can express the fixed design minimax regret (4.9) via the so-called *Shtarkov sum* discussed next. Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a hypothesis class and $\mathbf{x}^T$ be any given instances. The *Shtarkov sum* of $\mathcal{H}$ conditioning on $\mathbf{x}^T$ is defined as

$$\mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T) = \sum_{y^T \in \mathcal{Y}^T} \sup_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t] \tag{4.10}$$

Example 4.1. Let $\mathcal{H}$ be a finite class, we have for any $\mathbf{x}^T$ that

$$\begin{aligned}
\mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T) &= \sum_{y^T \in \mathcal{Y}^T} \sup_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t] \\
&\le \sum_{y^T \in \mathcal{Y}^T} \sum_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t] \\
&= \sum_{h \in \mathcal{H}} \sum_{y^T \in \mathcal{Y}^T} \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t] \overset{(a)}{\le} \sum_{h \in \mathcal{H}} 1 = |\mathcal{H}|
\end{aligned}$$

where (a) follows from the fact that the second sum adds to 1 since it represents a distribution.

The next theorem shows that the fixed design minimax regret for logarithmic loss can be computed via the Shtarkov sum.

Theorem 4.3. Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be any hypothesis class, and let $\mathbf{x}^T$ be any given instances. Then

$$\mathsf{reg}_T^{\mathsf{fix}}(\mathcal{H} \mid \mathbf{x}^T) = \log \mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T). \tag{4.11}$$

Proof. We introduce the following short-hand notations

$$P_h(y^T \mid \mathbf{x}^T) = \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t], \qquad \hat{Q}(y^T) = \prod_{t=1}^{T} \hat{p}_t[y_t].$$

Observe, by definition of log-loss, that

$$\begin{aligned}
\mathsf{reg}_T^{\mathsf{fix}}(\mathcal{H} \mid \mathbf{x}^T) &= \inf_{\hat{Q}} \sup_{y^T} \left[-\log \hat{Q}(y^T) + \log \sup_h P_h(y^T \mid \mathbf{x}^T) \right] \\
&= \inf_{\hat{Q}} \sup_{y^T} \left[-\log \hat{Q}(y^T) + \log P^*(y^T \mid \mathbf{x}^T) \right] \\
&\quad + \log \sum_{y^T} \sup_h P_h(y^T \mid \mathbf{x}^T) \\
&\overset{(a)}{=} \log \sum_{y^T} \sup_h P_h(y^T \mid \mathbf{x}^T) = \log \mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T),
\end{aligned}$$

where $P^*(y^T \mid \mathbf{x}^T) := \frac{\sup_h P_h(y^T|\mathbf{x}^T)}{\sum \sup_h P_h(y^T|\mathbf{x}^T)}$ and (a) follows setting optimal $\hat{Q}(\cdot) \equiv P^*(\cdot \mid \mathbf{x}^T)$. □

A by-product of our previous proof shows that the minimax optimal predictor satisfies equality

$$\hat{Q}(\cdot) \equiv P^*(\cdot \mid \mathbf{x}^T),$$

where

$$P^*(y^T \mid \mathbf{x}^T) := \frac{\sup_h P_h(y^T \mid \mathbf{x}^T)}{\sum_{y^T} \sup_h P_h(y^T \mid \mathbf{x}^T)}$$

and $\hat{Q}(y^T) = \prod_{t=1}^{T} \hat{p}_t[y_t]$. To satisfy the equality, we can define

$$\hat{p}_t[y] = \frac{\sum_{y^{T-t}} P^*(y^{t-1} y y^{T-t} \mid \mathbf{x}^T)}{\sum_{y^{T-t+1}} P^*(y^{t-1} y^{T-t+1} \mid \mathbf{x}^T)}.$$

This is known as the Normalized Maximum Likelihood (NML) predictor.

Finally, we made two remarks. For a finite class $\mathcal{H}$, we immediately have

$$\mathsf{reg}_T^{\mathsf{fix}}(\mathcal{H} \mid \mathbf{x}^T) = \log \mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T) \leq \log |\mathcal{H}|.$$

Furthermore, the Shtarkov sum forms a lower bound for the (sequential) minimax regret:

$$\mathsf{reg}_T(\mathcal{H}) \geq \sup_{\mathbf{x}^T} \mathsf{reg}_T^{\mathsf{fix}}(\mathcal{H} \mid \mathbf{x}^T) \geq \sup_{\mathbf{x}^T} \log \mathsf{Sht}(\mathcal{H} \mid \mathbf{x}^T).$$

Minimax Regret via Contextual Shtarkov. Thus far, we have only dealt with fixed design minimax regret for logarithmic loss. However, it can also be extended to the sequential minimax regret. For this, we need a new concept, recently introduced by [28], called the contextual Shtarkov sum.

Let $\tau : \bigcup_{t=0}^{T} \mathcal{Y}^t \to \mathcal{X}$ be an $\mathcal{X}$-valued $|\mathcal{Y}|$-ary tree of depth T. The contextual Shtarkov sum w.r.t. τ is defined as

$$\mathsf{Sht}(\mathcal{H} \mid \tau) = \sum_{y^T} \sup_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\tau(y^{t-1}))[y_t]. \tag{4.12}$$

It turns out that the contextual Shtarkov is the needed tool to express the sequential minimax regret as shown in the below theorem.

Theorem 4.4. Let $\mathcal{H} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be any hypothesis class. Then:

$$\mathsf{reg}_T(\mathcal{H}) = \sup_{\tau} \log \mathsf{Sht}(\mathcal{H} \mid \tau).$$

Sketch of Proof. We provide only the high-level idea (for a detailed proof see [28]):

Step One: Using the minimax switching trick as in Theorem 2.4 and a truncation argument, we obtain the following Bayesian representation:

$$\sup_{\mathbf{x}_1, p_1} \mathbb{E}_{y_1 \sim p_1} \cdots \sup_{\mathbf{x}_T, p_T} \mathbb{E}_{y_T \sim p_T} \left[\sum_{t=1}^{T} \inf_{\hat{p}_t} \mathbb{E}_{y_t \sim p_t} \left[\ell^{\mathsf{log}}(\hat{p}_t, y_t) \right] - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\mathsf{log}}(h(\mathbf{x}_t), y_t) \right].$$

Step Two: Observe that:

$$\inf_{\hat{p}_t} \mathbb{E}_{y_t \sim p_t} \left[\ell^{\mathsf{log}}(\hat{p}_t, y_t) \right] = H(p_t),$$

where $H(p_t)$ is the Shannon entropy.

Step Three: Via Skolemization Lemma 2.2 the expression reduces to:

$$\sup_{\tau} \sup_{P} \mathbb{E}_{y^T \sim P} \left[H(P) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\mathsf{log}}(h(\tau(y^{t-1})), y_t) \right],$$

where τ runs over all trees $\tau : \bigcup_{t=1}^{T} \mathcal{Y}^t \to \mathcal{X}$ and $P \in \Delta(\mathcal{Y}^T)$.

Step Four: Denote $\mathbf{x}_t = \tau(y^{t-1})$, and let $P_h(y^T|\mathbf{x}^T) = \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t]$. We have

$$\inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell^{\mathsf{log}}(h(\tau(y^{t-1})), y_t) = \inf_{h} - \log P_h(y^T|\mathbf{x}^T) = - \sup_{h} \log P_h(y^T|\mathbf{x}^T).$$

Therefore, we find

$$\sup_{P} \mathbb{E}_{y^T \sim P} \left[H(P) + \log \sup_{h} P_h(y^T|\mathbf{x}^T) \right] = \sup_{P} \mathbb{E} \left[- \log P(y^T) + \log \sup_{h} P_h(y^T|\mathbf{x}^T) \right]$$

$$= \sup_{P} \mathbb{E} \left[- \log P(y^T) + \log P^*(y^T|\mathbf{x}^T) \right] + \log \sum_{y^T} \sup_{h} \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t]$$

$$= \underbrace{\sup_{P} -\mathsf{KL}(P, P^*)}_{=0} + \log \sum_{y^T} \sup_{h} \prod_{t=1}^{T} h(\mathbf{x}_t)[y_t].$$

Here, $P^*(y^T|\mathbf{x}^T) = \frac{\sup_h P_h(y^T|\mathbf{x}^T)}{\sum_{y^T} \sup_h P_h(y^T|\mathbf{x}^T)}$, and the last equality is attained at $P = P^*$. □

4.4 Bibliographical Notes

There is a vast literature on redundancy and regret in information theory [1], [29], [30], [2], [6]–[8], [13]. The notions of regret and minimax regret in machine learning are well known and have been extensively discussed; see, e.g., [17], [18].

Theorem 4.1 appears to be known (see [9]); however, the proof presented in this chapter seems to be new. The fixed-design minimax regret has been previously studied in the machine learning literature under the name *Transductive Online Learning*; see [31]. The Bayesian representation of minimax regret in Theorem 4.2 was developed by [27].

The logarithmic loss is extensively discussed in the information theory literature in the context of source coding. In 1984, Shtarkov [12] introduced the worst-case minimax redundancy for source coding and used the maximum likelihood distribution to derive a precise inequality on the worst-case redundancy via the Shtarkov sum. This was translated in [2] into an exact expression for worst-case redundancy, where sequences modulo 1 are used to characterize redundancy for Shannon and Huffman codes. A novel extension of the Shtarkov sum to fixed design regret was introduced in [32] and to the sequential minimax regret in [28], where the *contextual Shtarkov sum* is introduced.

5

Minimax Regret with Log-loss

In this chapter, we present novel results for minimax regret with logarithmic loss (log-loss), as defined in (4.8). We consider only binary-valued labels $\mathcal{Y} = \{0, 1\}$, for which the log-loss can be equivalently written as

$$\ell(\hat{y}_t, y_t) = -y_t \log(\hat{y}_t) - (1 - y_t) \log(1 - \hat{y}_t),$$

where $y_t \in \mathcal{Y}$ and $\hat{y}_t \in [0, 1]$ is the prediction, interpreted as the probability assigned to label 1. Our main tool in this section will be the Shtarkov sum, already discussed in Chapter 4, which we evaluate precisely to obtain tight bounds on the minimax regret.

5.1 Bayesian Predictor

We begin with a general Bayesian predictor for log-loss, which we will later refine to obtain tighter bounds. Let $\mathcal{G}$ be any reference class of (sequential) functions mapping $\mathcal{X}^* \to [0, 1]$, where $\mathcal{X}^*$ denotes finite sequences over $\mathcal{X}$, and let $\mathcal{Y} = \{0, 1\}$. Let $\mathcal{W}$ be an index set for $\mathcal{G}$, and let μ be an arbitrary finite measure over $\mathcal{W}$. The standard Bayesian predictor with prior μ is presented in Algorithm 5.1. Based on this algorithm, the following two lemmas are used to establish most of the upper bounds presented in this chapter.

Algorithm 5.1 Bayesian predictor

```
Input: Reference class $\mathcal{G} := \{g_w : w \in \mathcal{W}\}$ with index set $\mathcal{W}$ and prior $\mu$ over $\mathcal{W}$
1: Set $p_w(y^0 \mid \mathbf{x}^0) = 1$ for all $w \in \mathcal{W}$.
2: for $t = 1, \cdots, T$ do
3:     Receive feature vector $\mathbf{x}_t$
4:     Make prediction with the following equation:
           $$\hat{y}_t = \frac{\int_{\mathcal{W}} g_w(\mathbf{x}^t) p_w(y^{t-1} \mid \mathbf{x}^{t-1}) \mathrm{d}\mu}{\int_{\mathcal{W}} p_w(y^{t-1} \mid \mathbf{x}^{t-1}) \mathrm{d}\mu}. \tag{5.1}$$
5:     Receive label $y_t$
6:     For all $w \in \mathcal{W}$, update:
           $$p_w(y^t \mid \mathbf{x}^t) = e^{-\ell(g_w(\mathbf{x}^t), y_t)} p_w(y^{t-1} \mid \mathbf{x}^{t-1}). \tag{5.2}$$
7: end for
```

Lemma 5.1. Let $\mathcal{G}$ be a collection of functions $g_w : \mathcal{X}^* \to [0,1], w \in \mathcal{W}$. Let $\hat{y}_t$ be the Bayesian prediction rule (5.1) as in Step 4 of Algorithm 5.1 with prior μ. Then, for any $\mathbf{x}^T$ and y^T we have

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq -\log \frac{\int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}} 1 \mathrm{d}\mu}$$

where

$$p_w(y^T \mid \mathbf{x}^T) = e^{-\sum_{t=1}^{T} \ell(g_w(\mathbf{x}^t), y_t)} = \prod_{t=1}^{T} g_w(x^t)^{y_t} (1 - g_w(x^t))^{1-y_t}$$

and ℓ is the log-loss as in equation (4.8).

Proof. We first observe that for any $y \in \{0,1\}$ we have $e^{-\ell(\cdot, y)}$ is concave over $[0,1]$. Let

$$\lambda_{t-1}(w) = \frac{p_w(y^{t-1} \mid \mathbf{x}^{t-1})}{\int_{\mathcal{W}} p_w(y^{t-1} \mid \mathbf{x}^{t-1}) \mathrm{d}\mu}.$$

Note that $\lambda_{t-1}(w)$ forms a probability density over $\mathcal{W}$ w.r.t. μ. By definition of $\hat{y}_t$, we have $\hat{y}_t = \mathbb{E}_{\lambda_{t-1}}[g_w(\mathbf{x}^t)]$, where the expectation is over the density of $\lambda_{t-1}(w)$. Therefore, by Jensen's inequality and the update procedure as in item 6 of Algorithm 5.1, we have

$$e^{-\ell(\hat{y}_t,y_t)} = e^{-\ell(\mathbb{E}[g_w(\mathbf{x}^t)],y_t)} \geq \mathbb{E}[e^{-\ell(g_w(\mathbf{x}^t),y_t)}] = \frac{\int_{\mathcal{W}} p_w(y^t \mid \mathbf{x}^t)\mathrm{d}\mu}{\int_{\mathcal{W}} p_w(y^{t-1} \mid \mathbf{x}^{t-1})\mathrm{d}\mu}.$$

By telescoping the sum, we find

$$e^{-\sum_{t=1}^{T} \ell(\hat{y}_t,y_t)} \geq \frac{\int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T)\mathrm{d}\mu}{\int_{\mathcal{W}} 1\mathrm{d}\mu}.$$

This implies

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq -\log \frac{\int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T)\mathrm{d}\mu}{\int_{\mathcal{W}} 1\mathrm{d}\mu}$$

and completes the proof. □

The following lemma bounds the regret under log-loss of finite classes.

Lemma 5.2. For any finite class of experts $\mathcal{G}$, we have $\mathsf{reg}_T(\mathcal{G}) \leq \log|\mathcal{G}|$.

Proof. Let $\mu(w) = \frac{1}{|\mathcal{W}|}$ as in Lemma 5.1 and $\hat{y}_t$ be the Bayesian predictor with input $\mathcal{G}$ and μ. Then

$$\begin{aligned}\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) &\leq -\log \frac{\int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T)\mathrm{d}\mu}{\int_{\mathcal{W}} 1\mathrm{d}\mu} \\ &= -\log \int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T)\mathrm{d}\mu + \log 1 \\ &= -\log \int_{\mathcal{W}} p_w(y^T \mid \mathbf{x}^T)\mathrm{d}\mu \\ &\leq -\log p_{w^*}(y^T \mid \mathbf{x}^T) + \log|\mathcal{W}|, \;\; w^* \text{ maximizes } p_w(y^T \mid \mathbf{x}^T) \\ &= \sum_{t=1}^{T} \ell(g_{w^*}(\mathbf{x}^t), y_t) + \log|\mathcal{G}|, \quad \text{since } |\mathcal{W}| = |\mathcal{G}|.\end{aligned}$$

This concludes the proof. □

We should remark that the regrets established in this section hold for (sequential) function classes $\mathcal{G}$, which generate outputs using the entire feature history $\mathbf{x}^t$ at each time step t. This is in contrast to the static function class $\mathcal{H}$ discussed earlier, which generates outputs based only on the current feature $\mathbf{x}_t$. This notion is also sometimes referred to as *dynamic functions* as in [33].

5.2 General Upper Bound via Smoothing

We begin by introducing a notion of a covering, called the *global sequential cover*, which was implicitly used in [27, Section 6.1] to derive regret bounds under the absolute loss, and traces back to ideas in [24].

Definition 5.1 (Global sequential covering)**.** For any $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, we say class $\mathcal{G}$ of functions map $\mathcal{X}^* \to [0,1]$ is a global *sequential* α-covering of $\mathcal{H}$ at scale α if for any $\mathbf{x}^T \in \mathcal{X}^T$ and $h \in \mathcal{H}$, there exists $g \in \mathcal{G}$ such that $\forall t \in [T]$,

$$|h(\mathbf{x}_t) - g(\mathbf{x}^t)| \leq \alpha.$$

Throughout we assume that $0 \leq \alpha \leq 1$.

To apply the global sequential cover we need to modify the standard Bayesian Algorithm 5.1 to the so called smooth truncated Bayesian predictor presented in Algorithm 5.2.

Algorithm 5.2 Smooth truncated Bayesian predictor

Input: Reference class $\mathcal{G}$ with index set $\mathcal{W}$ and prior μ over $\mathcal{W}$, and truncation parameter α.

```
Let $p_w(y^0 \mid \mathbf{x}^0) = 1$ for all $w \in \mathcal{W}$
for $t = 1, \cdots, T$ do
    Receive feature $\mathbf{x}_t$
    For all $w \in \mathcal{W}$, set
        $\tilde{g}_w(\mathbf{x}^t) = \frac{g_w(\mathbf{x}^t) + \alpha}{1 + 2\alpha}$    (5.3)
    Make prediction
        $\hat{y}_t = \frac{\int_{\mathcal{W}} \tilde{g}_w(\mathbf{x}^t) p_w(y^{t-1} \mid \mathbf{x}^{t-1}) \mathrm{d}\mu}{\int_{\mathcal{W}} p_w(y^{t-1} \mid \mathbf{x}^{t-1}) \mathrm{d}\mu}$    (5.4)
    Receive label $y_t$
    For all $w \in \mathcal{W}$, update:
        $p_w(y^t \mid \mathbf{x}^t) = e^{-\ell(\tilde{g}_w(\mathbf{x}^t), y_t)} p_w(y^{t-1} \mid \mathbf{x}^{t-1}).$    (5.5)
end for
```

We are now in the position to state our first main general finding.

Theorem 5.3. If, for any $\alpha > 0$, there exists a global sequential α-covering set $\mathcal{G}_\alpha$ of $\mathcal{H}$, then for log-loss, the minimax regret satisfies

$$\mathsf{reg}_T(\mathcal{H}) \le \inf_{0<\alpha<1} \{T\log(1+2\alpha) + \log|\mathcal{G}_\alpha|\} \le \inf_{0<\alpha<1} \{2\alpha T + \log|\mathcal{G}_\alpha|\}, \tag{5.6}$$

and this bound is achieved by Algorithm 5.2.

The proof is based on the following key lemma.

Lemma 5.4. Suppose $\mathcal{H}$ has a global sequential α-covering set $\mathcal{G}$ for some $\alpha \in [0,1]$. Then, there exists a truncated set $\tilde{\mathcal{G}}$ of $\mathcal{G}$ with $|\tilde{\mathcal{G}}| = |\mathcal{G}|$ such that for all $\mathbf{x}^T, y^T$ and $h \in \mathcal{H}$ there exists a $\tilde{g} \in \tilde{\mathcal{G}}$ satisfying

$$\frac{p_h(y^T \mid \mathbf{x}^T)}{p_{\tilde{g}}(y^T \mid \mathbf{x}^T)} \le (1+2\alpha)^T, \tag{5.7}$$

where

$$p_h(y^T \mid \mathbf{x}^T) = \prod_{t=1}^T h(\mathbf{x}_t)^{y_t}(1-h(\mathbf{x}_t))^{1-y_t}$$

and

$$p_{\tilde{g}}(y^T \mid \mathbf{x}^T) = \prod_{t=1}^T \tilde{g}(\mathbf{x}^t)^{y_t}(1-\tilde{g}(\mathbf{x}^t))^{1-y_t}.$$

Proof. We construct the set $\tilde{\mathcal{G}}$ as in Algorithm 5.2. For any $g \in \mathcal{G}$, we define a smooth truncated function $\tilde{g}$ such that for any $\mathbf{x}^t \in \mathcal{X}^*$

$$\tilde{g}(\mathbf{x}^t) = \frac{g(\mathbf{x}^t)+\alpha}{1+2\alpha}.$$

We introduce the following short-hand notation; for any function f, we define $f(y_t) = f(\mathbf{x}^t)^{y_t}(1-f(\mathbf{x}^t))^{1-y_t}$. For any $\mathbf{x}^T, y^T$, and $h \in \mathcal{H}$, let $g \in \mathcal{G}$ be a α-covering of h and $\tilde{g}$ be the truncated function as defined above. Now, the key observation is that for any $y_t \in \{0,1\}$, we have $h(y_t) \le g(y_t) + \alpha$ since g α-covers h. This implies that

$$\frac{h(y_t)}{\tilde{g}(y_t)} \le \frac{g(y_t)+\alpha}{(g(y_t)+\alpha)/(1+2\alpha)} = 1+2\alpha.$$

Therefore, we have

$$\begin{aligned}\frac{p_h(y^T \mid \mathbf{x}^T)}{p_{\tilde{g}}(y^T \mid \mathbf{x}^T)} &= \prod_{t=1}^{T} \frac{h(y_t)}{\tilde{g}(y_t)} \\ &\leq (1+2\alpha)^T .\end{aligned}$$

This completes the proof of Lemma 5.4. □

Proof of Theorem 5.3. We show that for any $0 < \alpha < 1$ if an α-covering set $\mathcal{G}_\alpha$ exists, then one can achieve the claimed bound for such an α. To do so, we run the Smooth truncated Bayesian Algorithm 5.2 on $\mathcal{G}_\alpha$ with uniform prior and truncation parameter α. We denote by $\tilde{\mathcal{G}}_\alpha$ to be the truncated class of $\mathcal{G}_\alpha$ as in Lemma 5.4. We now fix $\mathbf{x}^T, y^T$. By Lemma 5.2 (with $\mathcal{G}$ being $\tilde{\mathcal{G}}_\alpha$), we have

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{\tilde{g}\in\tilde{\mathcal{G}}_\alpha} \sum_{t=1}^{T} \ell(\tilde{g}(\mathbf{x}^t), y_t) + \log|\tilde{\mathcal{G}}_\alpha| = \inf_{\tilde{g}\in\tilde{\mathcal{G}}_\alpha} \sum_{t=1}^{T} \ell(\tilde{g}(\mathbf{x}^t), y_t) + \log|\mathcal{G}_\alpha|,$$

the last equality follows from $|\mathcal{G}_\alpha| = |\tilde{\mathcal{G}}_\alpha|$. Since $\sum_{t=1}^{T} \ell(f(\mathbf{x}^t), y_t) = -\log p_f(y^T \mid \mathbf{x}^T)$ for any function f, then by Lemma 5.4 we conclude that

$$\inf_{h\in\mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \geq \inf_{\tilde{g}\in\tilde{\mathcal{G}}_\alpha} \sum_{t=1}^{T} \ell(\tilde{g}(\mathbf{x}^t), y_t) - T\log(1+2\alpha) .$$

The result follows by combining the inequalities and noticing that $\log(1+x) \leq x$ for all $x \geq -1$. □

We will demonstrate how Theorem 5.3 can be applied in various contexts to obtain tight regret upper bounds by appropriately designing the covering set $\mathcal{G}$ in the following sections.

5.3 Lipschitz Parametric Class.

We now consider a Lipschitz parametric function class. Given a function $f : \mathcal{W} \times \mathcal{X} \to [0,1]$, define the following class

$$\mathcal{H}_f = \{f(\mathbf{w}, \cdot) \in [0,1]^{\mathcal{X}} \ : \ \mathbf{w} \in \mathcal{W}\},$$

where $\mathbf{w} \in \mathcal{W}$ is often a d-dimensional vector in $\mathbb{R}^d$.

We will assume that $f(\mathbf{w}, \mathbf{x})$ is L-Lipschitz on $\mathbf{w}$ for every $\mathbf{x}$, where $L \in \mathbb{R}^+$. More formally, $\forall \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$ and $\mathbf{x} \in \mathcal{X}$ we have

$$|f(\mathbf{w}_1, \mathbf{x}) - f(\mathbf{w}_2, \mathbf{x})| \leq L||\mathbf{w}_1 - \mathbf{w}_2||,$$

where $||\cdot||$ is some norm on $\mathcal{W}$. For example, if we take $\mathcal{W} \subset \mathbb{R}^d$ then the norm can be ℓ_1, ℓ_2 or ℓ_∞ norm. For any specific norm $||\cdot||$, we write $\mathcal{B}(R)$ for the ball under such norm with radius R in $\mathcal{W}$. In particular, we denote by $\mathcal{B}_s^d(R)$ the ball in $\mathbb{R}^d$ of radius R under ℓ_s norm centered at the origin.

Theorem 5.5. Let $f : \mathcal{B}_s^d(R) \times \mathbb{R}^d \to [0, 1]$ be a L-Lipschitz function under ℓ_s norm. Then

$$\mathsf{reg}_T(\mathcal{H}_f) \leq \min\left\{d \log\left(\frac{2RLT}{d} + 1\right) + 2d, T\right\}. \tag{5.8}$$

Proof. By L-Lipschitz condition, to find an α-covering in the sense of Definition 5.1, we only need to find a covering of $\mathcal{B}_s^d(R)$ with radius α/L. By standard result (see e.g. Lemma 5.7 and Example 5.8 of [34]) we know that the covering size is upper bounded by

$$\left(\frac{2RL}{\alpha} + 1\right)^d.$$

By Theorem 5.3, we find

$$\mathsf{reg}_T(\mathcal{H}_f) \leq \inf_{0<\alpha<1}\left\{2\alpha T + d \log\left(\frac{2RL}{\alpha} + 1\right)\right\}.$$

Taking $\alpha = d/T$, we conclude

$$\mathsf{reg}_T(\mathcal{H}_f) \leq d \log\left(\frac{2RLT}{d} + 1\right) + 2d.$$

This completes the proof for $T \geq d$. The upper bound T is achieved by predicting $\frac{1}{2}$ every time. □

Example 5.1. For logistic function $f(\mathbf{w}, \mathbf{x}) = (1 + e^{-\langle \mathbf{w}, \mathbf{x} \rangle})^{-1}$, and $\mathbf{w} \in \mathcal{B}_2^d(R)$ with $\mathbf{x} \in \mathcal{B}_2^d(1)$ our result recovers those of [35], but with a better leading constant (the bound in [35] has a constant 5). Note that, the result in [36] also provides a sub-optimal constant $c \sim 4$. Moreover, our bounds have a logarithmic dependency on Lipschitz constant L.

The question arises whether the factor in front of $\log T$ can be improved to $d/2$ instead of d as discussed in some recent papers [32], [37], [38]. In Theorem 5.6 below, we show that, in general, it cannot unless we further strengthen our assumption (see Theorem 5.9). For the ease of presentation, we only consider the parameters restricted to ℓ_2 norm.

Theorem 5.6. For any d, T, R, L such that $T \gg d\log(RLT)$, there exists L-Lipschitz function $f : \mathcal{B}_2^d(R) \times \mathbb{R}^d \to [0,1]$ such that

$$\mathsf{reg}_T(\mathcal{H}_f) \geq d\log\left(\frac{RLT}{d}\right) - d\log 64 - d\log\log(RLT). \tag{5.9}$$

We need the following two lemmas to prove Theorem 5.6.

Lemma 5.7. Let $\mathcal{P}$ be a finite class of distributions over the same domain Ω. Denote

$$S = \sum_{\omega\in\Omega} \max_{p\in\mathcal{P}} p(\omega)$$

to be the Shtarkov sum. Then for any estimation rule $\Phi : \Omega \to \mathcal{P}$ we have:

$$S \geq |\mathcal{P}| \cdot \left(1 - \max_{p\in\mathcal{P}} p\left(\{\omega : \Phi(\omega) \neq p\}\right)\right)$$

Proof. Note that Φ partitions Ω into $|\mathcal{P}|$ disjoint parts. For any $p \in \mathcal{P}$, we denote $\Omega_p = \{\omega \in \Omega : \Phi(\omega) = p\}$ to be the partition corresponding to p. We have:

$$\sum_{p\in\mathcal{P}} p(\Omega_p) = \sum_{\omega\in\Omega} p_\omega(\omega) \leq \sum_{\omega\in\Omega} \max_{p\in\mathcal{P}} p(\omega) = S,$$

where $p_\omega \in \mathcal{P}$ is the distribution such that $\omega \in \Omega_{p_\omega}$. This implies

$$\min_{p\in\mathcal{P}} p(\Omega_p) \leq \frac{S}{|\mathcal{P}|}.$$

The result follows by taking the complements of Ω_p. □

Lemma 5.8. For any $M \leq e^{T/8}$, there exist M vectors $v_1, v_2, \cdots, v_M \in \{0,1\}^T$ such that for any $i \neq j \in [M]$ we have:

$$\sum_{t=1}^{T} 1\{v_i[t] \neq v_j[t]\} \geq T/4.$$

Proof. This follows from standard packing number estimates of the Boolean cube, see [34, Example 5.3]. □

Proof of Theorem 5.6. Let $\mathbf{x}_1, \cdots, \mathbf{x}_T \in \mathbb{R}^d$ be any distinct points. We will construct a L-Lipschitz function $f(\mathbf{w}, \mathbf{x})$ such that the regret restricted only on $\mathbf{x}^T$ is large. To do so, we consider a maximum packing M of the parameter space $\mathcal{B}_2^d(R)$ of radius $\alpha/L > 0$ (where α is to be determined latter). Standard volume argument (see Chapter 5 of [34]) yields that

$$|M| \geq \left(\frac{LR}{2\alpha}\right)^d.$$

Now, we will define a L-Lipschitz functions $f(\mathbf{w}, \mathbf{x})$ only on $\mathbf{w} \in M$ and $\mathbf{x} \in \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$. By Lemma 5.8 (assume for now the conditions are satisfied), we can find $|M|$ binary vectors $V \subset \{0,1\}^T$ such that any pair of the vectors has Hamming distance lower bounded by $T/4$. For each of the vector $v \in V$, we define a vector $u \in [0,1]^T$ in the following way, for all $t \in [T]$

1. If $v[t] = 0$ then set $u[t] = 0$;
2. If $v[t] = 1$ then set $u[t] = \alpha$.

Denote by U be the set of all such vectors u. Note that $|U| = M$. For any $\mathbf{w} \in M$, we can associate a unique $u \in U$ such that for all $t \in [T]$

$$f(\mathbf{w}, \mathbf{x}_t) = u[t].$$

We now show that f is indeed L-Lipschitz restricted on M for all $\mathbf{x}_t \in \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$. This is because for any $\mathbf{w}_1 \neq \mathbf{w}_2 \in M$ we have $|f(\mathbf{w}_1, \mathbf{x}_t) - f(\mathbf{w}_2, \mathbf{x}_t)| \leq \alpha$ by definition of U and $||\mathbf{w}_1 - \mathbf{w}_2||_2 \geq \alpha/L$ since M is a packing.

We now view the vectors in $u \in U$ as a product of Bernoulli distributions with each coordinate t independently sampled from $\text{Bern}(u[t])$. We show that the sources in U are identifiable. To see this, we note that for any distinct pairs $u_1, u_2 \in U$, there exist a set $I \in [T]$ such that u_1 and u_2 differ on I and $|I| \geq T/4$. This further implies that there exists a set $J \subset I$ with $|J| \geq T/8$ such that u_1 takes all 0 on J and u_2 takes all α on J (or vice versa). We can then distinguish u_1, u_2 by checking if the samples on J are all 0s or not. The probability of making error is upper bounded by

$$(1-\alpha)^{T/8} \leq e^{-\alpha T/8}.$$

Since there are only $|M|^2$ such pairs, we have the probability of wrongly identifying the source upper bounded by

$$|M|^2 e^{-\alpha T/8}.$$

Taking $\alpha = \frac{16d\log(RLT)}{T}$, the error probability is upper bounded by

$$\left(\frac{RLT}{32d\log(RLT)}\right)^{2d} e^{-2d\log(RLT)} \leq \left(\frac{1}{32d\log(RLT)}\right)^{2d} \leq \frac{1}{2},$$

for sufficient large d, T, where we have use the fact that

$$|M| \leq (\frac{RLT}{32d\log(RLT)})^d.$$

Note that we only showed a lower bound on $|M|$ before, but this is not a problem since we can always remove some points from M to make the upper bound holds as well.

By Lemma 5.7, we know that the Shtarkov sum of sources in U is lower bounded by $|M|/2$. Therefore, we have

$$\begin{aligned}\mathsf{reg}_T(\mathcal{H}_f) \geq \mathsf{reg}_T^*(\mathcal{H}_f) &\geq \log(|M|/2) \\ &\geq d\log\left(RLT/d\right) - d\log 64 - d\log\log(RLT).\end{aligned}$$

Now, we have to extend the function to the whole set $\mathcal{B}_2^d(R)$ and keep the L-Lipschitz property. This follows from a classical result in real analysis (see [39, Theorem 1]) by defining for all $\mathbf{w} \in \mathcal{B}_2^d(R)$ and $\mathbf{x}_t \in \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$

$$f(\mathbf{w}, \mathbf{x}_t) = \sup_{\mathbf{w}' \in M} \{f(\mathbf{w}', \mathbf{x}_t) - L||\mathbf{w} - \mathbf{w}'||_2\}.$$

For the $\mathbf{x} \notin \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$, we can simply let $f(\mathbf{w}, \mathbf{x}) = 0$ for all $\mathbf{w}$.

Finally, we need to check that the condition of Lemma 5.8 holds for our choice of α, this is satisfied by our assumption $T \gg d\log(RLT)$. □

5.3.1 Lipschitz Class with Bounded Hessian.

As we have demonstrated in Theorem 5.6 the leading constant 1 of the regret for Lipschitz parametric classes can not be improved in general. We now show that for some special function $f \in \mathcal{H}_f$ one can improve the constant to $\frac{1}{2}$, as already noticed in [32], [37], [38]. For any function $f : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$, we say the Hessian of $\log f$ is uniformly bounded on $\mathcal{X} \subset \mathbb{R}^d$, if there exists a constant C such that for any $\mathbf{w} \in \mathbb{R}^d$ and $\mathbf{x} \in \mathcal{X}$ and $y \in \{0,1\}$ we have

$$\sup_{||\mathbf{u}||_2 \leq 1} |\mathbf{u}^T \nabla^2_{\mathbf{w}} \log f(\mathbf{w},\mathbf{x})^y (1 - f(\mathbf{w},\mathbf{x}))^{1-y} \mathbf{u}| \leq C,$$

where $\nabla^2_{\mathbf{w}}$ is the Hessian at $\mathbf{w}$.

Theorem 5.9. Let $f : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$ be a function such that the Hessian of $\log f$ is uniformly bounded by C on $\mathcal{X}$. Let

$$\mathcal{H}_f = \{f(\mathbf{w},\mathbf{x}) : \mathbf{w} \in \mathcal{W}, \mathbf{x} \in \mathcal{X}\}$$

be such that f is restricted to some compact set $\mathcal{W} \subset \mathbb{R}^d$. Then for any $\epsilon > 0$,

$$\mathsf{reg}_T(\mathcal{H}_f) \leq \log \frac{\mathrm{Vol}(\mathcal{W}^*)}{\mathrm{Vol}(\mathcal{B}_2^d(\epsilon))} + \frac{1}{2} C T \epsilon^2 + \log 2. \tag{5.10}$$

where $\mathcal{W}^* = \{\mathbf{w} + \mathbf{u} \mid \mathbf{w} \in \mathcal{W},\ \mathbf{u} \in \mathcal{B}_2^d(\epsilon)\}$, $\mathrm{Vol}(\cdot)$ is volume under Lebesgue measure. In particular, for $\mathcal{W} = \mathcal{B}_2^d(R)$ and $\epsilon = \sqrt{d/CT}$, we have

$$\mathsf{reg}_T(\mathcal{H}_f) \leq \frac{d}{2} \log \left(\frac{2CR^2T}{d} + 2 \right) + d/2 + \log 2.$$

Note that, Theorem 5.9 subsumes the results of [37], [40], where the authors considered the function of form $f(\langle \mathbf{w}, \mathbf{x} \rangle)$ and requires that the second derivative of $\log f$ is bounded, see also [9, Chapter 11.10]. However, the KL-divergence-based argument of [40] can not be used directly in the setup of Theorem 5.9 since we *do not* assume the function f has a linear structure. Our main proof technique of Theorem 5.9 is a direct application of Lemma 5.1 and an estimation of the integrals via Taylor expansion.

Proof of Theorem 5.9. The proof resembles that of [35] but running the Bayesian predictor (Algorithm 5.1) over $\mathcal{W}^*$ instead of $\mathcal{W}$ with $\mathcal{G}$ being $\mathcal{H}_f$ and μ being Lebesgue measure. Let $\mathbf{x}^T$, y^T and $\hat{y}^T$ be the feature, label and predictions of the Bayesian predictor respectively. By Lemma 5.1

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq -\log \frac{\int_{\mathcal{W}^*} p_{\mathbf{w}}(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}^*} 1 \mathrm{d}\mu}, \tag{5.11}$$

where μ is the Lebesgue measure and

$$p_{\mathbf{w}}(y^T \mid \mathbf{x}^T) = \prod_{t=1}^{T} f(\mathbf{w}, \mathbf{x}_t)^{y_t} (1 - f(\mathbf{w}, \mathbf{x}_t))^{1-y_t}.$$

We now write $h_t(\mathbf{w}) \overset{\text{def}}{=} \log f(\mathbf{w}, \mathbf{x}_t)^{y_t}(1 - f(\mathbf{w}, \mathbf{x}_t))^{1-y_t}$ to simplify notation. It is easy to see that $\ell(f(\mathbf{w}, \mathbf{x}_t), y_t) = -h_t(\mathbf{w})$. Let $\mathbf{w}^*$ be the point in $\mathcal{W}$ that maximizes

$$h(\mathbf{w}) \overset{\text{def}}{=} \sum_{t=1}^{T} h_t(\mathbf{w}).$$

Let $\mathbf{u} = \nabla h(\mathbf{w}^*)$ be the gradient of h at $\mathbf{w}^*$. By Taylor theorem, we have for any $\mathbf{w} \in \mathcal{W}^*$

$$h(\mathbf{w}) = h(\mathbf{w}^*) + \mathbf{u}^\tau (\mathbf{w} - \mathbf{w}^*) + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^\tau \nabla^2_{\mathbf{w}'} h(\mathbf{w}')(\mathbf{w} - \mathbf{w}^*),$$

where $\mathbf{w}'$ is a convex combination of $\mathbf{w}$ and $\mathbf{w}^*$ and $\mathbf{u}^\tau$ is the transpose of $\mathbf{u}$.

Now, the key observation is that for any point $\mathbf{w}$ such that $\mathbf{u}^\tau(\mathbf{w} - \mathbf{w}^*) \geq 0$ we have

$$h(\mathbf{w}) \geq h(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^\tau \nabla^2_{\mathbf{w}'} h(\mathbf{w}')(\mathbf{w} - \mathbf{w}^*) \geq h(\mathbf{w}^*) - \frac{1}{2} CT ||\mathbf{w} - \mathbf{w}^*||_2^2, \tag{5.12}$$

where the last inequality follows from our assumption about the bounded Hessian of $\log f$. Let B be the half ball of radius ϵ centered at $\mathbf{w}^*$ such that for all $\mathbf{w} \in B$ we have $\mathbf{u}^T(\mathbf{w} - \mathbf{w}^*) \geq 0$. By (5.12), for all $\mathbf{w} \in B$

$$h(\mathbf{w}) \geq h(\mathbf{w}^*) - \frac{1}{2} CT\epsilon^2. \tag{5.13}$$

Note that $B \subset \mathcal{W}^*$. Then using above observations we arrive at

$$\begin{aligned}
\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) &\leq -\log \frac{\int_{\mathcal{W}^*} p_{\mathbf{w}}(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}^*} 1 \mathrm{d}\mu} && (5.14)\\
&\leq -\log \frac{\int_{B} p_{\mathbf{w}}(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}^*} 1 \mathrm{d}\mu}, \text{ since } B \subset \mathcal{W}^* && (5.15)\\
&\leq -\log \frac{e^{-CT\epsilon^2/2} \int_{B} p_{\mathbf{w}^*}(y^T \mid \mathbf{x}^T) \mathrm{d}\mu}{\int_{\mathcal{W}^*} 1 \mathrm{d}\mu} && (5.16)\\
&= -\log p_{\mathbf{w}^*}(y^T \mid \mathbf{x}^T) + CT\epsilon^2/2 - \log \frac{\mathrm{Vol}(B)}{\mathrm{Vol}(\mathcal{W}^*)} && (5.17)\\
&= -\log p_{\mathbf{w}^*}(y^T \mid \mathbf{x}^T) + CT\epsilon^2/2 - \log \frac{\frac{1}{2}\mathrm{Vol}(\mathcal{B}_2^d(\epsilon))}{\mathrm{Vol}(\mathcal{W}^*)} && (5.18)\\
&= \sum_{t=1}^{T} \ell(f(\mathbf{w}^*, \mathbf{x}_t), y_t) + \log \frac{\mathrm{Vol}(\mathcal{W}^*)}{\mathrm{Vol}(\mathcal{B}_2^d(\epsilon))} + CT\epsilon^2/2 + \log 2. && (5.19)
\end{aligned}$$

This completes the proof of the general bound. The last part of the theorem follows from the fact that if $\mathcal{W} = \mathcal{B}_2^d(R)$ then $\mathcal{W}^* = \mathcal{B}_2^d(R+\epsilon)$ and noticing that

$$\frac{\mathrm{Vol}(\mathcal{B}_2^d(R+\epsilon))}{\mathrm{Vol}(\mathcal{B}_2^d(\epsilon))} \leq (R/\epsilon + 1)^d$$

as desired. □

Remark 5.1. When compared to the technique in [41], Theorem 5.9 does not assume that the gradient critical point of the loss is zero (e.g., the minimum may occur on the boundary). This is why we need to restrict to the half ball B in order to discard the linear term of Taylor expansion in Equation (5.13). Moreover, in the proof we work directly on the continuous space instead of a discretized cover, giving an efficient algorithm provided the posterior is efficiently samplable (by e.g., assuming some log-concavity of f as in [35]).

We complete this section with the following lower bound for generalized linear functions under unit ℓ_s balls.

Theorem 5.10. Let $f : \mathbb{R} \to [0, 1]$ be an arbitrary function such that there exists $c_1, c_2 \in (0, 1)$ and for all $r > 0$ we have $[c_1 - c_2 d^{-r}, c_1 + c_2 d^{-r}] \subset f([-d^{-r}, d^{-r}])$ for all sufficiently large d. Let

$$\mathcal{H}_f = \{f(\langle \mathbf{w}, \mathbf{x}\rangle) : \mathbf{w} \in \mathcal{B}_s^d(1), \mathbf{x} \in \mathcal{B}_s^d(1)\}$$

where $s > 0$. Then

$$\mathsf{reg}_T(\mathcal{H}_f) \geq \frac{d}{2} \log\left(\frac{T}{d^{(s+2)/s}}\right) - O(d) \tag{5.20}$$

where O hides some absolute constant that is independent of d, T.

Example 5.2 (Logistic function)**.** Note that for the logistic function $f(x) = (1 + e^{-x})^{-1}$ Theorem 5.10 holds with $c_1 = \frac{1}{2}$ and $c_2 = \frac{1}{5}$. Therefore,
1. If $s = 1$, then

$$\mathsf{reg}_T(\mathcal{H}_f) \geq \frac{d}{2} \log\left(\frac{T}{d^3}\right) - O(d).$$

2. If $s = 2$, then

$$\mathsf{reg}_T(\mathcal{H}_f) \geq \frac{d}{2} \log\left(\frac{T}{d^2}\right) - O(d).$$

3. If $s = \infty$, then

$$\mathsf{reg}_T(\mathcal{H}_f) \geq \frac{d}{2} \log\left(\frac{T}{d}\right) - O(d).$$

This recovers all the lower bounds from [37]. We note that a simple sufficient condition for Theorem 5.10 to hold is to require $f'(0) \neq 0$ if $f(x)$ is differentiable.

Remark 5.2. We should remark that the leading constants of Theorem 5.9 and 5.10 are matching only if d growth sub-polynomially w.r.t. T (e.g., when $d = e^{\sqrt{\log T}}$). However, when d growth polynomailly w.r.t. T the leading constants will no longer match, though the bounds are still having the same asymptotic rate $\Theta(d \log T)$ for $d \ll T^{s/s+2}$. Moreover, for $s = 2$ the condition $d \ll \sqrt{T}$ cannot be relaxed since for any function f with $\log f$ Lipschitz one can achieve a $\tilde{O}(\sqrt{T})$ upper bounds independent of dimension d, see [35, Example 2].

The proof of Theorem 5.10 is based on the following technical lemma.

Lemma 5.11. The following inequality holds, for $r > 0$:

$$\sum_{\mathbf{y}\in\{0,1\}^{T/d}} \sup_{w\in[c_1-c_2d^{-r},c_1+c_2d^{-r}]} P(\mathbf{y} \mid w) \geq \Omega(\sqrt{T/d^{2r+1}}), \tag{5.21}$$

where $P(\mathbf{y} \mid w) = w^k(1-w)^{T/d-k}$ with k being the number of 1s in $\mathbf{y}$.

Proof. By Stirling approximation, for all $k \in [T/d]$, there exists a constant $C \in \mathbb{R}^+$ such that

$$\begin{aligned} B(k, T/d) &\overset{\text{def}}{=} \binom{T/d}{k} \left(\frac{k}{T/d}\right)^k \left(1 - \frac{k}{T/d}\right)^{T/d-k} \\ &\geq C\sqrt{\frac{T/d}{k(T/d-k)}}. \end{aligned}$$

Since $P(\mathbf{y} \mid w)$ achieves maximum at $w = k \cdot d/T$, we have

$$\sum_{\mathbf{y}\in\{0,1\}^{T/d}} \sup_{w\in[c_1-c_2d^{-r},c_1+c_2d^{-r}]} p(\mathbf{y} \mid w) \geq \sum_{k=c_1T/d-c_2T/d^{r+1}}^{c_1T/d+c_2T/d^{r+1}} B(k, T/d).$$

Therefore, for each k in the above summation, we have that

$$B(k, T/d) \geq \frac{1}{\sqrt{k(T/d-k)}} \geq \sqrt{(c_1 + c_2d^{-r})(1 - c_1 - c_2d^{-r})d/T}.$$

Therefore, the LHS of (5.21) is lower bounded by

$$C\sqrt{(c_1 + c_2d^{-r})(1 - c_1 - c_2d^{-r})}\sqrt{\frac{T}{d}\frac{2c_2}{d^r}} = \Omega(\sqrt{T/d^{2r+1}})$$

for sufficient large d. □

Proof of Theorem 5.10. Now we are ready to prove Theorem 5.10. We choose a particular $\mathbf{x}^T$: We split the $\mathbf{x}^T$ into d blocks each with length of T/d. With that, the ith part of the inputs and the outputs are denoted by $\mathbf{x}^{(i)} = (\mathbf{x}_{(T/d)\cdot(i-1)+1}, \cdots, \mathbf{x}_{(T/d)\cdot i})$ and $\mathbf{y}^{(i)} = (y_{(T/d)\cdot(i-1)+1}, \cdots, y_{(T/d)\cdot i})$, respectively. For any any $\mathbf{x}_t$ we set $\mathbf{x}^{(i)} = \mathbf{e}_i$ where $\mathbf{e}_i$ is the standard d base of $\mathbb{R}^d$ with 1 in position i and 0s otherwise. Note that, with this choice of $\mathbf{x}_t$s, we have $\langle \mathbf{w}, \mathbf{x}_t \rangle = w_i$, where w_i is the ith coordinate of $\mathbf{w}$ and $\mathbf{x}_t \in \mathbf{x}^{(i)}$.

We will lower bound $r_T^*(\mathcal{H}_f \mid \mathbf{x}^T)$, which will automatically give a lower bound on $\mathsf{reg}_T(\mathcal{H}_f)$. We only need to compute the following Shtarkov sum

$$S_T(\mathcal{H}_f|\mathbf{x}^T) = \sum_{y^T\in\{0,1\}^T} \sup_{\mathbf{w}\in\mathcal{B}_s^d(1)} \prod_{i=1}^d P_f(\mathbf{y}^{(i)}|w_i), \tag{5.22}$$

where $P_f(\mathbf{y}^{(i)}|w_i) = f(w_i)^{k_i}(1-f(w_i))^{T/d-k_i}$ with k_i being the number of 1s in $\mathbf{y}^{(i)}$. We observe

$$\begin{aligned} S_T(\mathcal{H}_f|\mathbf{x}^T) &\geq \sum_{y^T\in\{0,1\}^T} \prod_{i=1}^d \sup_{w_i\in[-d^{-1/s},d^{-1/s}]} P_f(\mathbf{y}^{(i)}|w_i) \\ &= \prod_{i=1}^d \sum_{\mathbf{y}^{(i)}\in\{0,1\}^{T/d}} \sup_{w_i\in[-d^{-1/s},d^{-1/s}]} P_f(\mathbf{y}^{(i)}|w_i) \\ &= \Big(\sum_{\mathbf{y}\in\{0,1\}^{T/d}} \sup_{w\in[-d^{-1/s},d^{-1/s}]} P_f(\mathbf{y}|w)\Big)^d \\ &\geq \left(\sum_{\mathbf{y}\in\{T/d\}} \sup_{w\in[c_1-c_2d^{-1/s},c_1+c_2d^{-1/s}]} P(\mathbf{y}\mid w)\right)^d \end{aligned}$$

where $P(\mathbf{y} \mid w)$ is as in Lemma 5.11 and the last inequality holds since $[c_1 - c_2d^{-1/s}, c_1 + c_2d^{-1/s}] \subset f([d^{-1/s}, d^{-1/s}])$ by the assumption. Now, Lemma 5.11 implies that

$$S_T(\mathcal{H}_f \mid \mathbf{x}^T) \geq c^d \left(\frac{T}{d^{(s+2)/s}}\right)^{d/2},$$

where c is some absolute constant that is independent of d, T. We conclude

$$\mathsf{reg}_T(\mathcal{H}_f) \geq \mathsf{reg}_T^*(\mathcal{H}_f) \geq \log S_T(\mathcal{H}_f|\mathbf{x}^T) \geq \frac{d}{2}\log\left(\frac{T}{d^{(s+2)/s}}\right) - O(d)$$

which completes the proof. □

5.3.2 Tight Constants Beyond Bounded Hessian

In the preceding sections, we analyzed the Bayesian predictor under a *uniform* prior—equivalently, a uniform ϵ-cover—on the parameter space.

In this section, we show that for certain hypothesis classes, strictly better regret bounds can be achieved by averaging over a carefully chosen *non-uniform* cover. This phenomenon is well known in information theory (cf. [42]) for the fixed-design regret. However, for the sequential case, the situation is more complicated, since we need to deal with the $\mathbf{x}^T$ without knowing it in advance.

We start with the following simple mixture class of two hypotheses :

Lemma 5.12. Let h_1, h_2 be arbitrary functions map $\mathcal{X} \to [0,1]$. We define a function class $\mathcal{H} = \{h_w = wh_1 + (1-w)h_2 : w \in [0,1]\}$. Then

$$\mathsf{reg}_T(\mathcal{H}) \leq \frac{1}{2}\log T + \log\log T + O(1).$$

Proof. For any given $\mathbf{x}^T, y^T$, we denote $\tilde{h}_i(\mathbf{x}_t) = (1-y_t)(1-h_i(\mathbf{x}_t)) + y_t h_i(\mathbf{x}_t)$ for $i \in \{1,2\}$. We also denote $\tilde{h}_w(\mathbf{x}_t) = w\tilde{h}_1(\mathbf{x}_t) + (1-w)\tilde{h}_2(\mathbf{x}_t)$. The log-loss over $\mathbf{x}^T, y^T$ against h_w is

$$\sum_{t=1}^{T} \ell(h_w(\mathbf{x}_t), y_t) = \sum_{t=1}^{T} -\log \tilde{h}_w(\mathbf{x}_t).$$

Let $g_t(w) = -\log \tilde{h}_w(\mathbf{x}_t)$, we have

$$g_t''(w) = \frac{(\tilde{h}_1(\mathbf{x}_t) - \tilde{h}_2(\mathbf{x}_t))^2}{\tilde{h}_w(\mathbf{x}_t)^2}.$$

We now consider two cases:

1. If $\tilde{h}_1(\mathbf{x}_t) \geq \tilde{h}_2(\mathbf{x}_t)$, we have (using elementary algebra):

$$|g_t''(w)| = \frac{1}{(w + \tilde{h}_2(\mathbf{x}_t)/(\tilde{h}_1(\mathbf{x}_t) - \tilde{h}_2(\mathbf{x}_t)))^2} \leq \frac{1}{w^2}.$$

2 If $\tilde{h}_1(\mathbf{x}_t) \leq \tilde{h}_2(\mathbf{x}_t)$, we have (using elementary algebra):

$$|g_t''(w)| = \frac{1}{(1 - w + \tilde{h}_1(\mathbf{x}_t)/(\tilde{h}_2(\mathbf{x}_t) - \tilde{h}_1(\mathbf{x}_t)))^2} \leq \frac{1}{(1-w)^2}.$$

Let E be a *non-uniform* cover that is constructed as follows. Let $e_0 = \frac{1}{T}$, we define recursively

$$e_n = \left(1 + \sqrt{\frac{1}{T}}\right) e_{n-1}.$$

Let N be the maximum number for which $e_N \leq \frac{1}{2}$. It is easy to verify that $N \leq O(\sqrt{T}\log T)$. We define $E = \{e_n, 1-e_n : n \in \{0, \cdots, N\}\} \cup \{0,1\}$. By construction of E, we have for any $w \in [1/T, 1-1/T]$ there exists $e \in E$ such that

$$|e-w| \leq \min\{w\sqrt{1/T}, (1-w)\sqrt{1/T}\}.$$

Let $w^* \in [0,1]$ be the point that achieves the minimal of $g(w) = \sum_{t=1}^{T} g_t(w)$. We have either $w^* = 0, 1$ or $g'(w^*) = 0$. For the latter case, we have by Taylor expansion, for all e (where w' is convex combination of e, w^*):

$$g(e) = g(w^*) + g''(w')(e-w^*)^2.$$

Taking $e \in E$ be such that $|e-w^*| \leq \min\{w^*\sqrt{1/T}, (1-w^*)\sqrt{1/T}\}$ (assume for now $w^* \in [1/T, 1-1/T]$) and using the property for $g''(w)$ above, we have

$$g(e) \leq g(w^*) + O(1).$$

By applying the Bayesian algorithm over E with uniform prior, we obtain the following regret bound

$$\mathsf{reg}_T(\mathcal{H}) \leq \log|E| + O(1) = \frac{1}{2}\log T + \log\log T + O(1).$$

Finally, we observe that the case for $w^* \leq 1/T$ or $\geq 1-1/T$ does not affect the result, since by taking $e = 1/T$ or $1-1/T$ (respectively) guarantees that $g(e) \leq g(w^*) + O(1)$. □

Note that a uniform cover as in Theorem 5.5 would only give a $\log T$ term here, not the improved $\frac{1}{2}\log T$ rate.

Equipped with Lemma 5.12, we prove the following general result:

Theorem 5.13. Let $\{h_1, \ldots, h_s\}$ be any s functions and define

$$\mathcal{H} = \left\{\sum_{j=1}^{s} w_j h_j \;:\; w_j \in [0,1], \sum_{j=1}^{s} w_j = 1\right\}.$$

Then

$$\mathsf{reg}_T(\mathcal{H}) \leq \frac{s-1}{2}\log T + (s-1)\log\log T + O(s-1).$$

Proof. We define a recursive representation of mixture functions:

$$G(h^{j+1}, w^j) = \begin{cases} (1 - w_j)h_{j+1} + w_j G(h^{j-1}, w^{j-1}), & \text{if } j \geq 2, \\ w_1 h_1 + (1 - w_1)h_2, & \text{if } j = 1. \end{cases}$$

Then the hypothesis class can be written as

$$\mathcal{H} = \left\{ G(h^s, w^{s-1}) \ : \ w_j \in [0, 1] \text{ for all } j \right\}.$$

Let E be the covering set from Lemma 5.12. Define the approximate hypothesis class

$$\widetilde{\mathcal{H}} = \left\{ G(h^s, e^{s-1}) \ : \ e_j \in E \text{ for all } j \right\}.$$

Since $|E| = O(\sqrt{T} \log T)$, we have

$$\log |\widetilde{\mathcal{H}}| \leq \frac{s-1}{2} \log T + (s-1) \log \log T.$$

Let $L(h) = \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)$ be the cumulative log-loss. We will prove by induction on s that

$$\inf_{\tilde{h} \in \widetilde{\mathcal{H}}} L(\tilde{h}) \leq \inf_{h \in \mathcal{H}} L(h) + O(s-1). \tag{5.23}$$

The base case $s = 2$ is exactly Lemma 5.12. Assume the result holds for $s - 1$. Let w^* be the sequence of weights minimizing $L(h)$ over $\mathcal{H}$, and choose $e_{s-1} \in E$ such that

$$L(G(h^s, w_1^*, \ldots, w_{s-2}^*, e_{s-1})) \leq L(G(h^s, w^*)) + O(1).$$

Note that this function can be rewritten as

$$G(h^s, w_1^*, \ldots, w_{s-2}^*, e_{s-1}) = G(\tilde{h}^{s-1}, w_1^*, \ldots, w_{s-2}^*),$$

where each $\tilde{h}_j = (1 - e_{s-1})h_s + e_{s-1}h_j$. By the induction hypothesis, there exist $e_1, \ldots, e_{s-2} \in E$ such that

$$L(G(\tilde{h}^{s-1}, e_1, \ldots, e_{s-2})) \leq L(G(\tilde{h}^{s-1}, w_1^*, \ldots, w_{s-2}^*)) + O(s-2).$$

Finally, observe that

$$L(G(\tilde{h}^{s-1}, e_1, \ldots, e_{s-2})) = L(G(h^s, e_1, \ldots, e_{s-1})).$$

This completes the proof of inequality (5.23).

Applying the Bayesian aggregation over $\widetilde{\mathcal{H}}$ with a uniform prior yields the stated regret bound. □

Corollary 5.14. Let $f = \langle \mathbf{w}, \mathbf{x} \rangle$ with $\mathbf{w} \in \mathcal{B}_1^d(1) \cap [0,1]^d$ and $\mathbf{x} \in [0,1]^d$. Then we have

$$\mathsf{reg}_T(\mathcal{H}_f) \leq \frac{d}{2} \log T + d \log \log T + O(d).$$

Proof. Note that $\mathcal{H}_f$ is a convex combination of $\langle \mathbf{e}_j, \mathbf{x} \rangle$ with $j \in [d]$ and the all-zero valued function, where $\mathbf{e}_j$ is the standard base of $\mathbb{R}^d$ with value 1 at position j and zeros elsewhere. The result follows by Theorem 5.13 by taking $s = d + 1$. □

Remark 5.3. Note that the function $\log f$ in Corollary 5.14 is neither Lipschitz nor has bounded Hessian, yet we can still achieve a $\frac{1}{2}$ leading constant. It is worth noting that using the result in [9, Chapter 9.10] and the expected majorizing martingale characterization for r_T under Lipschitz loss as established in [27], one can show that the $\frac{1}{2}$ constant can be also achieved if we only assume $\log f$ is Lipschitz (i.e., no requirement on Hessian). This does not apply to the case in Corollary 5.14, since in our case the function $\log f$ is not even Lipschitz.

5.4 Large Growth.

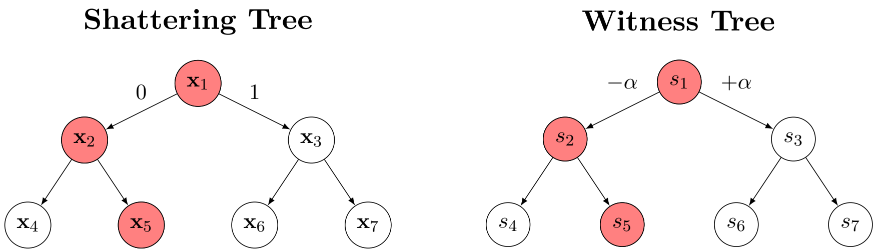
We now introduce results for the setting where the dimension d grows faster than the time horizon T. This is achieved by bounding the size of the *global* sequential covering (see Definition 5.1) using the notion of the sequential fat-shattering dimension.

We begin by introducing the definition of the sequential fat-shattering dimension from [27] (see also Section 2.2).

Definition 5.2 (Sequential Fat-Shattering). : Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$. We say a $\mathcal{X}$-valued binary tree $\tau : \bigcup_{i=0}^d \{0,1\}^i \to \mathcal{X}$ is α-fat-shattered by $\mathcal{H}$, witnessed by a $\mathbb{R}$-valued binary tree $s : \bigcup_{i=0}^d \{0,1\}^i \to \mathbb{R}$, if for any $\epsilon^d \in \{0,1\}^d$, there exists $h \in \mathcal{H}$ such that:

1. If $\epsilon_t = 0$, then $h(\tau(\epsilon^{t-1})) \leq s(\epsilon^{t-1}) - \alpha$;
2. If $\epsilon_t = 1$, then $h(\tau(\epsilon^{t-1})) \geq s(\epsilon^{t-1}) + \alpha$.

Example 5.3. We illustrate the sequential α-fat-shattering dimension with an example below.


Shattering Tree

Witness Tree

Consider a path $\{0, 1\}$, the α-fat shattering ensures $\exists h \in \mathcal{H}$ such that: (1) $h(\mathbf{x}_1) \leq s_1 - \alpha$ and (2) $h(\mathbf{x}_2) \geq s_2 + \alpha$.

Definition 5.3. The sequential α-fat shattering dimension of $\mathcal{H}$ is defined to be the maximum number $d(\alpha)$ such that $\mathcal{H}$ α-fat shatters a tree τ of depth $d := d(\alpha)$. We also write $\mathsf{sfat}_\alpha(\mathcal{H}) =: d(\alpha)$ if we want to emphasize that the fat dimmension depdends on $\mathcal{H}$.

In the below lemma, we present an upper bound for the cardinally of the global covering set with algorithmically constructed cover set $\mathcal{G}_\alpha$, see [27, Section 6.1] for proof.

Lemma 5.15. Let $\mathcal{H}$ be any class $\mathcal{X} \to [0, 1]$ and $d(\alpha)$ be the sequential α-fat shattering dimension of $\mathcal{H}$. Then there exists a global sequential α-covering set $\mathcal{G}_\alpha$ of $\mathcal{H}$ as in Definition 5.3 such that

$$|\mathcal{G}_\alpha| \leq \sum_{t=0}^{d(\alpha/2)} \binom{T}{t} \left\lceil \frac{1}{\alpha} \right\rceil^t \leq \left\lceil \frac{T}{\alpha} \right\rceil^{d(\alpha/2)+1}. \tag{5.24}$$

Example 5.4. By [27] we know that the sequential α-fat shattering number of linear functions $f(\mathbf{w}, \mathbf{x}) = |\langle \mathbf{w}, \mathbf{x} \rangle|$ with $\mathbf{w}, \mathbf{x} \in \mathcal{B}_2^d(1)$ is of order $\tilde{O}(\alpha^{-2})$ where in $\tilde{O}$ we hide a polylog factor. Lemma 5.15 implies that the global sequential α-covering number is upper bounded by

$$\left\lceil \frac{(T)}{(\alpha)} \right\rceil^{\tilde{O}(\alpha^{-2})}.$$

By Theorem 5.3, we have

$$\mathsf{reg}_T(\mathcal{H}_f) \leq \inf_{0<\alpha<1} \left\{ 2\alpha T + \tilde{O}\left(\frac{1}{\alpha^2}\right) \right\} \leq \tilde{O}(T^{2/3}),$$

by taking $\alpha = T^{-1/3}$. This bound is *independent* of the data dimension d.

Remark 5.4. Observe that for any class $\mathcal{H}$ with sequential fat-shattering dimension of order α^{-s} one can achieve a regret upper bound of order $\tilde{O}(T^{s/s+1})$ by Theorem 5.3. We refer to [27], [43] for the estimations of the sequential fat-shattering dimension of a variety of classes.

We now present the following general lower bound:

Theorem 5.16. For any $s \geq 1$, we define

$$\mathcal{D}_s = \left\{ \mathbf{p} \in [0,1]^T : \sum_{t=1}^{T} p_t^s \leq 1 \right\}.$$

We can view the vectors in $\mathcal{D}_s$ as functions mapping $[T] \to [0,1]$. Then

$$\mathsf{reg}_T(\mathcal{D}_s) \geq \mathsf{reg}_T^*(\mathcal{D}_s) \geq \Omega(T^{s/s+1}). \tag{5.25}$$

Proof. By Theorem 4.3, it is sufficient to compute the Shtarkov sum. For any $y^T \in \{0,1\}^T$ with k 1s, we claim that

$$\sup_{\mathbf{p} \in \mathcal{D}_s} p(y^T) = \frac{1}{k^{k/s}},$$

where

$$p(y^T) = \prod_{t=1}^{T} p_t^{y_t} (1 - p_t)^{1 - y_t}.$$

To see this, we use a *perturbation* argument. Denote I to be the positions in y^T that takes value 1 and let $|I| = k$. For any $\mathbf{p}$ such that $p(y^T)$ is maximum, we must have $p_j = 0$ for all $j \notin I$. Suppose otherwise, we then can move some probability mass on p_j to some $p_i < 1$ with $i \in I$, which will increase the value of $p(y^T)$, thus a contradiction. Now, we need to show that

$$\prod_{i \in I} p_i \leq \frac{1}{k^{k/s}},$$

this follows easily by AM-GM (i.e., arithmetic mean vs geometric mean) inequality since $\sum_{i \in I} p_i^s \leq 1$ and it is an equality when $p_i = \frac{1}{k^{1/s}}$ for all $i \in I$. Now, the Shtarkov sum can be written as

$$\sum_{k=0}^{T} \binom{T}{k} \frac{1}{k^{k/s}}. \tag{5.26}$$

To find a lower bound, we only need to estimate the maximum term in the summation. We have

$$\max_k \binom{T}{k} \frac{1}{k^{k/s}} \geq \max_k \frac{T^k}{k^{(1+1/s)k}} \geq e^{\frac{s+1}{s \cdot e} T^{s/s+1}},$$

where the last inequality follows by taking $k = \frac{1}{e} T^{s/s+1}$, and we also use the fact that

$$\binom{T}{k} \geq \frac{T^k}{k^k}.$$

Therefore, we have

$$\mathsf{reg}^*_T(\mathcal{D}_s) \geq \frac{s+1}{s \cdot e} T^{s/s+1} = \Omega(T^{s/s+1})$$

which completes the proof. □

To see why Theorem 5.16 implies a lower bound for $f(\mathbf{w}, \mathbf{x}) = |\langle \mathbf{w}, \mathbf{x} \rangle|$ with $d \geq T$, as in Example 5.4, we take $\mathbf{w}, \mathbf{x} \in \mathcal{B}_2^T(1)$ (i.e., with $d = T$) and define $\mathbf{x}_t = \mathbf{e}_t$ with $\mathbf{e}_t$ being the standard base of $\mathbb{R}^T$ that takes value 1 at position t and zeros otherwise. Note that the functions of $\mathcal{H}_f$ with $f(\mathbf{w}, \mathbf{x}) = |\langle \mathbf{w}, \mathbf{x} \rangle|$ restricted on $\mathbf{x}^T$ is exactly $\mathcal{D}_2$. Then

$$\mathsf{reg}_T(\mathcal{H}_f) \geq \mathsf{reg}^*_T(\mathcal{H}_f) \geq \mathsf{reg}^*_T(\mathcal{D}_2) \geq \Omega(T^{2/3})$$

and this is a matching lower bound of Example 5.4. Note that, it is proved in [44] that for function $f(\mathbf{w}, \mathbf{x}) = \frac{\langle \mathbf{w}, \mathbf{x} \rangle + 1}{2}$, one can achieve the regret of form $\tilde{O}(\sqrt{T})$[1]. Example 5.4 implies that the generalized linear functions of form $f(\langle \mathbf{w}, \mathbf{x} \rangle)$ can have different regrets with polynomial gap even with a simple shift on the value (though they have the same covering number!).

5.4.1 Tighter Lower Bounds for Generalized Linear Functions.

We now provide additional lower bounds through the fixed design regret. As we have demonstrated in Theorems 5.10 and 5.16 that lower bounds can be derived by selecting some appropriate $\mathbf{x}^T$ that maximizes $r^*_T(\mathcal{H} \mid \mathbf{x}^T)$, where we only choose $\mathbf{x}^T$ to be some combinations of the

[1] A $\tilde{\Omega}(\sqrt{T})$ lower bound for $d \geq \sqrt{T}$ can be derived from Theorem 5.10, recovering [44, Lemma 8].

standard base of $\mathbb{R}^d$. In this section, we present examples where a more sophisticated selection of $\mathbf{x}^T$ leads to better lower bounds.

The following theorem shows that the leading constant 1 in Theorem 5.6 holds even for generalized *linear* functions with Lipschitz transform function:

Theorem 5.17. For any $s \geq 1$ and $d \ll T^{s/(s+1)}/\log T$, there exists a 1-Lipschitz function $f : [0,1] \to [0,1]$ such that for hypothesis class:

$$\mathcal{H}_f = \{f(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathcal{B}_s^d(1), \mathbf{x} \in \mathcal{B}_{s/(s-1)}^d(1)\},$$

we have

$$\mathsf{reg}_T^*(\mathcal{H}_f) \geq d \log \left(\frac{T}{d^{(s+1)/s}} \right) - O(d \log \log T).$$

The proof of Theorem 5.17 is based on the following technical lemma, which can be viewed as a correlated version of Lemma 5.8.

Lemma 5.18. There exists a binary sequence $b_1, b_2, \cdots, b_{2T} \in \{0,1\}$ such that for any $i \neq j \in [T]$ we have

$$\sum_{t=1}^{T} 1\{b_{i+t} \neq b_{j+t}\} \geq \frac{T}{8}.$$

Proof. We use the probabilistic method to construct sequence b^{2T}. To do so, we select B^{2T} uniformly at random from $\{0,1\}^{2T}$ and show that the event of the lemma happens with positive probability. For any $i < j \in [T]$, we construct an *i.i.d.* sequence $X_0, \cdots, X_{T/2-1}$ with uniform distribution over $\{0,1\}$ such that:

$$\sum_{t=0}^{T-1} 1\{B_{i+t} \neq B_{j+t}\} \geq \sum_{t=0}^{T/2-1} X_t.$$

To do so, we maintain an index set I and a set $\mathbf{X}$ of random variables, initial $I, \mathbf{X} = \emptyset$ and $t = 0$. For each $t \in [T-1]$, if the index $i + t \in I$, we remove $i + t$ from I and continue to $t + 1$; else, we add the indicator $1\{B_{i+t} \neq B_{j+t}\}$ to $\mathbf{X}$ and $j + t$ to I and continue to $t + 1$. Clearly all the random variables in $\mathbf{X}$ are mutually independent and distributed uniformly over $\{0,1\}$, since we add the indicator to $\mathbf{X}$ only when there is no overlap on the indexes and B^{2T} are i.i.d. random variables (notice that if $i+t \notin I$ then $j+t$ also does not appear in the previous indexes, since $i<j$). We claim that $|I| \geq T/2$. This follows from the fact that the conflict period must not be more than the non-conflict period. The lemma now follows by Chernoff bound, to show that $\sum_{t=0}^{T/2-1} X_t$ tightly concentrates on $T/4$ and a union bound on all pairs (i,j). □

Proof of Theorem 5.17. We partition the $\mathbf{x}^T$ into d parts each of length T/d (assume w.l.o.g. that d divides T [2]), and denote $K = T/d$. For the ith part in the partition, we will select the K features of form $x_1\mathbf{e}_i, \cdots, x_K\mathbf{e}_i$, where $x_j \in [0,1]$ and $\mathbf{e}_i$ is the standard base of $\mathbb{R}^d$ with position i being 1 and zeros elsewhere. Note that all of the parts in the partition have the *same* scalars x_js but different vector $\mathbf{e}_i$. Clearly, the value of $f(\langle \mathbf{w}, \mathbf{x}\rangle)$ with $\mathbf{x}$ in the ith part depends only on w_i, where w_i is the ith coordinate of $\mathbf{w}$. Since $\mathbf{w} \in \mathcal{B}_s^d(1)$, we may assume that each w_i takes values in the full range $[-d^{-1/s}, d^{-1/s}]$ and selected independently. Therefore, it is sufficient to construct a function $f : [0,1] \to [0,1]$ such that the class

$$\mathcal{Q} = \{f(w \cdot x) : w \in [-d^{-1/s}, d^{-1/s}],\ x \in [0,1]\}$$

has large regret.

Let $\gamma \in [0,1]$ be a small real number depending on K that is to be determined later. For any number $t \leq 1/\gamma$, we define $z_t = (1-\gamma)^{t-1}$. Let $\mathcal{Z} = \{z_t : t \leq 1/\gamma\}$. We now select the $x_1, \cdots, x_K$ to be elements in $\mathcal{Z}$ such that each repeats at least $\lfloor \gamma K \rfloor$ times (there can be some elements in $\mathcal{Z}$ that repeat more than $\lfloor \gamma K \rfloor$ times). We also define a sequence z'_t with $t \leq 2/\gamma$ such that $z'_t = d^{-1/s}(1-\gamma)^{t-1}$. We observe that $|z'_t - z'_{t+1}| \geq d^{-1/s} \cdot e^{-3} \cdot \gamma$ for all $t \leq 2/\gamma$, since $(1-\gamma)^{2/\gamma} \sim e^{-2}$ for γ small enough. For any number $t \leq 1/\gamma$, we define $w_t = d^{-1/s}(1-\gamma)^{t-1} \in [-d^{-1/s}, d^{-1/s}]$.

Let $M = \lfloor 1/\gamma \rfloor$ and b^{2M} be the binary sequence as in Lemma 5.18 with $T = M$. Define f to be the function over z'_ts with $t \leq 2/\gamma$, such that $f(z'_t) = 0$ if $b_t = 0$ and $f(z'_t) = d^{-1/s} \cdot e^{-3} \cdot \gamma$ if $b_t = 1$. Clearly, f is 1-Lipschitz over the z'_ts. By Lemma 5.18, for any w_i and w_j as defined above with $i \neq j \leq 1/\gamma$, there are at least $M/16$ positions $t \leq 1/\gamma$ such

[2] Otherwise, we round T to be some $T' \leq T$ that is divisible by d, this only incurs a $O(d)$ regret loss.

that $f(w_i \cdot z_t)$ take all 0s and $f(w_j \cdot z_t)$ take all values $d^{-1/s} \cdot e^{-3} \cdot \gamma$ (or vice versa). Note that, any of the w_i defines a product distribution over $\{0,1\}^K$ such that each coordinate $t \leq K$ is an independent Bernoulli random variable with parameter $f(w_i \cdot x_t)$. Since each z_t appears least $\lfloor \gamma K \rfloor$ times in the x_ts, we have the probability of wrongly identifying the source (of the product distribution) with parameters w_i and w_j being upper bounded by

$$(1 - d^{-1/s} \cdot e^{-3} \cdot \gamma)^{(M/16) \cdot \lfloor \gamma K \rfloor} \leq e^{-d^{-1/s} \cdot e^{-3} \cdot \gamma (K - 1/\gamma - \gamma K)/16}.$$

Taking $\gamma = 64 \cdot e^3 \cdot (d^{1/s} \cdot \log K)/K$ and applying a union bound on all the pairs w_i, w_j (there are at most K^2 such pairs), we can make error probability upper bounded by $1/2$ for sufficiently large K. We now verify that $\gamma \to 0$ as $K \to \infty$, this is guaranteed by our assumption that $d \ll T^{s/(s+1)}/\log T$. By Lemma 5.7, we have

$$\mathsf{reg}^*_T(\mathcal{Q}) \geq \log(K/(d^{1/s} \log K)) - O(1).$$

Since $K = T/d$ and each of the parts of the partition of $\mathbf{x}^T$ are independent, we have

$$\mathsf{reg}^*_T(\mathcal{H}) \geq d \cdot \mathsf{reg}^*_T(\mathcal{Q}) \geq d \log\left(\frac{T}{d^{(s+1)/s}}\right) - O(d \log\log T).$$

Finally, by [39, Theorem 1], we can extend f to the whole set $[0,1]$ while keep the Lipshitz condition. □

Remark 5.5. Note that the condition $d \ll T^{s/(s+1)}$ cannot be relaxed (upto poly-log factors in general) by Example 5.4. Our results in Theorem 5.17, Theorem 5.5, and Example 5.4 imply an interesting threshold phenomenon (for the generalized linear functions with worst case Lipschitz transform function), i.e., when $d \ll T^{(s/s+1)-\epsilon}$ with $\epsilon > 0$, the regret grows as $\Theta(d \log T)$, while for $d \gg T^{s/s+1}$, the regret grows as $\tilde{O}(T^{s/s+1})$. Moreover the leading constant is exactly 1 for the $d \log T$ term if d is sub-polynomial w.r.t. T (e.g., $d = e^{\sqrt{\log T}}$).

5.4.2 Additional Large Classes.

In this section we consider a general (including non-parametric) class $\mathcal{H} = \{h \in [0,1]^{\mathcal{B}^d_s(1)} : \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{B}^d_s(1),\ |h(\mathbf{x}_1) - h(\mathbf{x}_2)| \leq ||x_1 - x_2||_s\}$

of all Lipschitz functions mapping from a unit ℓ_s ball to $[0, 1]$. We also assume that the Lipschitz condition is under the norm ℓ_s as well [3]. The following theorem establishes a lower bound for such function classes (a matching upper bound – up to poly-log factors – can be derived using Theorem 5.3 and the classical uniform covering numbers as in [45, Lemma 5.2], see also [36]).

Theorem 5.19. For any $d, T, s \geq 1$ such that $d \ll \log T$, we let $\mathcal{H} = \{h \in [0,1]^{\mathcal{B}_s^d(1)} : \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{B}_s^d(1),\ |h(\mathbf{x}_1) - h(\mathbf{x}_2)| \leq ||x_1 - x_2||_s\}$. Then

$$\mathsf{reg}_T^*(\mathcal{H}) \geq \Omega(T^{d/d+1}).$$

Moreover, if $d \geq \Omega(\log T)$, then $\mathsf{reg}_T^*(\mathcal{H}) \geq \Omega(T)$.

Proof. Let M be a maximum packing of $\mathcal{B}_s^d(1)$ under ℓ_s norm with radius T^{-r}, where $r = \frac{1}{d+1}$. Standard volume argument yields that:

$$|M| \geq T^{dr}.$$

Note that the packing number is independent of s, since we are packing a ℓ_s ball under the same ℓ_s norm. Assume w.l.o.g. that $|M| = \lfloor T^{dr} \rfloor$. We now select $\mathbf{x}^T$ to be all the elements in M such that each of them is repeated at least $\lfloor T^{1-dr} \rfloor$ times in $\mathbf{x}^T$. We now select a class $\mathcal{F} \subset \mathcal{H}$ of functions that map $M \to \{0, T^{-r}\}$ such that any two functions differ by at least $\lfloor T^{dr}/4 \rfloor$ elements in M. By Lemma 5.8, we know that there are at least $2^{T^{dr}/8}$ such functions. By removing some functions, we may assume that there are $\lfloor 2^{T^{dr}/16} \rfloor$ functions in $\mathcal{F}$.

$$(1 - T^{-r})^{\lfloor T^{1-dr} \rfloor \cdot \lfloor T^{dr}/4 \rfloor / 2} \leq e^{-T^{-r} \cdot \lfloor T^{1-dr} \rfloor \cdot \lfloor T^{dr}/4 \rfloor / 2} \leq e^{-(1/8 - o(1)) T^{1-r}},$$

where we used the fact that each element in M repeats at least $\lfloor T^{1-dr} \rfloor$ times in $\mathbf{x}^T$ and $d \ll \log T$. In order for a union bound over all pairs in $\mathcal{F}$ to work (there are at most $2^{T^{dr}/8}$ such pairs), it is sufficient to have (since $2 < e$):

$$T^{dr}/8 \leq T^{1-r}/8.$$

This holds when $r = \frac{1}{1+d}$. Therefore, by union bound, we have for sufficient large T, one can identify the sources in $\mathcal{F}$ with error probability

[3] Note that our technique can be generalized to cases when the Lipschitz conditions are defined in a different norm.

upper bounded by $\frac{1}{2}$. Invoking Lemma 5.7, we conclude that

$$\mathsf{reg}^*_T(\mathcal{H}) \geq \mathsf{reg}^*_T(\mathcal{F}) \geq \frac{1}{16\log 2} T^{d/d+1} - \log(2).$$

By definition, the functions in $\mathcal{F}$ restricted on M are 1-Lipschitz under ℓ_s norm. By [39, Theorem 1], we can extend them to the whole set of $\mathcal{B}^d_s(1)$ while keeping the Lipschitz property. The last part of the theorem follows from the fact that when $d = c\log T$ for any constant c, we have $T^{-1/(1+d)} \sim e^{-1/c}$. We can therefore let c be small enough so that $e^{-1/c} \leq \frac{1}{16}$, which will allow our argument above to go thorough (by reducing the size of $\mathcal{F}$ by a constant on the exponent) and result in a regret bound $\Omega(T^{d/(d+1)}) \geq \Omega(T)$. Note that regret is monotone increasing according to d, thus the result follows. □

5.5 Bibliographical Notes

In this chapter we study the sequential minimax regret for a general online regression with logarithmic loss using tools of information theory, in particular the universal source coding (lower bounds) [2]–[5], [7], [11], [13], [30] and sequential covering (upper bounds).

Most of the existing works in online regression deal with the logistic regression. We first mention the work of [46] who studied the pointwise regret of the logistic regression for the *proper* setting. Unlike the *improper* learning, studied in this paper, where feature $\mathbf{x}_t$ at time t is also available to the learner, [46] showed that the pointwise regret is $\Theta(T^{1/3})$ for $d = 1$ and $O(\sqrt{T})$ for $d > 1$. Furthermore, [40] demonstrates results that regret for logistic regression grows like $O(d\log T/d)$. This was further generalized in [35]. These results were strengthened in [37], which also provides the matching lower bounds. Precise asymptotics for the fixed design minimax regret were recently presented in [32], [38] for $d = o(T^{1/3})$.

Regret bounds under logarithmic loss for general expert class $\mathcal{H}$ was first investigated by Vovk under the framework of mixable losses [47], [48]. In particular, Vovk showed that for finite class $\mathcal{H}$, the regret growth is $\log|\mathcal{H}|$ via the *aggregating algorithm* (i.e., the Bayesian algorithm that we will discuss below). We refer the reader to [9, Chapter 3.5, 3.6] and the references therein for more results on this topic. Cesa-Bianchi and Lugosi [9] were the first to investigate log-loss under general (infinite) expert class $\mathcal{H}$ [9, Chapter 9.10, 9.11], where they derived a general upper bound using the concept of covering number and a two-stage prediction scheme. In particular, Cesa-Bianchi and Lugosi showed that for Lipschitz parametric classes with values bounded away from $\{0, 1\}$, one can achieve a regret bound of the form $d/2 \log(T/d)$. When the values are close to $\{0, 1\}$, they used a *hard* truncation approach, which gives a sub-optimal bound of the form $2d \log(T/d)$. Moreover, the approach of [9] only works for the fixed design regret (or *simulatable* in their context). In [44], the authors extended the result of [9, Chapter 9.10] to the sequential case via the machinery of sequential covering that was established in [49]. However, [44] also used the same *hard* truncation as in [9] resulting in suboptimal upper bounds. In [36], the authors obtained an upper bound similar to the upper bound presented in Theorem 5.3 via the observation that the log function is self-concordance. In particular, this allows them to resolve the tight bounds for non-parametric Lipschitz functions map $[0, 1]^s \to [0, 1]$. However, their bounds are proved *non-constructively*, i.e., the proof does not provide an algorithm that achieves such bounds. We note also a recent work in [50] which used similar idea of smoothing for controlling the unboundedness of log-loss, however, they are assumed that the features $\mathbf{x}^T$ are presented *i.i.d.*. More importantly, the results in [50] only holds for the *average case* regret.

This chapter is largely based on [51] and [52].

6

Minimax Regret with Lipschitz Losses

In this section, we present results on the minimax regret when the loss function is both bounded and Lipschitz. Furthermore, we extend our analysis beyond finite hypothesis classes $\mathcal{H}$, as considered in Theorem 3.4. The main tool used throughout this chapter is the notion of sequential covering, introduced in Definition 5.1.

6.1 Absolute Loss

We start with the absolute loss. Let $\mathcal{Y} = \{0, 1\}$ and $\hat{\mathcal{Y}} = [0, 1]$, the absolute loss function is defined as

$$\ell(\hat{y}, y) = |\hat{y} - y|.$$

Observe that $|\hat{y} - y| = \mathbb{E}_{y' \sim \mathsf{Bern}(\hat{y})}[1\{y' \neq y\}]$, i.e., it measures the *expected miss-classification* loss when sampling from a Bernoulli source of parameter $\hat{y}$.

We recall from Theorem 3.4 that for a *finite* class $\mathcal{H}$ the minimax regret of $\mathcal{H}$ under the absolute loss is upper bounded by

$$\mathsf{reg}_T(\mathcal{H}) \leq O(\sqrt{T \log |\mathcal{H}|}),$$

which is achieved by the (generalized) EWA Algorithm 3.3.

We now consider hypothesis classes $\mathcal{H}$ that are not finite and address the question: under what conditions are such classes learnable? We begin with the following example.

Example 6.1. Consider the following class of *threshold functions*:

$$\mathcal{H}^{\mathsf{thres}} := \{h_a(x) = 1\{x \geq a\} : a, x \in [0,1]\}.$$

For any learner Φ, we construct the following strategy for *Nature*:

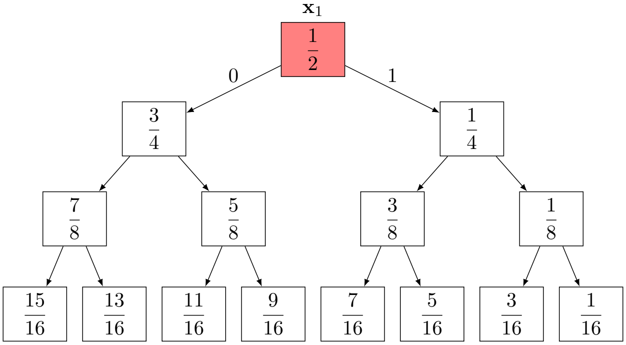
1. At every step t, we select label $y_t \in \{0,1\}$ such that $|y_t - \hat{y}_t| \geq \frac{1}{2}$.
2. Select instances $\mathbf{x}_t$ from the set of *dyadic rationals*, starting with $\mathbf{x}_1 = \frac{1}{2}$ and updating (according to learner's prediction $\hat{y}_{t-1}$) as:

$$\mathbf{x}_t = \begin{cases} \mathbf{x}_{t-1} + \frac{1}{2^t}, & \text{if } \hat{y}_{t-1} \geq 0.5, \\ \mathbf{x}_{t-1} - \frac{1}{2^t}, & \text{else.} \end{cases}$$

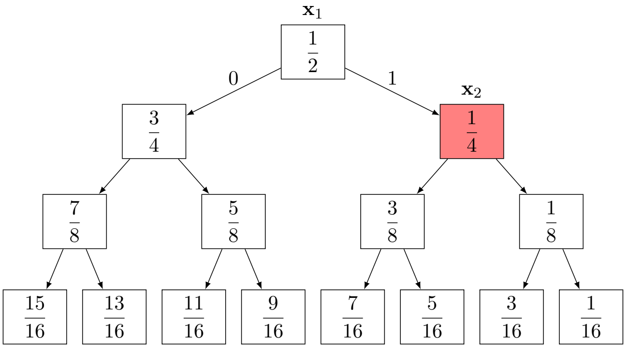
This strategy ensures that the cumulative loss incurred by the learner is at least $T/2$, while for all $t \leq T$, the threshold function $h_{\mathbf{x}_{T+1}}$ with parameter $a = \mathbf{x}_{T+1}$ satisfies $h_{\mathbf{x}_{T+1}}(\mathbf{x}_t) = y_t$—that is, it incurs zero loss. Therefore, we conclude that

$$\mathsf{reg}_T(\mathcal{H}^{\mathsf{thres}}) \geq T/2.$$

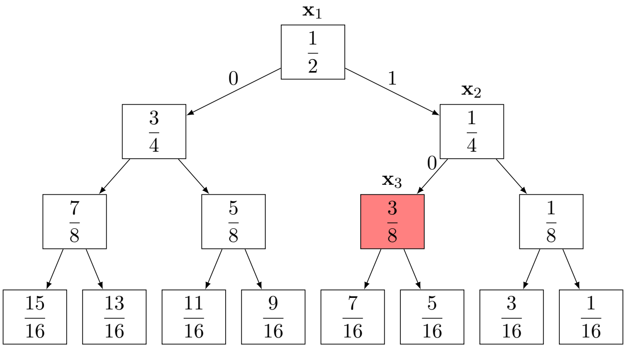
To make the argument even clearer, we illustrate the adversarial strategy using a tree-based process. Let the learner's prediction sequence be $\{0,1,1\}$. The strategy employed by Nature proceeds as follows:



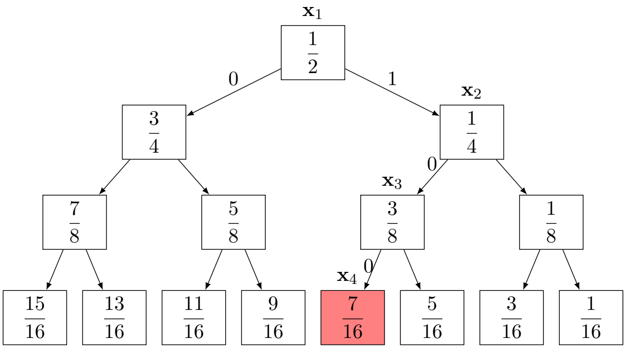
When x_2 shows up we go down the tree to see:



Upon seeing x_3 we ended up with the following:



Finally, when x_4 arrives we ended up at the following:



The function $h_{\mathbf{x}_4}(\mathbf{x}) := 1\{\mathbf{x} \geq \frac{7}{16}\}$ consistent with all true labels, but the learner errs at every step.

6.1.1 Shattering Trees and Littlestone Dimension

We have shown that even for simple threshold functions, achieving sublinear regret is not possible. This naturally raises the question: *What intrinsic structure of $\mathcal{H}$ leads to this failure?*

To address this, we introduce the concept of shattering trees, previously discussed in Section 5.4. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued hypothesis class. A $\mathcal{X}$-valued binary tree of depth d is a mapping

$$\tau : \bigcup_{i \le d} \{0,1\}^i \to \mathcal{X}.$$

We say that τ is shattered by $\mathcal{H}$ if for every $\epsilon^d \in \{0,1\}^d$, there exists $h \in \mathcal{H}$ such that

$$\forall i \le d, \quad h(\tau(\epsilon^{i-1})) = \epsilon_i.$$

Note that the tree formed by the dyadic rationals is shattered by the class $\mathcal{H}^{\mathsf{thres}}$.

We now observe a simple fact:

Lemma 6.1. For any binary-valued class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$, if there exists a $\mathcal{X}$-valued binary tree of depth d that can be shattered by $\mathcal{H}$, then:

$$\mathsf{reg}_T(\mathcal{H}) \ge \frac{1}{2} \min\{d, T\}.$$

Proof. Indeed, it is enough to select the labels opposite to learner's prediction, and the instances by following the shattering tree τ, similar to the threshold function case as discussed in Example 6.1 □

From the discussion above, it should be clear that we need some constraints on the class $\mathcal{H}$ for it to be learnable. We are now ready to introduce the Littlestone dimension.

Definition 6.1 (Littlestone Dimension). Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued hypothesis class. The *Littlestone dimension* of $\mathcal{H}$ is defined as the maximum number d such that there exists a $\mathcal{X}$-valued binary tree of depth d that can be shattered by $\mathcal{H}$.

We will denote $\mathsf{Ldim}(\mathcal{H})$ as the Littlestone dimension of $\mathcal{H}$. It is clear from our previous discussion that $\mathsf{reg}_T(\mathcal{H}) \ge \frac{1}{2}\min\{\mathsf{Ldim}(\mathcal{H}), T\}$. Therefore, Littlestone dimension forms an intrinsic barrier for the minimax regret.

Example 6.2. We now present several examples of hypothesis classes $\mathcal{H}$ and discuss their corresponding Littlestone dimensions.

1. For the threshold functions $\mathcal{H}^{\mathsf{thres}}$, we have $\mathsf{Ldim}(\mathcal{H}^{\mathsf{thres}}) = \infty$.
2. For any finite hypothesis class $\mathcal{H}$, it holds that $\mathsf{Ldim}(\mathcal{H}) \leq \log|\mathcal{H}|$.
3. Consider the class of indicator functions defined by
$$\mathcal{H}^{\mathsf{ind}} := \{h_a(x) := 1\{x = a\} : x, a \in [0,1]\}.$$
Then $\mathsf{Ldim}(\mathcal{H}^{\mathsf{ind}}) = 1$.

We have shown that the Littlestone dimension forms a natural lower bound for the minimax regret. Can we achieve an upper bound as well?

Algorithm 6.1 The Standard Optimal Algorithm (SOA)

```
Initialize $\mathcal{H}^{(0)} \leftarrow \mathcal{H}$
for each time step $t = 1, 2, \ldots$ do
    for each $y \in \{0,1\}$ do
        $\mathcal{H}_y^{(t)} \leftarrow \{h \in \mathcal{H}^{(t-1)} : h(\mathbf{x}_t) = y\}$
    end for
    Predict $\hat{y}_t \leftarrow \arg\max_{y\in\{0,1\}}\{\mathsf{Ldim}(\mathcal{H}_y^{(t)})\}$
    Receive true label $y_t$
    Update $\mathcal{H}^{(t)} \leftarrow \mathcal{H}_{y_t}^{(t)}$
end for
```

Lemma 6.2. For any data $\mathbf{x}^T, y^T$ that is realizable w.r.t. a binary-valued class $\mathcal{H}$, i.e., $\exists h^* \in \mathcal{H}$ such that $\forall t \leq T$, $h^*(\mathbf{x}_t) = y_t$, the SOA predictor enjoys the following mistake bound

$$\sum_{t=1}^{T} 1\{\hat{y}_t \neq y_t\} \leq \mathsf{Ldim}(\mathcal{H}).$$

Proof. At each round t, if the SOA predictor errs $\hat{y}_t \neq y_t$, it updates to $\mathcal{H}_{y_t}^{(t)} = \{h \in \mathcal{H}^{(t)} : h(x_t) = y_t\}$. By the definition of Littlestone dimension, $\min_{y\in\{0,1\}}\{\mathsf{Ldim}(\mathcal{H}_y^{(t-1)})\} \leq \mathsf{Ldim}(\mathcal{H}^{(t-1)}) - 1$, so each mistake decreases the dimension by at least one. Since the initial dimension is at most $\mathsf{Ldim}(\mathcal{H})$ and cannot become negative, the total number of mistakes is at most $\mathsf{Ldim}(\mathcal{H})$. □

6.1.2 Sequential Cover

Note that Lemma 6.2 holds under the realizability assumption. To extend our analysis to the agnostic setting, we recall the notion of a *sequential cover*, as introduced in Definition 5.1, and now specialize it to binary-valued hypothesis classes.

Definition 6.2 (Sequential Cover). Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary-valued class, and $\mathcal{G} \subset \{0,1\}^{\mathcal{X}^*}$ be a class mapping $\mathcal{X}^* \to \{0,1\}$. We say that the class $\mathcal{G}$ sequentially covers $\mathcal{H}$ up to step T if, for any $\mathbf{x}^T \in \mathcal{X}^T$ and $h \in \mathcal{H}$, there exists $g \in \mathcal{G}$ such that

$$\forall t \leq T,\ g(\mathbf{x}^t) = h(\mathbf{x}_t).$$

We observe that the cover happens locally, depending on any given $\mathbf{x}^T$, unlike the classical uniform cover where each h is covered by a fixed g. Sequential cover allows the covering function g to depend on $\mathbf{x}^T$ as well. Furthermore, infinite classes $\mathcal{H}$ can be sequentially covered by a finite class $\mathcal{G}$. Indeed, consider the class $\mathcal{H}^{\mathsf{ind}} := \{h_a(x) := 1\{x = a\} : x, a \in [0,1]\}$. For any $i \leq [T] \cup \{0\}$, define the sequential function:

$$g_i(\mathbf{x}^t) = \begin{cases} 1\{t \geq i \text{ and } \mathbf{x}_t = \mathbf{x}_i\}, & \text{if } i \neq 0 \\ 0, & \text{if } i = 0 \end{cases}.$$

The class $\mathcal{G} := \{g_i : i \in [T] \cup \{0\}\}$ sequentially covers $\mathcal{H}^{\mathsf{ind}}$.

The next lemma is the key to proving the final bounds on regret.

Lemma 6.3. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any binary-valued class. If there exists a predictor for $\mathcal{H}$ that achieves mistake bound err_T in the realizable case, then there exists a sequential cover $\mathcal{G}$ of $\mathcal{H}$ up to step T such that

$$\log |\mathcal{G}| \leq \log \sum_{i=0}^{\mathsf{err}_T} \binom{T}{i} \leq O(\mathsf{err}_T \cdot \log T).$$

Proof. Let Φ achieve err_T mistakes for $\mathcal{H}$ in the realizable case. For any $I \subset [T]$, we recursively define the sequential function

$$g_I(\mathbf{x}^t) = \begin{cases} \Phi(\mathbf{x}^t, g_I(\mathbf{x}^1), \cdots, g_I(\mathbf{x}^{t-1})), & \text{if } t \notin I \\ 1 - \Phi(\mathbf{x}^t, g_I(\mathbf{x}^1), \cdots, g_I(\mathbf{x}^{t-1})), & \text{if } t \in I \end{cases}.$$

The class $\mathcal{G} := \{g_I : I \subset [T],\ |I| \leq \mathsf{err}_T\}$ sequentially covers $\mathcal{H}$, since for any $\mathbf{x}^T$ and h we can pick I being the time steps where Φ errs. Finally, we have $|\mathcal{G}| \leq \sum_{i=0}^{\mathsf{err}_T} \binom{T}{i}$ by counting the size of $\{I \subset [T] : |I| \leq \mathsf{err}_T\}$. □

This leads to the main result of this section, which establishes tight minimax regret bound in the agnostic setting.

Theorem 6.4. For any binary-valued class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ with finite Littlestone dimension $\mathsf{Ldim}(\mathcal{H})$, the minimax regret of $\mathcal{H}$ satisfies

$$\Omega(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T}) \leq \mathsf{reg}_T(\mathcal{H}) \leq O(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T \log T}).$$

Proof. The upper bound is easy. From Lemma 6.2 we know that the class admits a mistake bound of $\mathsf{Ldim}(\mathcal{H})$ in the realizable case. This implies, by Lemma 6.3, a sequential cover $\mathcal{G}$ of size

$$\log |\mathcal{G}| \leq O(\mathsf{Ldim}(\mathcal{H}) \cdot \log T).$$

Applying the EWA algorithm over $\mathcal{G}$ and using the property of sequential covering, we deduce, from Theorem 3.3, the regret upper bound

$$\mathsf{reg}_T(\mathcal{H}) \leq O(\sqrt{T \log |\mathcal{G}|}) \leq O(\sqrt{\mathsf{Ldim}(\mathcal{H}) \cdot T \log T}).$$

The *lower bound* is more intricate. We first prove a simpler $\Omega(\sqrt{T})$ lower bound and assume that $|\mathcal{H}| \geq 2$. Taking any $\mathbf{x} \in \mathcal{X}$ such that there exist $h_0, h_1 \in \mathcal{H}$ so that $h_i(\mathbf{x}) = i$. We now select y^T uniformly over $\{0,1\}^T$ and select $\mathbf{x}_t := \mathbf{x}$ for all $t \leq T$. We have for any prediction rule Φ that $\mathbb{E}_{y^T}\left[\sum_{t=1}^T |\hat{y}_t - y_t|\right] = \frac{T}{2}$. Let k be the number of 1's in y^T. We have

$$\inf_{h \in \{h_0, h_1\}} \sum_{t=1}^T |h(\mathbf{x}) - y_t| = \min\{k, T-k\}.$$

Let ϵ^T be uniform over $\{\pm 1\}^T$, we have $\sum_{t=1}^T \epsilon_t$ distributed equally as $2k - T$. Note that $|k - \frac{T}{2}| = \frac{T}{2} - \min\{k, T-k\}$, we have by Khinchine's Inequality (Lemma 2.13) that

$$\mathbb{E}[\min\{k, T-k\}] \leq \frac{T}{2} - \frac{1}{\sqrt{8}}\sqrt{T}.$$

Therefore, the regret is lower bounded by $\sqrt{T/8}$.

The $\Omega(\sqrt{\mathsf{Ldim}(\mathcal{H})T})$ lower bound follows by a more careful selection of the $\mathbf{x}^T$. Assume that T is divisible by $\mathsf{Ldim}(\mathcal{H})$ (otherwise we truncate T). We partition $\mathbf{x}^T, y^T$ into $\mathsf{Ldim}(\mathcal{H})$ blocks each of size $\frac{T}{\mathsf{Ldim}(\mathcal{H})}$, and denote k_i as the number of 1's in the i'th block of y^T.

Let τ be a $\mathcal{X}$-valued binary tree of depth $\mathsf{Ldim}(\mathcal{H})$ that can be shattered by $\mathcal{H}$. We now select y^T uniformly over $\{0,1\}^T$ and select $\mathbf{x}^T$ by traversing τ:

1. We assign the same value within each block of $\mathbf{x}^T$, with the first block being the value of the root v_1 of τ.

2. Let v_i be the node in τ for the i's block. If $k_i \geq \frac{T}{2\mathsf{Ldim}(\mathcal{H})}$ we set v_{i+1} being left child of v_i, and set to the right child otherwise.

By definition of shattering, $\exists h \in \mathcal{H}$ that achieves $\min\{k_i, \frac{T}{\mathsf{Ldim}(\mathcal{H})} - k_i\}$ losses for all i simultaneously. The regret is then lower bounded by

$$\Omega(\mathsf{Ldim}(\mathcal{H}) \cdot \sqrt{T/\mathsf{Ldim}(\mathcal{H})}) = \Omega(\sqrt{\mathsf{Ldim}(\mathcal{H})T})$$

which completes the proof. □

6.2 Minimax Regret for Real-Valued Class

In this section, we consider classes of real-valued functions $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$. To analyze their learnability, we begin by introducing the notion of *sequential Rademacher complexity.*

6.2.1 Sequential Rademacher Complexity

We start with the definition of sequential Rademacher complexity.

Definition 6.3 (Sequential Rademacher Complexity). For any real-valued class $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$, we define the *sequential Rademacher complexity* of $\mathcal{H}$ as

$$\mathsf{sRad}_T(\mathcal{H}) = \sup_{\tau} \mathbb{E}_{\epsilon^T}\left[\sup_{h\in\mathcal{H}} \sum_{t=1}^{T} \epsilon_t h(\tau(\epsilon^{t-1}))\right],$$

where $\tau : \bigcup_{i=0}^{T}\{-1,+1\}^i \to \mathcal{X}$ runs over all $\mathcal{X}$-valued binary trees of depth T, and ϵ^T is sampled uniformly over $\{-1,+1\}^T$.

Note that the sequential Rademacher complexity is similar to classical Rademacher complexity, except that the optimizing is over trees instead of sequences.

Example 6.3. Let $\mathcal{H}^{\mathsf{lin}} := \{h_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle : \mathbf{w} \in B_2\}$ be the class of linear functions with weight $\mathbf{w}$ lying in a unit L_2 ball. Let $\mathcal{X} := B_2$ as well, we have

$$\mathsf{sRad}_T(\mathcal{H}^{\mathsf{lin}}) \leq \sqrt{T}.$$

Indeed, fix any tree τ and denote $\mathbf{x}_t := \tau(\epsilon^{t-1})$, we have:

$$\begin{aligned}
\mathsf{sRad}_T(\mathcal{H}^{\mathsf{lin}}) &= \sup_{\tau} \mathbb{E}_{\epsilon^T}\left[\sup_{\mathbf{w}\in B_2} \sum_{t=1}^{T} \epsilon_t \langle \mathbf{w}, \mathbf{x}_t)\rangle\right] \\
&= \sup_{\tau} \mathbb{E}_{\epsilon^T}\left[\sup_{\mathbf{w}\in B_2} \left\langle \mathbf{w}, \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t) \right\rangle\right] \\
&\leq \sup_{\tau} \mathbb{E}_{\epsilon^T} \sqrt{\left\langle \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t, \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t \right\rangle}, \\
&\leq \sup_{\tau} \sqrt{\mathbb{E}_{\epsilon^T} \left\langle \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t, \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t \right\rangle}, \text{ by Jensen's inequality} \\
&\leq \sup_{\tau} \sqrt{\mathbb{E}_{\epsilon^T}\left[T + \sum_{i\neq j\leq T} \epsilon_i \epsilon_j \mathbf{x}_i^T \mathbf{x}_j\right]}, \text{ by } \|\mathbf{x}_t\|_2 \leq 1 \\
&= \sqrt{T}
\end{aligned}$$

as expected.

We now introduce a general approach for reducing the minimax regret to sequential Rademacher complexity. From Theorem 4.2, we know that the minimax regret can be reduced to

$$\begin{aligned}
\mathsf{reg}_T(\mathcal{H}) &= \sup_{\mu} \mathbb{E}\left[\sum_{t=1}^{T} \inf_{\hat{y}\in\hat{\mathcal{Y}}} \mathbb{E}_t[\ell(\hat{y}_t, y_t)] - \inf_{h\in\mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)\right] \\
&= \sup_{\mu} \mathbb{E}\left[\sup_{h\in\mathcal{H}} \left\{\sum_{t=1}^{T} \inf_{\hat{y}\in\hat{\mathcal{Y}}} \mathbb{E}_t[\ell(\hat{y}_t, y_t)] - \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)\right\}\right] \\
&\leq \sup_{\mu} \mathbb{E}\left[\sup_{h\in\mathcal{H}} \left\{\sum_{t=1}^{T} \mathbb{E}_t[\ell(h(\mathbf{x}_t), y_t)] - \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t)\right\}\right].
\end{aligned}$$

Denote $h^\ell(\mathbf{z}_t) := \ell(h(\mathbf{x}_t), y_t)$ where $\mathbf{z}_t = (\mathbf{x}_t, y_t)$. We obtain the following upper bound

$$\sup_\mu \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\sum_{t=1}^T \mathbb{E}_t[h^\ell(\mathbf{z}_t)] - h^\ell(\mathbf{z}_t)\right\}\right].$$

We now introduce a tangent sequence $\mathbf{z}'_1, \cdots, \mathbf{z}'_T$ such that $\mathbf{z}'_t = (\mathbf{x}'_t, y'_t)$ with $\mathbf{x}'_t = \mathbf{x}_t$ and y'_t being an *i.i.d.* copy of y_t conditioning on $\mathbf{x}^t, y^{t-1}$. The upper bound can be expresses as

$$\begin{aligned}
&\sup_\mu \mathbb{E}_{\mathbf{z}^T}\left[\sup_{h\in\mathcal{H}}\left\{\sum_{t=1}^T \mathbb{E}_t[h^\ell(\mathbf{z}'_t)] - h^\ell(\mathbf{z}_t)\right\}\right], \text{ by definition of } \mathbf{z}'^T \\
&\quad\le \sup_\mu \mathbb{E}_{\mathbf{z}^T}\mathbb{E}_{\mathbf{z}'^T}\left[\sup_{h\in\mathcal{H}}\left\{\sum_{t=1}^T h^\ell(\mathbf{z}'_t) - h^\ell(\mathbf{z}_t)\right\}\right], \text{ by } \sup\mathbb{E} \le \mathbb{E}\sup \\
&\quad\overset{(a)}{=} \sup_\mu \mathbb{E}_{\mathbf{x}_1}\mathbb{E}_{y_1,y'_1}\mathbb{E}_{\epsilon_1}\cdots\mathbb{E}_{\mathbf{x}_T}\mathbb{E}_{y_T,y'_T}\mathbb{E}_{\epsilon_T}\left[\sup_{h\in\mathcal{H}}\left\{\sum_{t=1}^T \epsilon_t(h^\ell(\mathbf{z}'_t) - h^\ell(\mathbf{z}_t))\right\}\right] \\
&\quad\overset{(b)}{\le} 2\sup_\mu \mathbb{E}_{\mathbf{x}_1}\mathbb{E}_{y_1}\mathbb{E}_{\epsilon_1}\cdots\mathbb{E}_{\mathbf{x}_T}\mathbb{E}_{y_T}\mathbb{E}_{\epsilon_T}\left[\sup_{h\in\mathcal{H}}\left\{\sum_{t=1}^T \epsilon_t h^\ell(\mathbf{z}_t)\right\}\right]
\end{aligned}$$

where ϵ_t is uniform over $\{\pm1\}$ and is (conditional) independent of y_t, y'_t. Here (a) follows by the conditional symmetries of y_t, y'_t and (b) follows by $\sup(A+B) \le \sup A + \sup B$ and symmetries between y_t, y'_t.

Note that, the following operator inequality holds (by $\mathbb{E} \le \sup$):

$$\mathbb{E}_{\mathbf{x}_1}\mathbb{E}_{y_1}\mathbb{E}_{\epsilon_1}\cdots\mathbb{E}_{\mathbf{x}_T}\mathbb{E}_{y_T}\mathbb{E}_{\epsilon_T} \le \sup_{\mathbf{x}_1,y_1}\mathbb{E}_{\epsilon_1}\cdots\sup_{\mathbf{x}_T,y_T}\mathbb{E}_{\epsilon_T}.$$

By Skolemization as in Lemma 2.2, the upper bound then equals

$$\sup_{\mathbf{x}_1,y_1}\mathbb{E}_{\epsilon_1}\cdots\sup_{\mathbf{x}_T,y_T}\mathbb{E}_{\epsilon_T}\left[\sup_{h\in\mathcal{H}}\left\{\sum_{t=1}^T \epsilon_t h^\ell(\mathbf{z}_t)\right\}\right] = \underbrace{\sup_\tau \mathbb{E}_{\epsilon^T}\left[\sup_{h\in\mathcal{H}}\left\{\sum_{t=1}^T \epsilon_t h^\ell(\tau(\epsilon^{t-1}))\right\}\right]}_{\mathsf{sRad}(\mathcal{H}^\ell)},$$

where τ runs over all $(\mathcal{X}\times\mathcal{Y})$-valued binary trees.

Putting everything together, we have proved the following lemma.

Lemma 6.5. The following holds

$$\mathsf{reg}_T(\mathcal{H}) \le 2\cdot\mathsf{sRad}_T(\mathcal{H}^\ell),$$

where $\mathcal{H}^\ell := \{\ell(h(\mathbf{x}), y) : h \in \mathcal{H}\} \in \hat{\mathcal{Y}}^{(\mathcal{X}\times\mathcal{Y})}$.

To relate the sequential Rademacher complexity to the minimax regret, we need the following Lipschitz contraction inequality for sequential Rademacher complexity from [27, Lemma 12].

Lemma 6.6. Let $\mathcal{H} \subset \mathbb{R}^{\mathcal{Z}}$ and $\phi : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$. If for all $\mathbf{z} \in \mathcal{Z}$, $\phi(\cdot, \mathbf{z})$ is a L-Lipschitz function, then

$$\mathsf{sRad}_T(\phi(\mathcal{H})) \leq O(L \cdot \log^{3/2} T) \cdot \mathsf{sRad}_T(\mathcal{H}),$$

where $\phi(\mathcal{H}) = \{\mathbf{z} \to \phi(h(\mathbf{z}), z) : h \in \mathcal{H}\}$.

6.2.2 Regret via Sequential Rademacher Complexity

Now we are in the position to present general bounds for minimax regret when the loss function is bounded and Lipschitz.

Theorem 6.7. Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0, 1]$ and $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ be a real-valued class. If the loss function ℓ is bounded, convex, and Lipschitz in its first argument, then:

$$\mathsf{reg}_T(\mathcal{H}) \leq \tilde{O}(\mathsf{sRad}_T(\mathcal{H})).$$

Moreover, for the absolute loss $\ell(\hat{y}, y) = |\hat{y} - y|$, we have

$$\mathsf{reg}_T(\mathcal{H}) \geq \Omega(\mathsf{sRad}_T(\mathcal{H})).$$

Proof. The upper bound follows directly from Lemma 6.6 and Lemma 6.5.

We focus now on the lower bound. Let $\tau : \bigcup_{i=0}^{T}\{0, 1\}^i \to \mathcal{X}$ be any $\mathcal{X}$-valued binary tree of depth T. We define a specific distribution μ over $(\mathcal{X} \times \mathcal{Y})^T$ as follows:

1. Sample y^T uniformly from $\{0, 1\}^T$;
2. Let $\mathbf{x}_t = \tau(y^{t-1})$.

Note that $\inf_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_t[|\hat{y}_t - y_t|] = \frac{1}{2}$, since y_t is uniform over $\{0, 1\}$ conditioning on $\mathbf{x}^t, y^{t-1}$. That is the Bayesian optimal risk equals $\frac{T}{2}$. Moreover, $|h(\mathbf{x}_t) - y_t| = \epsilon_t h(\mathbf{x}_t) + (1 - \epsilon_t)/2$, where $\epsilon_t = 1 - 2y_t \in \{-1, +1\}$. Therefore, by Theorem 4.2, we have

$$\begin{aligned}\mathsf{reg}_T(\mathcal{H}) &\geq \mathbb{E}_{y^T}\left[\frac{T}{2} - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T}\left(\epsilon_t h(\mathbf{x}_t) + \frac{1 - \epsilon_t}{2}\right)\right] \\ &= \mathbb{E}_{\epsilon^T}\left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} \epsilon_t h(\mathbf{x}_t)\right],\end{aligned}$$

where the equality follows by $\mathbb{E}_{y_t}[(1-\epsilon_t)/2] = \frac{1}{2}$ and changing measure to ϵ^T. Since τ is selected arbitrary, the inequality remains holds when taking $\sup_\tau$. We conclude that $\mathsf{reg}_T(\mathcal{H}) \geq \mathsf{sRad}_T(\mathcal{H})$, as needed. □

It should be noted that the logarithmic factor in the upper bound can be improved by a constant factor via a direct minimax switching analysis of the regret (without going through Lemma 6.5); see [53].

6.2.3 Minimax Regret for General Lipschitz Loss Functions

We have shown in Theorem 6.7 that for Lipschitz losses, the minimax regret is tightly characterized by the sequential Rademacher complexity. To obtain concrete bounds, we need to bound the sequential Rademacher complexity as well via the sequential fat-shattering dimension, already introduced in Section 5.4.

The *sequential α-fat-shattering dimension* $\mathsf{sfat}_\alpha(\mathcal{H})$ for a class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ is defined as the maximal number d such that $\mathcal{H}$ can α-fat-shatter certain trees τ, s of depth d. We recall that in Section 5.4 we also denoted this dimension as $d(\alpha)$ but in this section we rather use $\mathsf{sfat}_\alpha(\mathcal{H})$ to show its dependence on $\mathcal{H}$.

To proceed, we recall the real-valued sequential covering as introduced in Definition 5.1.

Definition 6.4 ((Real-valued) Sequential Cover)**.** Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ and $\mathcal{G} \subset [0,1]^{\mathcal{X}^*}$ be a class mapping $\mathcal{X}^* \to [0,1]$. We say that the class $\mathcal{G}$ sequentially α-covers $\mathcal{H}$ up to step T if, for any $\mathbf{x}^T \in \mathcal{X}^T$ and $h \in \mathcal{H}$, there exists $g \in \mathcal{G}$ such that

$$\forall t \leq T, \ |g(\mathbf{x}^t) - h(\mathbf{x}_t)| \leq \alpha.$$

Similar to the binary-valued case, we can bound the (real-valued) sequential cover via the sequential fat-shattering dimension as follows:

Lemma 6.8. For any class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ with sequential α-fat-shattering dimension $\mathsf{sfat}_\alpha(\mathcal{H})$, there exists a sequential α-cover $\mathcal{G}_\alpha$ of $\mathcal{H}$ such that

$$\log|\mathcal{G}_\alpha| \leq \tilde{O}(\mathsf{sfat}_{\alpha/3}(\mathcal{H})),$$

where $\tilde{O}$ hides poly-logarithmic factors in α and T.

Proof. Let $K = \{2i\alpha : i \leq [1/(2\alpha)]\}$ be a discretization of $[0, 1]$ such that for any $a \in [0, 1]$, there exists $b \in K$ where $|a - b| \leq \alpha$. For any $h \in \mathcal{H}$, we define a function $h' \in K^{\mathcal{X}}$ such that

$$h'(\mathbf{x}) = \arg\min_{\beta \in K} |h(\mathbf{x}) - \beta|.$$

Let $\mathcal{H}' = \{h' : h \in \mathcal{H}\} \subset K^{\mathcal{X}}$. It is easy to observe that any sequential 2α-cover of $\mathcal{H}'$ implies a sequential 3α-cover of $\mathcal{H}$. Our primary goal is now reduced to bounding the 2α-covering set size of $\mathcal{H}'$. To achieve this, we introduce the following concept:

1-Shattering Dimension: The 1-shattering dimension of $\mathcal{H}'$ is defined as the maximum number d such that there exist a $\mathcal{X}$-valued tree τ and a K-valued tree s, both of depth d, such that $\forall \epsilon^d \in \{0, 1\}^d$, $\exists h' \in \mathcal{H}'$ we have:

1. If $\epsilon_t = 0$, then $h'(\tau(\epsilon^{t-1})) \leq s(\epsilon^{t-1}) - 2\alpha$;
2. If $\epsilon_t = 1$, then $h'(\tau(\epsilon^{t-1})) \geq s(\epsilon^{t-1}) + 2\alpha$.

We denote $\mathsf{FAT}_1(\mathcal{H}')$ as the 1-shattering dimension of $\mathcal{H}'$. It is easy to observe that $\mathsf{FAT}_1(\mathcal{H}') \leq \mathsf{sfat}_\alpha(\mathcal{H})$.

We now introduce the M-SOA algorithm:

Algorithm 6.2 The M-SOA Algorithm

```
Initialize the running hypothesis class: $\mathcal{H}^{(0)} \leftarrow \mathcal{H}'$
for $t = 1$ to $T$ do
    for each $\beta \in K$ do
        Define $\mathcal{H}^{(t)}_\beta \leftarrow \{h \in \mathcal{H}^{(t-1)} : h(\mathbf{x}_t) = \beta\}$
    end for
    Predict $\hat{y}_t \leftarrow \arg\max_{\beta \in K} \left\{\mathsf{FAT}_1(\mathcal{H}^{(t)}_\beta)\right\}$
    Observe true label $y_t$
    if $|\hat{y}_t - y_t| > 2\alpha$ then
        Update $\mathcal{H}^{(t)} \leftarrow \mathcal{H}^{(t)}_{y_t}$
    else
        Keep $\mathcal{H}^{(t)} \leftarrow \mathcal{H}^{(t-1)}$
    end if
end for
```

Claim 1. The M-SOA algorithm satisfies the following realizable risk bound:

$$\sup_{\mathbf{x}^T} \sup_{h' \in \mathcal{H}'} \sum_{t=1}^{T} 1\{|\hat{y}_t - h'(\mathbf{x}_t)| > 2\alpha\} \leq \mathsf{FAT}_1(\mathcal{H}').$$

Proof. At each time step t such that $|\hat{y}_t - y_t| > 2\alpha$, the algorithm updates the hypothesis class by setting $\mathcal{H}^{(t)} = \mathcal{H}^{(t)}_{y_t} \subseteq \mathcal{H}^{(t-1)}$. Since the prediction $\hat{y}_t$ was chosen to maximize $\mathsf{FAT}_1(\mathcal{H}^{(t)}_{\beta})$ over $\beta \in K$, and the prediction was incorrect by more than 2α, this update necessarily eliminates hypotheses that could otherwise achieve a higher fat-shattering dimension. Therefore, $\mathsf{FAT}_1(\mathcal{H}^{(t)})$ decreases by at least 1 at each such step. Since the initial class $\mathcal{H}^{(0)} = \mathcal{H}'$ has fat-shattering dimension $\mathsf{FAT}_1(\mathcal{H}')$, the number of mistakes of this type is at most $\mathsf{FAT}_1(\mathcal{H}')$. □

For any $I \subset [T]$ and $\{\beta_t\}_{t\in I} \in K^{|I|}$, we define a sequential function by simulating the M-SOA algorithm with the following modification at steps 8–12:

1. If $t \in I$, update $\mathcal{H}^{(t)} = \mathcal{H}^{(t)}_{\beta_t}$;
2. If $t \notin I$, proceed as in M-SOA.

Let $\mathcal{G}$ denote the collection of all such sequential functions with $|I| \leq \mathsf{FAT}_1(\mathcal{H})$. We now observe that the class $\mathcal{G}$ sequentially 2α-covers $\mathcal{H}'$, and

$$\log|\mathcal{G}| \leq O\left(\mathsf{FAT}_1(\mathcal{H}')\log(|K|T)\right).$$

Here, the covering property follows from the risk bound in Claim 1 and a similar argument as in Lemma 6.3. While the upper bound on the size follows by counting the number of such sets I and corresponding choices of $\{\beta_t\}_{t\in I}$.

The lemma follows by combining all the preceding results. □

We conclude this section with the following theorem, which relates all of the concepts introduced.

Theorem 6.9. Let $\mathcal{Y} = \hat{\mathcal{Y}} := [0,1]$, and let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$. Assume the loss function ℓ is bounded, convex, and Lipschitz in its first argument. Then, the following statements are equivalent for given $p \geq 2$ (where $\tilde{O}$ hides poly-logarithmic factors in α and T):

1. The *Sequential Fat-Shattering Dimension* $\mathsf{sfat}_\alpha(\mathcal{H}) = \tilde{\Theta}(\alpha^{-p})$;
2. There exists a *Sequential α-cover* $\mathcal{G}_\alpha$ with $\log|\mathcal{G}_\alpha| = \tilde{\Theta}(\alpha^{-p})$;
3. The *Sequential Rademacher Complexity* $\mathsf{sRad}_T(\mathcal{H}) = \tilde{\Theta}(T^{\frac{p-1}{p}})$;
4. The *minimax regret* $\mathsf{reg}_T(\mathcal{H}) = \tilde{\Theta}(T^{\frac{p-1}{p}})$.

In this section, we have shown that $1 \Rightarrow 2$ (Lemma 6.8) and $3 \Leftrightarrow 4$ (Theorem 6.7). The remaining implications are more technical and are addressed in [27]. A direct implication $2 \Rightarrow 4$ can be obtained via the EWA algorithm, yielding a regret bound of $\mathsf{reg}_T(\mathcal{H}) \leq \tilde{O}(T^{\frac{p+1}{p+2}})$. The tighter bound of $\tilde{O}(T^{\frac{p-1}{p}})$ follows from the use of chaining, through the path $2 \Rightarrow 3 \Rightarrow 4$. We omit the full proof and refer the reader to [27].

6.3 Bibliographical Notes

The Littlestone dimension was introduced in [54] (see also [55]), where the Standard Optimal Algorithm (SOA) was presented. It has since become a fundamental tool in online learning. The generalization to the non-realizable case was first studied in [24]. Extensions to real-valued function classes and the introduction of sequential Rademacher complexity were developed in [27], [53].

7

Expected Worst Case Minimax Regret

In this chapter, we study the problem of online learning in the *hybrid* setting, where features are drawn from a general unknown *stochastic* process while labels remain adversarial. This contrasts to the *worst-case* minimax regret considered in previous chapters and offers a broader and more relaxed modeling of learning environments.

The central performance measure in hybrid online learning is the expected worst-case minimax regret, introduced in (4.4) and recalled below in (7.1). To simplify notation, we will abbreviate the expected worst-case minimax regret as $\tilde{r}_T(\mathcal{H}) := \overline{\mathsf{reg}}_T(\mathcal{H})$. Additionally, we write $r_T^*(\mathcal{H} \mid \mathbf{x}^T) := \mathsf{reg}_T^*(\mathcal{H} \mid \mathbf{x}^T)$ and $r_T(\mathcal{H}) := \mathsf{reg}_T(\mathcal{H})$ (worst case regret), as previously discussed in Chapter 4.

7.1 Problem Formulation

Let $\mathcal{X}$ be a feature space, $\mathcal{Y}$ be the true label space, and $\hat{\mathcal{Y}} = [0,1]$ be the space of outputs of the learner. We denote by $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ a class of functions $\mathcal{X} \to \hat{\mathcal{Y}}$. For any time horizon T, we consider a class P of distributions (i.e., random processes) over $\mathcal{X}^T$. We consider the following game between Nature and the predictor. At the beginning, Nature selects a distribution $\boldsymbol{\nu}^T \in \mathsf{P}$ and samples an input sequence

$\mathbf{x}^T \sim \boldsymbol{\nu}^T$, where $\mathbf{x}^T \in \mathcal{X}^T$. At each time step $t \leq T$, Nature reveals the t-th sample $\mathbf{x}_t$ of $\mathbf{x}^T$ to the predictor. The predictor then makes a prediction $\hat{y}_t \in \mathcal{Y}$ using a strategy $\phi_t : \mathcal{X}^t \times \mathcal{Y}^{t-1} \to \hat{\mathcal{Y}}$ potentially using the history observed thus far, that is, $\hat{y}_t = \phi_t(\mathbf{x}^t, y^{t-1})$. After the prediction, Nature reveals the true label y_t and the predictor incurs a loss $\ell(\hat{y}_t, y_t)$ for some predefined convex loss function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to [0, \infty)$. We are interested in the following *expected worst-case* minimax regret:

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) = \inf_{\phi^T} \sup_{\boldsymbol{\nu}^T \in \mathsf{P}} \mathbb{E}_{\mathbf{x}^T \sim \boldsymbol{\nu}^T} \left[\sup_{y^T} \left(\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right) \right]. \tag{7.1}$$

We note that the expected worst-case minimax regret $\tilde{r}_T(\mathcal{H}, \mathsf{P})$ recovers the sequential and fixed-design minimax regrets discussed in Chapter 4 by selecting an appropriate distribution class P.

We also introduce the following expected *average case* minimax regret:

$$\bar{r}_T(\mathcal{H}, \mathsf{P}) = \inf_{\phi^T} \sup_{\boldsymbol{\nu}^T \in \mathsf{P}, h \in \mathcal{H}} \mathbb{E}_{\mathbf{x}^T \sim \boldsymbol{\nu}^T} \left[\sup_{y^T} \left(\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \ell(h(\mathbf{x}_t), y_t) \right) \right] \tag{7.2}$$

where the main difference with $\tilde{r}_T(\mathcal{H}, \mathsf{P})$ is the position of $\sup_h$.

The following observation is straightforward to prove and demonstrates the relationship among different notions of regret.

Proposition 7.1. If P is a class of all singleton distributions over $\mathcal{X}^T$, then $\tilde{r}_T(\mathcal{H}, \mathsf{P}) = \mathsf{reg}_T(\mathcal{H})$ (worst case regret) for all $\mathcal{H}$. If P is the singleton distribution that assigns probability 1 for $\mathbf{x}^T$, then $\tilde{r}_T(\mathcal{H}, \mathsf{P}) = r_T^*(\mathcal{H} \mid \mathbf{x}^T)$. Furthermore, $\tilde{r}_T(\mathcal{H}, \mathsf{P}) \geq \bar{r}_T(\mathcal{H}, \mathsf{P})$, for any $\mathcal{H}$ and P.

Example 7.1. To understand differences between $\tilde{r}_T$ and $\bar{r}_T$, we consider the following example. Let $\mathcal{H}$ be the class of all functions $[0, 1] \to \{0, 1\}$ that take value 1 on at most T positions and 0 otherwise. Let ν be the uniform distribution over $[0, 1]$, and $\ell(\hat{y}_t, y_t) = |\hat{y}_t - y_t|$, where $\hat{y}_t \in [0, 1]$ and $y_t \in \{0, 1\}$. We will denote by $\nu^{\otimes T}$ the *i.i.d* distribution of length T with marginal ν. We have $\bar{r}_T(\mathcal{H}, \nu^{\otimes T}) = 0$, since for any h, w.p. 1 we have $h(x_t) = 0$ for all $t \in [T]$, meaning that a strategy that predicts 0 all the time incurs 0 regret. However, we also have $\tilde{r}_T(\mathcal{H}, \nu^{\otimes T}) \geq \frac{T}{2}$. To see this, we choose $y^T \in \{0,1\}^T$ uniformly at random and observe that any strategy will make at least $\frac{T}{2}$ accumulated losses, however, for any $\mathbf{x}^T$ and y^T, there exists $h \in \mathcal{H}$ such that $\forall t \in [T],\ h(\mathbf{x}_t) = y_t$.

7.2 Stochastic Sequential Cover

Let $\mathcal{X}^*$ be the set of all finite sequences over $\mathcal{X}$. We introduce one of our main technical ingredients, i.e., the stochastic (global) sequential covering, as follows:

Definition 7.1 (Stochastic sequential cover). We say a class $\mathcal{G}$ of functions $\mathcal{X}^* \to [0,1]$ is a stochastic (global) sequential cover of a class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ w.r.t. the class P of distributions over $\mathcal{X}^T$ at scale $\alpha > 0$ and confidence $\delta > 0$, if for all $\boldsymbol{\nu}^T \in \mathsf{P}$

$$\Pr_{\mathbf{x}^T \sim \boldsymbol{\nu}^T}\left[\exists h \in \mathcal{H}\ \forall g \in \mathcal{G}\ \exists\ t \in [T]\ s.t.\ |h(\mathbf{x}_t) - g(\mathbf{x}^t)| > \alpha\right] \leq \delta.$$

We define the minimal size of $\mathcal{G}$ to be the stochastic global sequential covering number of $\mathcal{H}$.

Note that the distribution class P in Definition 7.1 is completely general and recovers the (classical) sequential covering discussed in Chapter 6 if P is the class of all *singleton* distributions over $\mathcal{X}^T$.

We first establish the following simple (but useful) composition property of stochastic sequential cover. Let $\mathcal{H}_1, \cdots, \mathcal{H}_m \subset [0,1]^{\mathcal{X}}$ be m function classes over the same domain and Θ be a parameter space equipped with some norm $||\cdot||$. For any function $F : [0,1]^m \times \Theta \to [0,1]$ such that $\forall \mathbf{z}_1, \mathbf{z}_2 \in [0,1]^m,\ \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$ we have $F(\mathbf{z}_1, \boldsymbol{\theta}_1) - F(z_2, \boldsymbol{\theta}_2) \leq L \max\{||\mathbf{z}_1 - \mathbf{z}_2||_\infty, ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||\}$ for some constant $L \in \mathbb{R}^+$, the F-composition of $\mathcal{H}_1, \cdots, \mathcal{H}_m$ and Θ is defined to be the class:

$$\mathcal{H} = \{h(\mathbf{x}) = F(h_1(\mathbf{x}), \cdots, h_m(\mathbf{x}), \boldsymbol{\theta}) : \forall i \in [m],\ h_i \in \mathcal{H}_i \text{ and } \boldsymbol{\theta} \in \Theta\}.$$

Proposition 7.2. Let $\mathcal{H}_1, \cdots, \mathcal{H}_m \subset [0,1]^{\mathcal{X}}$ be any classes, Θ be any parameter space equipped with norm $||\cdot||$, and F be any function satisfying the conditions above. If $\forall i \in [m]$, $\mathcal{H}_i$ admits a statistical sequential covering set $\mathcal{G}_i$ at scale α/L and confidence δ/m w.r.t. distribution class P, and Θ admits an α/L cover $\mathcal{C}$ under norm $||\cdot||$, then the F-composition class $\mathcal{H}$ of $\mathcal{H}_1, \cdots, \mathcal{H}_m$ and Θ admits a statistical sequential covering set $\mathcal{G}$ w.r.t. P at scale α and confidence δ such that:

$$|\mathcal{G}| \leq |\mathcal{C}| \prod_{i=1}^{m} |\mathcal{G}_i|.$$

Proof. For any tuple of indices $(j_1, \cdots, j_m)$ with $j_i \in [|\mathcal{G}_i|]$ and $\boldsymbol{\theta}' \in \mathcal{C}$, we construct a function g such that:

$$g(\mathbf{x}^t) = F(g_{j_1}(\mathbf{x}^t), \cdots, g_{j_m}(\mathbf{x}^t), \boldsymbol{\theta}'),$$

where g_{j_i} is the j_ith function in $\mathcal{G}_i$. The covering set $\mathcal{G}$ is defined to be the class containing of all such functions g. For any function $h \in \mathcal{H}$, there exist $h_1, \cdots, h_m$ with $h_i \in \mathcal{H}_i$ and $\boldsymbol{\theta} \in \Theta$ such that for all $\mathbf{x} \in \mathcal{X}$, $h(\mathbf{x}) = F(h_1(\mathbf{x}), \cdots h_m(\mathbf{x}), \boldsymbol{\theta})$. By union bound and definition of stochastic sequential covering of $\mathcal{G}_i$, w.p. $\geq \delta$ over $\mathbf{x}^T$, for all $i \in [m]$, there exist $g_{j_i} \in \mathcal{G}_i$ such that $\forall t \in [T]$, $|g_{j_i}(\mathbf{x}^t) - h_i(\mathbf{x}_t)| \leq \alpha/L$. One can verify that the function g corresponding to $(j_1, \cdots, j_m)$ and $\boldsymbol{\theta}' \in \mathcal{C}$ closest to $\boldsymbol{\theta}$ under $||\cdot||$ is the desired function that covers h on $\mathbf{x}^T$, due to the L-Lipschitz property of F. □

We demonstrate below how F-composition can be exploited to generate interesting complex classes from simple classes.

Example 7.2. Let $\Theta = [0,1]^2$ and $\mathcal{H}_1 \subset \{0,1\}^{\mathcal{X}}$ be a binary valued class of finite VC-dimension. If we take $F(y, \boldsymbol{\theta}) = y\theta_1 + (1-y)\theta_2$ for $y \in \{0,1\}$ and $\boldsymbol{\theta} \in [0,1]^2$, the F-composition class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ of $\mathcal{H}_1$ and Θ recovers the setup of [56]. We note that in this case the set Θ admits an α-covering set of size $O(\alpha^{-2})$ under L_∞ norm for all $\alpha > 0$ and F is 1-Lipschitz in the sense of Proposition 7.2.

Example 7.3. Let $\Theta = \{\theta_1 + \cdots + \theta_d \leq 1 : \boldsymbol{\theta} \in [0,1]^d\}$ for some $d \in \mathbb{N}^+$ and $\mathcal{H}_1, \cdots, \mathcal{H}_d \subset \{0,1\}^{\mathcal{X}}$ be d binary-valued classes of finite VC-dimension. If we take $F(\mathbf{y}^d, \boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{y}^d \rangle$ for $\mathbf{y}^d \in \{0,1\}^d$ and $\boldsymbol{\theta} \in \Theta$, the F-composition class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ of $\mathcal{H}_i$'s and Θ defines a natural class. We note that in this case Θ is α-covered by a set of size α^{-d} under L_1 norm and F is 1-Lipschitz in the sense of Proposition 7.2. Moreover, if we take $d = 2$ and $\mathcal{H}_2 = \{1 - h(\mathbf{x}) : h \in \mathcal{H}_1\}$ we subsume the setup of Example 7.2.

Example 7.4. Let Θ be empty, $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{H}_i = \{h_{[a,b]}(\mathbf{x}) = 1\{\mathbf{x}[i] \in [a,b]\} : [a,b] \subset [0,1],\ \mathbf{x} \in \mathbb{R}^d\}$ for $i \in [d]$, i.e., $\mathcal{H}_i$ is the class of indicators of intervals on the ith coordinate of $\mathbf{x}$. If we take $F(y_1, \cdots, y_d) = \prod_{i=1}^d y_i$ for $y^d \in \{0,1\}^d$, the F-composition class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ of $\mathcal{H}_i$'s defines the class of indicators of rectangular cuboids in $\mathbb{R}^d$ and F is 1-Lipschitz.

7.2.1 Upper bounds on regret via stochastic sequential covering

We now prove two general results below that demonstrate how a bound on the stochastic sequential covering number implies bounds on the expected worst-case regret $\tilde{r}_T$ in an algorithmic fashion.

Theorem 7.1. Let $\mathcal{H}$ be a set of functions mapping $\mathcal{X} \to [0,1]$ and $\mathcal{G}_\alpha$ be a stochastic sequential covering of $\mathcal{H}$ at scale α and confidence $\delta = 1/T$ w.r.t. distribution class P. If $\ell(\cdot, y)$ is convex, L-Lipschitz, and bounded by 1 on $\hat{\mathcal{Y}}$ for any $y \in \mathcal{Y}$, then:

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq \inf_{0 \leq \alpha \leq 1} \left\{ \alpha L T + \sqrt{(T/2)\log|\mathcal{G}_\alpha|} + 1 \right\}.$$

Proof. We run the Exponential Weighted Average (EWA) Algorithm 3.3 on $\mathcal{G}_\alpha$. We split the regret into two parts, one that is incurred by the predictor against $\mathcal{G}_\alpha$ and the other that is incurred by the discrepancy between $\mathcal{G}_\alpha$ and $\mathcal{H}$. For the first term, using Theorem 3.4 we conclude that with probability 1 on $\mathbf{x}^T$:

$$\sum_{t=1}^T \ell(\hat{y}_t, y_t) \leq \inf_{g \in \mathcal{G}_\alpha} \sum_{t=1}^T \ell(g(\mathbf{x}^t), y_t) + \sqrt{(T/2)\log|\mathcal{G}_\alpha|}.$$

For the second term, we denote by A the event described in the probability of Definition 7.1. Since $\Pr[A] \leq \frac{1}{T}$ and $\ell(\hat{y}, y) \leq 1$ by assumption, the expected regret contributed by the event A is at most 1. We now condition on the event that A does not happen. By Definition 7.1, we obtain $\forall h \in \mathcal{H} \exists g \in \mathcal{G}_\alpha \forall t \in [T],\ |h(\mathbf{x}_t) - g(\mathbf{x}^t)| \leq \alpha$. Since ℓ is bounded by 1 and L-Lipschitz, we have:

$$\inf_{h \in \mathcal{H}} \sum_{t=1}^T \ell(h(\mathbf{x}_t), y_t) \geq \inf_{g \in \mathcal{G}_\alpha} \sum_{t=1}^T \ell(g(\mathbf{x}^t), y_t) - \alpha L T.$$

The result follows by combining these inequalities. □

Now we turn our attention to the *logarithmic loss.*

Theorem 7.2. Let $\mathcal{Y} = \{0,1\}$, $\hat{\mathcal{Y}} = [0,1]$ and ℓ be the logarithmic loss. If for all $\alpha \geq 0$ there exists a stochastic sequential covering set $\mathcal{G}_\alpha$ of class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ w.r.t. distribution class P at scale α and confidence $\delta = \frac{1}{T}$, then:

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq \inf_{0 \leq \alpha \leq 1} \{2\alpha T + \log(|\mathcal{G}_\alpha| + 1) + \log(|\mathcal{G}_\alpha| + 1)/T + 1\} .$$

Proof. The proof is similar to the proof of Theorem 7.1, but replacing the EWA algorithm with the Smooth truncated Bayesian Algorithm 5.2 and running the algorithm on $\mathcal{G}_\alpha \cup \{u\}$ with truncation parameter α and uniform prior, where u is the function that maps to $\frac{1}{2}$ for all $\mathbf{x}^t$. We again split the regret into two parts, one incurred by the Smooth truncated Bayesian Algorithm 5.2, and the other incurred by the error of covering. By Theorem 5.3 the first term is upper bounded by $2\alpha T + \log(|\mathcal{G}_\alpha| + 1)$. For the error term, we note that we have added all $\frac{1}{2}$ valued functions u into the expert class when running the Smooth truncated Bayesian Algorithm. This implies that the prediction rule can only incur the *actual* accumulated losses upper bounded by $T + \log(|\mathcal{G}_\alpha| + 1)$. Therefore, when the event A (defined in proof of Theorem 7.1) happens, the expected regret only contributes $(T + \log(|\mathcal{G}_\alpha| + 1)) \cdot \Pr[A] \leq (T + \log(|\mathcal{G}_\alpha| + 1))/T$. The result follows by combining the inequalities. □

7.3 Stochastic Cover for Binary Valued Classes

This section focuses on the stochastic sequential covering number of binary valued classes $\mathcal{H}$. We assume that P is the class of all *i.i.d.* distributions over $\mathcal{X}^T$; however, our results hold for general *exchangeable* processes [57] over $\mathcal{X}^T$ as well, i.e., distributions that are invariant under permutation of the indexes.

7.3.1 Stochastic sequential cover for finite VC-class

Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be binary valued class with finite VC-dimension. We write $\mathsf{VC}(\mathcal{H})$ for the VC-dimension of $\mathcal{H}$. We will show that the stochastic global sequential covering number can be upper bounded by $e^{O(\mathsf{VC}(\mathcal{H}) \log^2 T)}$ w.h.p. using the *1-inclusion graph* algorithm that was introduced in [58]. Without going into the technical details of the *1-inclusion graph* algorithm, we can understand it as a function that maps $(\mathcal{X} \times \{0,1\})^{t-1} \times \mathcal{X} \to \{0,1\}$, for any given $t \geq 1$. For $\mathcal{H}$ of finite VC-dimension and any function $\Phi : (\mathcal{X} \times \{0,1\})^{t-1} \times \mathcal{X} \to \{0,1\}$, we define the following quantity (here, we follow the notation in [58]):

$$\hat{\hat{M}}_{\Phi,\mathcal{H}}(t) = \sup_{\mathbf{x}^t \in \mathcal{X}^t} \sup_{h \in \mathcal{H}} \mathbb{E}_\sigma \left[1\{\Phi(\mathbf{x}^{\sigma(t)}, h(\{\mathbf{x}^{\sigma(t-1)}\})) \neq h(\mathbf{x}_{\sigma(t)})\} \right],$$

where σ is the uniform random permutation over $[t]$, we have $\mathbf{x}^{\sigma(t)} = \{\mathbf{x}_{\sigma(1)}, \cdots, \mathbf{x}_{\sigma(t)}\}$ and $h(\{\mathbf{x}^{\sigma(t-1)}\}) = \{h(\mathbf{x}_{\sigma(1)}), \cdots, h(\mathbf{x}_{\sigma(t-1)})\}$. The main result of [58] is stated as follows:

Theorem 7.3 (Haussler et al., Theorem 2.3(ii))**.** For any binary valued class $\mathcal{H}$ of finite VC-dimension and for any $t \geq 1$, there exists a function $\Phi : (\mathcal{X} \times \{0,1\})^{t-1} \times \mathcal{X} \to \{0,1\}$, i.e., the 1-inclusion graph algorithm, that satisfies

$$\hat{\hat{M}}_{\Phi,\mathcal{H}}(t) \leq \frac{\mathsf{VC}(\mathcal{H})}{t}.$$

Our main result for this part is as follows, with the proof presented below Lemma 7.6.

Theorem 7.4. For any binary valued class $\mathcal{H}$ with finite VC-dimension, there exists a global sequential covering set $\mathcal{G}$ of $\mathcal{H}$ w.r.t. the class of all *i.i.d.* distributions over $\mathcal{X}^T$ at scale $\alpha = 0$ and confidence δ such that for $T \geq e^5$ we have:

$$\log|\mathcal{G}| \leq 5\mathsf{VC}(\mathcal{H}) \log^2 T + \log T \log(1/\delta) + \log T.$$

The main idea for proving Theorem 7.4 is to show that for the 1-inclusion graph predictor Φ, we have w.p. $\geq 1-\delta$ over the sample $\mathbf{x}^T \overset{i.i.d}{\sim} \boldsymbol{\nu}^T$, the *cumulative* error is upper bounded by $O(\mathsf{VC}(\mathcal{H}) \log T + \log(1/\delta))$. Assuming this holds, one will be able to construct the covering functions in a similar fashion as [24, Lemma 12]. The bound will follow by counting the error patterns. However, a direct application of Theorem 7.3 will only give us an *expected* $\mathsf{VC}(\mathcal{H}) \log T$ error bound. The main challenge follows from the fact that even though the samples $\mathbf{x}^T$ are generated *i.i.d.*, the predictions made by the 1-inclusion predictor are *not* independent (neither a martingale), and therefore the standard concentration inequalities do not apply directly.

Our main proof technique exploits *permutation invariance* of the 1-inclusion graph predictor, which allows us to relate the cumulative error to a *reversed* martingale[1]. Using Bernstein inequality for martingales Lemma 2.10, we then establish the following key lemma.

Lemma 7.5. Let $\Phi : (\mathcal{X} \times \{0,1\})^* \times \mathcal{X} \to \{0,1\}$ and $h : \mathcal{X} \to \{0,1\}$ be functions such that Φ is permutation invariant on $(\mathcal{X} \times \{0,1\})^*$. If for all $t \in [T]$ and $\mathbf{x}^t \in \mathcal{X}^t$ we have:

$$\mathrm{Pr}_{\sigma_t}\left[\Phi(\mathbf{x}^{\sigma_t(t)}, h(\{\mathbf{x}^{\sigma_t(t-1)}\})) \neq h(\mathbf{x}_{\sigma_t(t)})\right] \leq \frac{C}{t}, \tag{7.3}$$

where σ_t is the uniform random permutation on $[t]$ and $C \in \mathbb{N}^+$, then for all $\delta > 0$ and $T \geq e^5$ we have

$$\Pr_{\sigma_T}\left[\sum_{t=1}^{T} \mathbf{1}\Big\{\Phi(\mathbf{x}_{\sigma_T(t)},\, h(\{\mathbf{x}_{\sigma_T(s)}\}_{s<t})) \neq h(\mathbf{x}_{\sigma_T(t)})\Big\} \right.$$
$$\left. \geq 4C\log T + \log\left(\frac{1}{\delta}\right)\right] \leq \delta.$$

Proof. For any $t \in [T]$, define the indicator variable

$$I_t = \mathbf{1}\left\{\Phi(\mathbf{x}^{\sigma(t)}, h(\{\mathbf{x}^{\sigma(t-1)}\})) \neq h(\mathbf{x}_{\sigma(t)})\right\},$$

where σ is a uniform random permutation over $[T]$. For each $t \in [T]$, we define the *reversed* sequence of indicators as $I'_t = I_{T-t+1}$. We observe that for any $t \in [T]$, the indicator I'_t depends only on the realizations of $\mathbf{x}_{\sigma(T)}, \mathbf{x}_{\sigma(T-1)}, \dots, \mathbf{x}_{\sigma(T-t+1)}$, since Φ is permutation-invariant over $\mathbf{x}_{\sigma(1)}, \dots, \mathbf{x}_{\sigma(T-t)}$. Therefore,

$$\mathbb{E}[I'_t \mid I'_1, \dots, I'_{t-1}] = \mathbb{E}[I'_t \mid \mathbf{x}_{\sigma(T)}, \dots, \mathbf{x}_{\sigma(T-t+2)}] \leq \min\left\{\frac{C}{T-t+1}, 1\right\},$$

where the inequality follows from the fact that conditioning on $\mathbf{x}_{\sigma(T-t+2)}^{\sigma(T)}$ the permutation σ restricted on $\mathbf{x}^T \backslash \{\mathbf{x}_{\sigma(T)}, \cdots, \mathbf{x}_{\sigma(T-t+2)}\}$ is also a uniform random permutation, so that Eq. (7.3) applies.

[1] Note that [59, Proposition 10.2] also considers a similar martingale based approach only for an almost sure rate.

For any realization $I'_1, \dots, I'_{t-1}$, define the centered random variables:

$$I''_t = I'_t - \mathbb{E}[I'_t \mid I'_1, \dots, I'_{t-1}].$$

These variables form a martingale difference sequence, i.e., for all $t \in [T]$, $\mathbb{E}[I''_t \mid I''_1, \dots, I''_{t-1}] = 0$. Applying the Bernstein inequality for martingales (Lemma 2.10), we obtain:

$$\Pr\left[\sum_{t=1}^{T} I''_t > k \text{ and } \Sigma^2 \leq v\right] \leq \exp\left(-\frac{k^2}{2(v + k/3)}\right),$$

where

$$\Sigma^2 = \sum_{t=1}^{T} \mathbb{E}[I''^2_t \mid I''_1, \dots, I''_{t-1}].$$

Note that, conditioned on $I'_1, \dots, I'_{t-1}$, the indicator I'_t is a Bernoulli random variable with parameter $p_t \leq \min\left\{\frac{C}{T-t+1}, 1\right\}$. Therefore, if $I'_t = 1$, then $I''_t \leq 1$, and if $I'_t = 0$, then $|I''_t| \leq p_t$. Using basic algebra, we have, with probability 1:

$$\sum_{t=1}^{T} \mathbb{E}[I''^2_t \mid I'_1, \dots, I'_{t-1}] \leq \sum_{t=1}^{T} p_t + (1 - p_t)p_t^2 \leq C \log T + 3C.$$

Substituting into the Bernstein inequality with

$$k = 2(C \log T + 3C) + \log(1/\delta) \quad \text{and} \quad v = C \log T + 3C,$$

we conclude that, with probability at least $1 - \delta$,

$$\begin{aligned}\sum_{t=1}^{T} I_t = \sum_{t=1}^{T} I'_t \leq \sum_{t=1}^{T} I''_t + \sum_{t=1}^{T} \mathbb{E}[I'_t \mid I'_1, \dots, I'_{t-1}] \\ \leq k + \sum_{t=1}^{T} \min\left\{\frac{C}{T - t + 1}, 1\right\} \leq 3C \log T + 5C + \log(1/\delta).\end{aligned}$$

Here, we used the following elementary inequality:

$$\forall a, b \geq 0, \quad \frac{(2a + b)^2}{2(a + (2a + b)/3)} \geq b.$$

The lemma now follows from the fact that $C \log T \geq 5C$ when $T \geq e^5$ and $C \geq 1$. □

Lemma 7.6 (From error bound to covering)**.** Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary valued class and $\mathbf{err} \in \mathbb{N}^+$. For any $\Omega \subset \mathcal{X}^T$, suppose there exists a prediction rule Φ such that

$$\forall h \in \mathcal{H},\ \forall \mathbf{x}^T \in \Omega,\ \sum_{t=1}^{T} 1\{\Phi(\mathbf{x}^t, h(\{\mathbf{x}^{t-1}\})) \neq h(\mathbf{x}_t)\} \leq \mathbf{err}.$$

Then, there exists a covering set $\mathcal{G} \subset \{0,1\}^{\mathcal{X}^*}$ such that for all $\mathbf{x}^T \in \Omega$ and $h \in \mathcal{H}$ one can find $g \in \mathcal{G}$ that satisfies $g(\mathbf{x}^t) = h(\mathbf{x}_t)$ for all $t \in [T]$, and

$$|\mathcal{G}| \leq \sum_{t=0}^{\mathbf{err}} \binom{T}{t} \leq T^{\mathbf{err}+1}.$$

Proof. The proof essentially replicates that of Lemma 6.3. For any $I \subset [T]$ with $|I| \leq \mathbf{err}$, we construct a function g_I as follows. Let $\mathbf{x}^t$ be the inputs at time t. We define

$$g_I(\mathbf{x}^t) = \begin{cases} 1 - \Phi(\mathbf{x}^t,\, g_I(\{\mathbf{x}^i\}_{i=1}^{t-1})) & \text{if } t \in I, \\ \Phi(\mathbf{x}^t,\, g_I(\{\mathbf{x}^i\}_{i=1}^{t-1})) & \text{otherwise.} \end{cases}$$

where $g_I(\{\mathbf{x}^i\}_{i=1}^{t-1}) = \{g_I(\mathbf{x}^1), \ldots, g_I(\mathbf{x}^{t-1})\}$. We claim that the set $\mathcal{G}$, consisting of all such functions g_I, is the desired covering set. To see this, fix any $h \in \mathcal{H}$ and any sequence $\mathbf{x}^T \in \Omega$. By assumption, we have

$$\sum_{t=1}^{T} \mathbf{1}\{\Phi(\mathbf{x}^t,\, h(\{\mathbf{x}^{t-1}\})) \neq h(\mathbf{x}_t)\} \leq \mathbf{err}.$$

Let $I \subset [T]$ be the set of positions i where $\Phi(\mathbf{x}^i,\, h(\{\mathbf{x}^{i-1}\})) \neq h(\mathbf{x}_i)$, so that $|I| \leq \mathbf{err}$. By construction, it is easy to verify that for all $t \in [T]$, we have $g_I(\mathbf{x}^t) = h(\mathbf{x}_t)$. The upper bound on $|\mathcal{G}|$ follows by counting the number of possible subsets I. □

Proof of Theorem 7.4. Let Φ be the 1-inclusion graph predictor. We have that Φ is permutation invariant, since the nodes in the 1-inclusion graph are determined by subsets of $\mathcal{X}$ that do not depend on the order of elements in the set. By symmetries of *i.i.d.* distributions, for any event $A(\mathbf{x}^T)$ on $\mathbf{x}^T \overset{i.i.d.}{\sim} \boldsymbol{\nu}^T$, we have:

$$\begin{aligned}\Pr[A(\mathbf{x}^T)] &= \mathbb{E}_\sigma[\Pr_{\mathbf{x}^T}[A(\mathbf{x}_{\sigma(1)}, \cdots, \mathbf{x}_{\sigma(T)})]] \\ &= \mathbb{E}_\sigma \mathbb{E}_{\mathbf{x}^T} 1\{A(\mathbf{x}_{\sigma(1)}, \cdots, \mathbf{x}_{\sigma(T)})\} \\ &= \mathbb{E}_{\mathbf{x}^T} \mathbb{E}_\sigma 1\{(A(\mathbf{x}_{\sigma(1)}, \cdots, \mathbf{x}_{\sigma(T)})\} \\ &\leq \sup_{\mathbf{x}^T} \Pr_\sigma[A(\mathbf{x}_{\sigma(1)}, \cdots, \mathbf{x}_{\sigma(T)})],\end{aligned}$$

where the interchange of the expectations follows from Fubini's theorem. It is therefore sufficient to show that for any $\mathbf{x}^T \in \mathcal{X}^T$, w.p. $\geq 1 - \delta$ over a random permutation σ on $[T]$,

$$\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} 1\{\Phi(\mathbf{x}^{\sigma(t)}, h(\{\mathbf{x}^{\sigma(t-1)}\})) \neq h(\mathbf{x}_{\sigma(t)})\} \leq 5\mathsf{VC}(\mathcal{H}) \log T + \log(1/\delta).$$

To see this, we observe that by Sauer's lemma (Lemma 2.1), there are at most $T^{\mathsf{VC}(\mathcal{H})}$ functions of $\mathcal{H}$ restricted on any given $\mathbf{x}^T$. Let now δ in Lemma 7.5 be $\frac{\delta}{T^{\mathsf{VC}(\mathcal{H})}}$ and $C = \mathsf{VC}(\mathcal{H})$. When applying Theorem 7.3 together with a union bound, the error bound w.p. $\geq 1 - \delta$ is of the form $5\mathsf{VC}(\mathcal{H}) \log T + \log(1/\delta)$.

The upper bound for the size of the covering set $\mathcal{G}$ follows from Lemma 7.6 by taking $\Omega \subset \mathcal{X}^T$ to be the set for which Φ makes at most $5\mathsf{VC}(\mathcal{H}) \log T + \log(1/\delta)$ accumulated errors, where $\Pr[\Omega] \geq 1 - \delta$. □

Theorem 7.4 and Theorem 7.1 immediately imply the following regret bound.

Corollary 7.7. Let $\mathcal{H} \subset \{0, 1\}^{\mathcal{X}}$ be a binary valued class with finite VC-dimension, P be the class of all *i.i.d.* distributions over $\mathcal{X}^T$ and $T \geq e^5$. If $\ell(\cdot, y)$ is convex, L-Lipschitz and bounded by 1 for all $y \in \mathcal{Y}$, then:

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq \sqrt{3T\mathsf{VC}(\mathcal{H}) \log^2 T} + O(1).$$

This result recovers [60] but with a worse $\log T$ term. However, our result establishes the (essentially) same result by using a completely different technique. Moreover, our technique can be applied to more general problems than the epoch-based approach of [60].

Indeed, for logarithmic loss, we have the following regret bound:

Corollary 7.8. Let $\mathcal{H}$ be a F-composition class of $\mathcal{H}_1, \cdots, \mathcal{H}_d \subset \{0,1\}^{\mathcal{X}}$ with Θ as in Example 7.3, P be the class of all *i.i.d.* distributions over $\mathcal{X}^T$, and $T \geq e^5$. If ℓ is the logarithmic loss, then:

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq O\left(\left(d + \sum_{i=1}^{d} \mathsf{VC}(\mathcal{H}_i)\right) \log^2 T\right).$$

Proof. Taking $\alpha = \frac{1}{T}$, we note that Θ can be α-covered by a set $\mathcal{C}$ of size upper bounded by T^d under L_1 norm. Applying Proposition 7.2 and Theorem 7.4 and noticing that the composition function F is 1-Lipschitz, there exists a stochastic sequential covering set $\mathcal{G}$ of $\mathcal{H}$ w.r.t. *i.i.d.* processes at scale $\alpha = \frac{1}{T}$ and confidence δ such that:

$$\log |\mathcal{G}| \leq d \log T + \left(5 \sum_{i=1}^{d} \mathsf{VC}(\mathcal{H}_i) \log^2 T\right) + d \log T \log(d/T) + d \log T.$$

The result follows by applying Theorem 7.2 and taking $\alpha = \delta = \frac{1}{T}$. □

For $d = 2$ and $\mathcal{H}$ being the class in Example 7.2, Corollary 7.8 improves upon the $O(\sqrt{T})$ regret bound established by [56]. Moreover, [50] derive an $O\left(\frac{\mathsf{VC}(\mathcal{H}) \log^2 T}{T}\right)$ *risk bound* under log-loss, which can be converted into an $O\left(\mathsf{VC}(\mathcal{H}) \log^3 T\right)$ regret bound using the epoch-based approach of [60]. This bound is off by a factor of $\log T$ compared to our regret bound in Corollary 7.8. In addition, our results apply to the general (worse-case) regret $\tilde{r}_T$, rather than the *average* and *well-specified* regret $\bar{r}_T$ considered in [50], [56].

7.3.2 Tight bounds for classes with finite star number

In the previous section, we demonstrated that the stochastic sequential covering number of finite VC class is upper bounded w.h.p. by $e^{O(\log^2 T)}$. We now show that if we assume additional structure on the class, we can improve the bound to $e^{O(\log T)}$, matching the naive fixed design lower bound for many non-trivial classes. It can be showed that even for 1-dimensional threshold functions the *realizable* cumulative error is lower bounded by $\Omega(\log T)$, thus arguing that the error pattern counting argument as in Lemma 7.6 cannot provide a bound better than $e^{O(\log^2 T)}$.

To resolve this issue, we introduce the notion of *star number* that was used originally in [61] for analyzing the sample complexity of active learning; however, we use it here in a completely different context. For any binary valued class $\mathcal{H}$ and $\mathbf{x}^d \in \mathcal{X}^d$, we say $\mathcal{H}$ Star-shatters $\mathbf{x}^d$ if there exist $h, h_1, \cdots, h_d \in \mathcal{H}$ such that for all $i, j \in [d]$ with $j \neq i$ we have:

$$h(\mathbf{x}_i) \neq h_i(\mathbf{x}_i) \text{ but } h(\mathbf{x}_j) = h_i(\mathbf{x}_j),$$

i.e., a sequence $\mathbf{x}^d$ is Star-shattered by $\mathcal{H}$ if there exists a function $h \in \mathcal{H}$ such that any labeling on $\mathbf{x}^d$, which differs by one position from the realization of h, is also realizable by some function $h_i \in \mathcal{H}$. Such a sequence $\mathbf{x}^d$ is called a *star* set of $\mathcal{H}$. The star number $\mathsf{Star}(\mathcal{H})$ of $\mathcal{H}$ is defined to be the maximum number d such that there exists $\mathbf{x}^d$ that is Star-shattered by $\mathcal{H}$. Clearly, we have $\mathsf{Star}(\mathcal{H}) \geq \mathsf{VC}(\mathcal{H})$ for all $\mathcal{H}$.

We now introduce a new notion of *certification*, which is the key for our following arguments. For any sequence $\mathbf{x}^t$ and $h \in \mathcal{H}$, we say $\mathbf{x}^{t-1}$ certifies $\mathbf{x}_t$ under h if:

$$\forall f \in \mathcal{H}, \text{ if } \forall i \in [t-1],\ f(\mathbf{x}_i) = h(\mathbf{x}_i) \text{ then } f(\mathbf{x}_t) = h(\mathbf{x}_t).$$

We have the following property of finite star number classes w.r.t. certification:

Lemma 7.9. If $\mathcal{H}$ has star number upper bounded by s, then for any $\mathbf{x}^t \in \mathcal{X}^t$ and $h \in \mathcal{H}$ we have:

$$\mathrm{Pr}_\sigma \left[\{\mathbf{x}_{\sigma(1)}, \cdots, \mathbf{x}_{\sigma(t-1)}\} \text{ certifies } \mathbf{x}_{\sigma(t)} \text{ under } h\}\right] \geq 1 - \frac{s}{t},$$

where σ is the uniform random permutation over $[t]$.

Proof. We only need to show that there are at most s points in $\mathbf{x}^t$ that can not be certified by the others under h. Suppose otherwise, that we have $s+1$ such points. Consider the realization of h on these points. By definition of certification, we can find functions $h_1, \cdots, h_{s+1}$ as in the definition of Star-shattering. This contradicts the fact that the star number is upper bounded by s. □

We now prove a high probability bound on the number of non-certified positions for a finite star number class, which is similar to Lemma 7.5.

Lemma 7.10. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a class with a finite star number and $T \geq e^5$. Then, with probability $\geq 1-\delta$ over $\mathbf{x}^T$ (sampled from some $i.i.d.$ distribution over $\mathcal{X}^T$) for all $h \in \mathcal{H}$:

$$\sum_{t=1}^{T} \mathbf{1}\left\{\mathbf{x}^{t-1} \text{ does not certify } \mathbf{x}_t \text{ under } h\right\} \leq$$
$$\mathsf{VC}(\mathcal{H}) \log T + 4\,\mathsf{Star}(\mathcal{H}) \log T + \log(1/\delta).$$

Proof. Note that the event $\{\mathbf{x}^{t-1}$ does not certify $\mathbf{x}_t$ under $h\}$ can be viewed as the event $\{\Phi$ makes an error at step $t\}$ as in Lemma 7.5 (since certification is permutation invariant). By Lemma 7.9 and Lemma 7.5 with $C = \mathsf{Star}(\mathcal{H})$, we have, for all $h \in \mathcal{H}$ and $\mathbf{x}^T \in \mathcal{X}^T$ w.p. $\geq 1-\delta$ over uniform random permutation σ on $[T]$:

$$\sum_{t=1}^{T} 1\{\mathbf{x}^{\sigma(t-1)} \text{ does not certify } \mathbf{x}_{\sigma(t)} \text{ under } h\} \leq 4\mathsf{Star}(\mathcal{H}) \log T + \log(1/\delta).$$

The result follows from a similar path as in the proof of Theorem 7.4 □

Lemma 7.10 allows us to construct the sequential covering set explicitly without relying on error pattern counting as shown next.

Theorem 7.11. Let $\mathcal{H}$ be a binary valued class with finite star number. Then, there exists a stochastic sequential covering set $\mathcal{G}$ of $\mathcal{H}$ w.r.t. the class of all $i.i.d.$ distributions over $\mathcal{X}^T$ at scale $\alpha = 0$ and confidence δ such that for $T \geq e^5$:

$$\log |\mathcal{G}| \leq 5\mathsf{Star}(\mathcal{H}) \log T + \log(1/\delta).$$

Proof. We will construct a covering set $\mathcal{G}$ directly without relying on the error pattern counting as in Lemma 7.6. This is the key to removing the extra $\log T$ factor. We will introduce a set K to index the functions in $\mathcal{G}$, we assume that K is fixed and $|K| = 2^M$ for some M to be chosen later. For any $k \in K$, we will construct a *sequential function* g_k as follows:

Let $\mathbf{x}^T$ be a realization of the sample from an $i.i.d.$ source. The realization tree $\mathcal{T}$ of $\mathcal{H}$ on $\mathbf{x}^T$ is a leveled binary tree of depth $T+1$, with each node at level t being labeled $\mathbf{x}_t$ (where level 1 has only the root v_1), each left edge being labeled 0 and each right edge being labeled 1, such that any node $v_t \in \mathcal{T}$ at level t has left (respectively right) child if and only if there exist $h \in \mathcal{H}$ such that $h(\mathbf{x}_t) = 0$ (respectively $h(\mathbf{x}_t) = 1$) and $h(\mathbf{x}_i) = L(v_i \to v_{i+1})$ for all $i \leq t-1$, where $v_1 \to v_2 \to \cdots \to v_t = v$ is the path from root v_1 to v and L is the edge label function. Note that different realizations of $\mathbf{x}^T$ will result in different realization trees.

We now assign values of the functions g_k with $k \in K$ using the following procedure. For any node v in the realization tree $\mathcal{T}$, we will associate a set $\mathcal{K}(v) \subset K$ using the following rule (starting from root):

1. If v is the root, then $\mathcal{K}(v) = K$;
2. If v has only one child u, then $\mathcal{K}(u) = \mathcal{K}(v)$;
3. If v has two children u_1, u_2, we assign the sets to u_1, u_2 being an arbitrary partition of $\mathcal{K}(v)$ of *equal* sizes, i.e., $|\mathcal{K}(u_1)| = |\mathcal{K}(u_2)|$, $\mathcal{K}(u_1) \cap \mathcal{K}(u_2) = \emptyset$ and $\mathcal{K}(u_1) \cup \mathcal{K}(u_2) = \mathcal{K}(v)$.

Clearly, the value $\mathcal{K}(v)$ for any node v at level t can be determined with only the realization of $\mathbf{x}^t$ and the values of $\mathcal{K}$ of all nodes at level t form a partition of K. The procedure $\mathcal{K}$ fails if there exists some node v with two children such that $|\mathcal{K}(v)| < 2$. Suppose the procedure $\mathcal{K}$ does not fail. We have for any $k \in K$, there exists a unique path $v_1 \to v_2 \to \cdots \to v_{T+1}$ with v_1 being the root, such that for all $t \leq T+1$ we have $k \in \mathcal{K}(v_t)$. For any such k, we assign the value of g_k on $\mathbf{x}^t$ as:

$$g_k(\mathbf{x}^t) = L(v_t \to v_{t+1}),$$

where L is the edge label function as discussed above. If the procedure $\mathcal{K}$ fails at some node v_t, we assign the value of $g_k(\mathbf{x}^j)$ arbitrarily for $j \geq t$.

By definition of the realization tree, for any $h \in \mathcal{H}$ there must be a unique path $v_1 \to \cdots \to v_{T+1}$, with v_1 being root such that $h(\mathbf{x}_t) = L(v_t \to v_{t+1})$ for all t. Therefore, if the procedure $\mathcal{K}$ does not fail, then for $k \in \mathcal{K}(v_{T+1})$, we have $h(\mathbf{x}_t) = g_k(\mathbf{x}^t)$ for all $t \leq T$ by definition of g_k. We now show that by setting $M = \lceil 5\mathsf{Star}(\mathcal{H}) + \log(1/\delta) \rceil$, w.p. $\geq 1-\delta$ over $\mathbf{x}^T$, the procedure $\mathcal{K}$ will not fail, thus proving that the class $\mathcal{G} = \{g_k : k \in K\}$ is a stochastic sequential covering of $\mathcal{H}$ with confidence δ. To see this, we note that the procedure $\mathcal{K}$ fails at node v_t at level t if and only if there are $\geq M+1$ nodes with two children in the (unique) path $v_1 \to \cdots \to v_t$, where v_1 is root, since only rule 3 will reduce the size of value of $\mathcal{K}$ by 1/2. Assume now the procedure $\mathcal{K}$ fails at node v_t. Let $h \in \mathcal{H}$ be a function such that $h(\mathbf{x}_i) = L(v_i \to v_{i+1})$ for all $i \leq t$, which must exist by definition of realization tree. Since any node v_j in the path $v_1 \to \cdots \to v_t$ with two children implies $\mathbf{x}^{j-1}$ *does not* certify $\mathbf{x}_j$ under h, we have that there are at least $M+1$ positions j (with $j \leq t$) such that $\mathbf{x}^{j-1}$ does not certify $\mathbf{x}_j$ under h. By Lemma 7.10 and selection of M, this happens with probability $\leq \delta$. This completes the proof. □

	h_1	h_2	h_3	h_4	h_5
$\mathbf{x}_1$	0	0	0	1	1
$\mathbf{x}_2$	0	0	1	1	1
$\mathbf{x}_3$	0	1	0	0	1

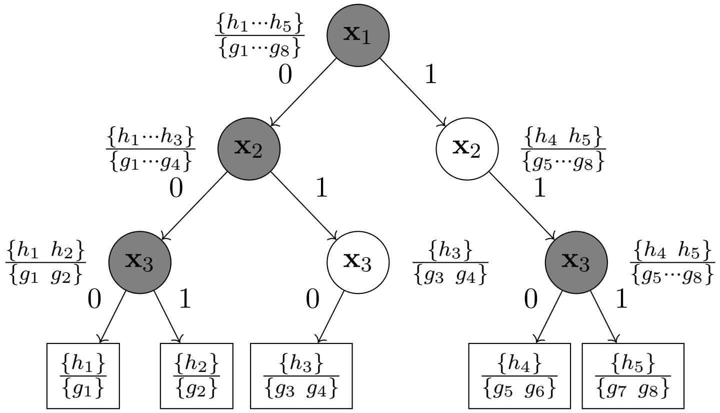

$\frac{\{h_1 \cdots h_5\}}{\{g_1 \cdots g_8\}}$
$\mathbf{x}_1$
0
1
$\frac{\{h_1 \cdots h_3\}}{\{g_1 \cdots g_4\}}$
$\mathbf{x}_2$
$\mathbf{x}_2$
$\frac{\{h_4\ h_5\}}{\{g_5 \cdots g_8\}}$
0
1
1
$\frac{\{h_1\ h_2\}}{\{g_1\ g_2\}}$
$\mathbf{x}_3$
$\mathbf{x}_3$
$\frac{\{h_3\}}{\{g_3\ g_4\}}$
$\mathbf{x}_3$
$\frac{\{h_4\ h_5\}}{\{g_5 \cdots g_8\}}$
0
1
0
0
1
$\frac{\{h_1\}}{\{g_1\}}$
$\frac{\{h_2\}}{\{g_2\}}$
$\frac{\{h_3\}}{\{g_3\ g_4\}}$
$\frac{\{h_4\}}{\{g_5\ g_6\}}$
$\frac{\{h_5\}}{\{g_7\ g_8\}}$

Figure 7.1: Realization tree of $\mathcal{H}$ defined by the table above and partitioning of $\mathcal{G}$.

Example 7.5. We illustrate the construction of the realization tree in this example. We set $\mathcal{H} = \{h_1, \cdots, h_5\}$, as shown in the table of Figure 7.1 with sample $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$. The realization tree is shown in Figure 7.1, where each function $h \in \mathcal{H}$ is consistent with some path of the tree, and each path has some function $h \in \mathcal{H}$ consistent with it. We assign a subset of $\mathcal{G}$ for each node in the tree denoted as $\{\cdot\}$. Observe that if a node has only one child then the child has the same assigned set as the parent, else we assign an arbitrary partition of the parent set with *equal* sizes to its two children. The final partitions of the set $\mathcal{G} = \{g_1, \cdots, g_8\}$ are in the leaf nodes of the tree. In the figure, binary nodes (i.e., nodes with two children) are in gray color. The maximum number of binary nodes in any path is 3, by selecting $|\mathcal{G}| \geq 2^3 = 8$, which guarantees that the assigning procedure does not fail until the leaf. Each g_k is associated with a unique path from root to the leaf with (the only) assigned sets on the nodes that contain g_k. The values of g_k are defined to be the labels of out edges along the path in the obvious way. One can verify that g_1 covers h_1, g_2 covers h_2, g_3, g_4 covers h_3, g_5, g_6 covers h_4, and g_7, g_8 covers h_5. Generally, by Lemma 7.10 the number of binary nodes in any path is of order $O(\log T)$ with high probability (i.e., setting $|\mathcal{G}| = 2^{O(\log T)}$ ensures the process success w.h.p.).

Corollary 7.12. Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be the F-composition class as in Example 7.2 with $\mathcal{H}_1 \subset \{0,1\}^{\mathcal{X}}$ being a class of finite star number, P being the class of all *i.i.d.* distributions over $\mathcal{X}^T$. If ℓ is the Log-loss, then:

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq O(\mathsf{Star}(\mathcal{H}_1) \log T).$$

Proof. By Proposition 7.2 and Theorem 7.11, $\mathcal{H}$ admits a stochastic sequential covering set $\mathcal{G}$ at scale α and confidence δ such that $\log |\mathcal{G}| \leq 2\log(1/\alpha) + 5\mathsf{Star}(\mathcal{H}_1)\log T + \log(1/\delta)$. Taking $\alpha = \delta = \frac{1}{T}$ and applying Theorem 7.2, the result follows. □

Note that a natural class that has finite star number is the threshold functions $\mathcal{H} = \{1\{x \geq a\} : x, a \in [0,1]\}$, which has star number 2. Corollary 7.12 implies the regret under Log-loss is upper bounded by $O(\log T)$. We refer to [61] for more non-trivial examples.

We note also that the $O(\log T)$ regret bound is not likely to be established by the epoch based approach (which [50], [56], [60] have used to establish their regret bounds), since the epochs will inevitably introduce an additional $\log T$ factor.

However, having a finite star number is not a necessary condition for achieving a covering number of size $e^{O(\log T)}$. To see this, consider the hypothesis class that labels exactly one input as 1 and assigns label 0 to all others. This class admits a worst-case sequential cover of size $T + 1 = e^{O(\log T)}$ (cf. Chapter 6.1.2), yet its star number is infinite.

7.3.3 Tight bounds with finite Star-Littlestone dimension

In this section, we introduce a new complexity measure that we call *Star-Littlestone* dimension. The main purpose of this measure is to incorporate the star number and Littlestone dimension that goes beyond simple finite star number, and allows us to expand the class of $\mathcal{H}$ with $e^{O(\log T)}$ cover.

Definition 7.2 (Star-Littlestone dimension)**.** Let $\bigcup_{i=0}^{d}\{0,1\}^i$ denote the set of all binary sequences of length at most d. For any integers d and s, we say that a binary tree $\tau : \bigcup_{i=0}^{d}\{0,1\}^i \to \mathcal{X}$ is *Star-Littlestone shattered* by $\mathcal{H}$ at star scale s if for every path $\epsilon^d \in \{0,1\}^d$, the star number $\mathsf{Star}(\mathcal{H}_{\epsilon^d}) > s$, where

$$\mathcal{H}_{\epsilon^d} = \left\{ h \in \mathcal{H} : \forall t \in [d],\ h(\tau(\epsilon^{t-1})) = \epsilon_t \right\}.$$

In words, Star-Littlestone shattering means that for every path in the tree, the class of hypotheses consistent with the path has star number greater than s. The *Star-Littlestone dimension* $\mathsf{SL}(s)$ of $\mathcal{H}$ at star scale s is defined as the largest d such that there exists a depth-d tree τ that is Star-Littlestone shattered at scale s by $\mathcal{H}$.

Applying Theorem 7.11 and the SOA argument as in [24], we establish our next main theorem.

Theorem 7.13. Let $\mathcal{H}$ be a binary valued class with Star-Littlestone dimension $\mathsf{SL}(s)$ at star scale s. Then, there exists a stochastic sequential covering set $\mathcal{G}$ of $\mathcal{H}$ w.r.t. the class of all *i.i.d.* distributions over $\mathcal{X}^T$ at scale $\alpha = 0$ and confidence δ such that:

$$\log |\mathcal{G}| \leq O(\max\{\mathsf{SL}(s) + 1, s\} \log T + \log(1/\delta)).$$

Proof. The proof will incorporate the SOA argument as in [24] and the result from Theorem 7.11. For notational convenience, we denote

$d = \mathsf{SL}(s) + 1$. For any $I \subset [T]$ with $|I| \leq d$, we will construct a set $\mathcal{G}_I$. Let Φ be the SOA predictor (cf. Algorithm 6.1) that predicts the label for which the remaining consistent subclass has maximum Star-Littlestone dimension at star scale s, if both subclasses have SL dimension 0 we predict the label for which the remaining consistent subclass has maximum star number (and break ties arbitrarily). We now construct functions in $\mathcal{G}_I$ as follows. The predictions of functions in $\mathcal{G}_I$ are partitioned into 2 phases (start with phase 1). At phase 1, all the functions in $\mathcal{G}_I$ use the same prediction rule as in Lemma 7.6, that is, if we are at time step $t \in I$, we predict using $1-\Phi$, else we use Φ to predict, where Φ is the SOA prediction rule described above. We enter phase 2 if the remaining consistent class has star number upper bounded by s; we then construct the prediction functions in $\mathcal{G}_I$ as in Theorem 7.11 with $\mathsf{Star}(\mathcal{H}) = s$, confidence δ/T^{d+1} and $|\mathcal{G}_I| \leq e^{5s\log T + \log(T^{d+1}/\delta)}$. The covering class $\mathcal{G}$ is defined to be:

$$\mathcal{G} = \bigcup_{I\subset[T],\ |I|\leq d} \mathcal{G}_I.$$

By Theorem 7.11 with $\mathsf{Star}(\mathcal{H}) = s$ and $\delta = \delta/T^{d+1}$ and computing the number of Is, we have

$$|\mathcal{G}| \leq T^{d+1} e^{5s\log T + \log(T^{d+1}/\delta)} \leq e^{O(\max\{d,s\}\log T + \log(1/\delta))}.$$

We now show that $\mathcal{G}$ is indeed a stochastic sequential covering of $\mathcal{H}$ with confidence δ. Let $\mathcal{H}_I$ be the (*random*) subclass of functions in $\mathcal{H}$ that are consistent with Φ with error pattern I before entering phase 2 [2] (it is possible that h remains on phase 1 until time T). Note that all functions in $\mathcal{H}_I$ agree on samples at phase 1. Note also that, with probability 1 we have $\mathcal{H} = \bigcup_{I\subset[T],|I|\leq d} \mathcal{H}_I$. To see this, we note that if h disagreed with the SOA then the remaining consistent class has $\mathsf{SL}(s)$ decreased by at least 1 (similar to Lemma 6.2) or has star number $\leq s$ if the current consistent class has $\mathsf{SL}(s) = 0$. This implies that any $h \in \mathcal{H}$ can be disagreed with SOA at most d times before entering phase 2, which must be in some $\mathcal{H}_I$ with $|I| \leq d$. Now, for any I with $|I| \leq d$ we

[2] Here, phase 1 and 2 corresponds to that the functions in $\mathcal{H}$ consistent with h on current sample has star number $> s$ and $\leq s$, respectively.

need to show that:

$$\Pr[\mathcal{G}_I \text{ covers } \mathcal{H}_I] \geq 1 - \frac{\delta}{T^{d+1}}.$$

Note that the main difficulty here is that $\mathcal{H}_I$ is a *random* subset. We show that conditioning on any realization of $\mathcal{H}_I$, the above inequality holds (the inequality will then hold by law of total probability). This follows from Theorem 7.11 by noticing that the samples in phase 2 are still *i.i.d.* and independent of samples in phase 1, and $\mathcal{G}_I$ trivially covers $\mathcal{H}_I$ in phase 1 by definition of $\mathcal{G}_I$ and $\mathcal{H}_I$. The theorem will now follow by a union bound on all the Is. □

Example 7.6. In this example, we show a class $\mathcal{H}$ that has both infinite star number and Littlestone dimension but finite Star-Littlestone dimension. Let $\mathcal{H} = \{h_{[a,b]}(x) = 1\{x \in [a,b]\} : [a,b] \subset [0,1]\}$ be the indicators of intervals. It is easy to see that $\mathcal{H}$ has both infinite star number and Littlestone dimension. However we can show that it has Star-Littlestone dimension 0 at star scale 4. To see this, consider any point $x \in [0,1]$ and the hypothesis class $\mathcal{H}_1 = \{h \in \mathcal{H} : \ h(x) = 1\}$. We show that the star number of $\mathcal{H}_1$ is ≤ 4. For any 5 points in $[0,1]$, there must be at least 3 points on the same side relative to x, the restriction of $\mathcal{H}_1$ on such points is equivalent to threshold functions (either of form $1\{x \geq a\}$ or $1\{x \leq b\}$), thus it cannot Star-shatter these 3 points. This implies that the global sequential covering size of $\mathcal{H}$ is upper bounded by $e^{O(\log T)}$ as in Theorem 7.13.

Example 7.7. Let

$$\mathcal{H} = \left\{ h_B(\mathbf{x}) = 1\{\mathbf{x} \in B\} : B = \prod_{i=1}^{d} [a_i, b_i] \subset \mathbb{R}^d \right\}$$

be the class of indicators of rectangular cuboids in $\mathbb{R}^d$. Note that $\mathcal{H}$ has infinite Star-Littlestone dimension for any finite star scale when $d \geq 2$ and the VC-dimension of $\mathcal{H}$ is upper bounded by $O(d)$. By Example 7.4, we have $\mathcal{H}$ can be expressed as a function in terms of indicators of intervals. Applying Proposition 7.2 and Example 7.6 we obtain a covering set $\mathcal{G}$ of $\mathcal{H}$ with $\log |\mathcal{G}| \leq O(d \log T + d \log(d/\delta))$. This implies a regret bound of mixable losses (including logarithmic loss) of order $O(d \log T + d \log d)$.

Remark 7.1. We leave it as an open problem to determine if the upper bound $e^{O(\log T)}$ can be achieved for any finite VC-dimensional class. Establishing such a result even for the threshold functions $\mathcal{H} = \{h_{\mathbf{w}}(\mathbf{x}) = 1\{\langle \mathbf{w}, \mathbf{x} \rangle \geq a\} : \mathbf{w}, \mathbf{x} \in \mathbb{R}^d,\ a \in \mathbb{R}\}$ with $d \geq 2$ seems to be a hard task.

7.4 Real Valued Class with Finite Fat-shattering

We have established tight stochastic sequential covering bounds for finite VC classes in the previous section. We now consider the more general setting where $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ is a class of real-valued functions taking values in $[0,1]$, and assume that $\mathcal{H}$ has a bounded fat-shattering dimension.

We first recall the notion of fat-shattering number, already discussed in Chapter 2.2. For any class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, we say $\mathcal{H}$ α-fat shatters $\mathbf{x}^d \in \mathcal{X}^d$ if there exists $s^d \in [0,1]^d$ such that for all $I \subset [d]$ there exists $h \in \mathcal{H}$ such that for all $t \in [d]$: (i) If $t \in I$, then $h(\mathbf{x}_t) \geq s_t + \alpha$; and (ii) If $t \notin I$, then $h(\mathbf{x}_t) \leq s_t - \alpha$. Then, the fat shattering number of $\mathcal{H}$ at scale α is defined to be the maximum number $d := d(\alpha)$ such that there exists $\mathbf{x}^d \in \mathcal{X}^d$ with $\mathcal{H}$ α-fat shatters $\mathbf{x}^d$.

We now state our main result for this section.

Theorem 7.14. Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to [0,1]$ with the α-fat shattering number $d(\alpha)$. Then there exists a stochastic global sequential covering set $\mathcal{G}$ of $\mathcal{H}$ w.r.t. the class of all *i.i.d.* distributions over $\mathcal{X}^T$ at scale α and confidence δ such that:

$$\log|\mathcal{G}| \leq O(d(\alpha/32)(\log T \log(4/\alpha))^4 \\ + (\log^2 T + \log T \log(4/\alpha)) \log(\log T/\delta)),$$

where O hides absolute constant which is independent of α, T, and δ.

We first introduce the notion of local α-covering. We say that a class $\mathcal{F}$ locally α-covers $\mathcal{H}$ at $\mathbf{x}^T \in \mathcal{X}^T$ if for all $h \in \mathcal{H}$ there exists $f \in \mathcal{F}$ such that:

$$\forall t \in [T],\ |h(\mathbf{x}_t) - f(\mathbf{x}_t)| \leq \alpha.$$

Here, we also assume that $\mathcal{F} \subset \mathcal{H}$ (we can always convert α-covering set $\mathcal{F}$ of $\mathcal{H}$ to a 2α-covering set $\tilde{\mathcal{F}} \subset \mathcal{H}$ such that $|\tilde{\mathcal{F}}| \leq |\mathcal{F}|$).

The following lemma upper bounds the local α-covering size w.r.t. the α-fat shattering number of $\mathcal{H}$, which is due to [62].

Lemma 7.15. Suppose the α fat-shattering number of $\mathcal{H}$ is $d(\alpha)$. Then for all $\mathbf{x}^T \in \mathcal{X}^T$, there exists $\mathcal{F}$ (which depends on $\mathbf{x}^T$) that locally α-covers $\mathcal{H}$ at $\mathbf{x}^T$ such that:

$$|\mathcal{F}| \leq 2\left(T\left(\frac{2}{\alpha}+1\right)^2\right)^{\lceil d(\alpha/4)\log\left(\frac{2eT}{\alpha d(\alpha/4)}\right)\rceil} \leq 2^{d(\alpha/4)(\log^2 T + 2\log^2(1/\alpha)+O(1))}.$$

Our proof of Theorem 7.14 is based on the following key lemma, which is an application of the classical symmetrization argument and an epoch approach similar to [60].

Lemma 7.16. Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be a class with α-fat shattering number $d(\alpha)$. Let S_1, S_2 be two *i.i.d.* samples from the same distribution over $\mathcal{X}$, both of size k. For any S_i with $i \in \{1,2\}$, we define a distance for all $h_1, h_2 \in \mathcal{H}$ as:

$$d^{\alpha}_{S_i}(h_1,h_2) = \sum_{s\in S_i} 1\{|h_1(s)-h_2(s)| \geq \alpha\}.$$

Then

$$\Pr_{S_1,S_2}\left[\exists h_1,h_2 \in \mathcal{H} \ s.t.\ d^{\alpha}_{S_1}(h_1,h_2)=0 \text{ and } d^{4\alpha}_{S_2}(h_1,h_2)\geq r\right] \leq 2^{\tilde{O}(d(\alpha/8))-r}$$

where $\tilde{O}(d(\alpha/8)) = 2d(\alpha/8)(\log^2 k + 2\log^2(1/\alpha)+O(1))$.

Proof. We use a symmetrization argument. We denote by A the event that $\exists h_1, h_2 \in \mathcal{H}$ such that $d^{\alpha}_{S_1}(h_1,h_2)=0$ but $d^{4\alpha}_{S_2}(h_1,h_2)\geq r$. Let σ be a random permutation that switches the ith positions of S_1, S_2 w.p. $\frac{1}{2}$ and independently for different $i \in [k]$. By symmetries, it is sufficient to fix S_1, S_2 and upper bound $\Pr_{\sigma}[A[\sigma(S_1,S_2)]]$. By Lemma 7.15, we know that there exists a set $\mathcal{F}$ that $\alpha/2$-covers $\mathcal{H}$ on $S_1 \cup S_2$ with:

$$|\mathcal{F}| \leq 2^{d(\alpha/8)(\log^2 k + 2\log^2(1/\alpha)+O(1))}.$$

If the event A happens, then there exist $f_1, f_2 \in \mathcal{F}$ such that (using property of covering):

$$d^{2\alpha}_{S_1}(f_1,f_2)=0 \text{ but } d^{3\alpha}_{S_2}(f_1,f_2)\geq r.$$

Clearly, in order for A to happen, any position $s \in S_2$ such that $|f_1(s) - f_2(s)| \geq 3\alpha$ must not be switched to S_1 under σ, which happens with probability upper bounded by 2^{-r}. Applying union bound over all pairs of $\mathcal{F}$, we have

$$\Pr_{S_1,S_2}[A] \leq 2^{2d(\alpha/8)(\log^2 k + 2\log^2(1/\alpha) + O(1)) - r}$$

which completes the proof. □

Proof of Theorem 7.14. We partition the time horizon into epochs, where each epoch s ranges from time step $2^{s-1}, \cdots, 2^s - 1$. For each epoch s, we will construct a covering set $\mathcal{G}_s$. The global covering set $\mathcal{G}$ will be constructed by considering all the combinations of functions in $\mathcal{G}_s$ with $s \in \{1, \cdots, \lceil \log T \rceil\}$.

For any epoch s, we construct $\mathcal{G}_s$ as follows. Let $\mathcal{F} \subset \mathcal{H}$ be the local α-covering set on the samples $\mathbf{x}_1, \cdots, \mathbf{x}_{2^{s-1}-1}$. By Lemma 7.15, we have

$$|\mathcal{F}| \leq 2^{d(\alpha/4)(s^2 + 2\log^2(1/\alpha) + O(1))}.$$

Let

$$r_s = 2d(\alpha/8)(s^2 + 2\log^2(1/\alpha) + O(1)) + \log(\log T/\delta).$$

By Lemma 7.16 w.p. $\geq 1 - \frac{\delta}{\log T}$ for any $h \in \mathcal{H}$ there exists $f \in \mathcal{F}$ such that f 4α-covers h on samples $\mathbf{x}_{2^{s-1}}, \cdots, \mathbf{x}_{2^s-1}$ except r_s positions (the $f \in \mathcal{F}$ that α-covers h on $\mathbf{x}^{2^{s-1}-1}$ is the desired function since $\mathcal{F}$ is a local α-covering). Let J be a discretization of interval $[0, 1]$ such that for any $a \in [0, 1]$, there exists $b \in J$ so that $|a - b| \leq 4\alpha$. We have $|J| \leq \lceil \frac{1}{8\alpha} \rceil$. Now, for any $I \subset \{2^{s-1}, \cdots, 2^s - 1\}$ with $|I| \leq r_s$, $\{k_i\}_{i \in I} \in J^{|I|}$ and $f \in \mathcal{F}$, we construct a function $f_{I,k^{|I|}}$ as follows:

1. If $t \in I$, we set $f_{I,k^{|I|}}(\mathbf{x}_t) = k_t$;

2. If $t \notin I$, we set $f_{I,k^{|I|}}(\mathbf{x}_t) = f(\mathbf{x}_t)$.

The class $\mathcal{G}_s$ is defined as the class of all such $f_{I,k^{|I|}}$. By definition of r_s and by Lemma 7.16, we have w.p. $\geq 1 - \frac{\delta}{\log T}$, for all $h \in \mathcal{H}$ there exists $g \in \mathcal{G}_s$ such that for all $t \in \{2^{s-1}, \cdots, 2^s - 1\}$ we have:

$$|g(\mathbf{x}_t) - h(\mathbf{x}_t)| \leq 4\alpha.$$

We observe that:

$$|\mathcal{G}_s| \leq |\mathcal{F}| \cdot (2^s|K|)^{r_s+1} \leq 2^{O(d(\alpha/8)((s\log(1/\alpha))^3)+(s+\log(1/\alpha))\log(\log T/\delta))}.$$

We now construct the global covering set $\mathcal{G}$ as follows. For any index $(j_1, \cdots, j_{\lceil \log T\rceil})$ with $j_s \in [|\mathcal{G}_s|]$, we define a function g such that it uses the j_s function in $\mathcal{G}_s$ to make prediction during epoch s. By union bound on the epochs, we have w.p. $\geq 1-\delta$ for any $h \in \mathcal{H}$, there exists g such that:

$$\forall t \in [T],\ |h(\mathbf{x}_t) - g(\mathbf{x}^t)| \leq 4\alpha.$$

This implies that $\mathcal{G}$ is a 4α global sequential covering set of $\mathcal{H}$. Thus

$$|\mathcal{G}| = \prod_{s=1}^{\lceil \log T\rceil} |\mathcal{G}_s| \leq 2^{O(d(\alpha/8)(\log T\log(1/\alpha))^4+(\log^2 T+\log T\log(1/\alpha))\log(\log T/\delta))}.$$

The result follows by taking α in the above expression to be $\alpha/4$. □

We complete this section with two results regarding the expected worst case minimax regret.

Corollary 7.17. Let $\mathcal{H}$ be a [0,1]-valued class with α-fat shattering number of order α^{-l} for some $l \geq 0$, and P be a class of all *i.i.d.* distributions over $\mathcal{X}^T$. If $\ell(\cdot, y)$ is convex, L-Lipschitz and bounded by 1 for all $y \in \mathcal{Y}$, then:

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq \tilde{O}((LT)^{(l+1)/(l+2)})$$

where $\tilde{O}$ hides a poly-log factor.

Proof. Apply Theorem 7.14 to Theorem 7.1 to find

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq \inf_{0\leq\alpha\leq 1} \left\{\alpha LT + \tilde{O}\left(\sqrt{T\alpha^{-l}}\right)\right\}$$

and taking $\alpha = (LT)^{-1/(l+2)}$ finishes the proof. □

Note that [63, Theorem 3] demonstrated that for *known i.i.d.* processes one can achieve an $\tilde{O}(T^{(l-1)/l})$ regret bound (in fact they establish the result for the *smooth adversary* processes). However, extending such an chaining based argument to our *unknown i.i.d.* processes as in Corollary 7.17 seems to be an non-trivial task, since for unknown *i.i.d.* processes one cannot express the expected worst case regret in the iterated minimax formulation as in [25]. We leave it as an open problem to determine if the bound in Corollary 7.17 is tight or not for the unknown *i.i.d.* processes.

Corollary 7.18. Let $\mathcal{H}$ be a [0,1]-valued class with α-fat shattering number of order α^{-l} with $l \geq 0$, and P be the class of all *i.i.d.* distributions over $\mathcal{X}^T$. If ℓ is Log-loss, then $\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq \tilde{O}(T^{l/l+1})$.

Proof. Applying Theorem 7.14 to Theorem 7.2, we have $\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq \inf_{0 \leq \alpha \leq 1} \left\{2\alpha T + \tilde{O}(\alpha^{-l})\right\}$, and taking $\alpha = T^{-1/(l+1)}$ completes the proof. □

We can show that the regret bound in Corollary 7.18 is actually *tight* upto poly-logarithmic factors for *general* classes of α-fat shattering number of order α^{-l} (with $l \geq 1$), see Proposition 7.3 in Section 7.5. However, it is known by [36], [51] that this bound is not tight for *all* classes even for the adversary case.

7.5 Lower Bounds For Regret

We now provide a general approach for lower bounding the regret $\tilde{r}(\mathcal{H}, \mathsf{P})$ using the fixed design regret defined in (4.5) and analyzed in Chapters 5 and 6. We will assume throughout this section that $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ is a general real valued function class and P is the class of all *i.i.d.* processes over $\mathcal{X}^T$. We first introduce the following well known tail bound for the coupon collector problem, see e.g. [64, Theorem 1.9.2].

Lemma 7.19. Let $X_1, X_2, \cdots$ be *i.i.d.* samples from the uniform distribution over $[T]$, and ρ be the first time such that $[T] \subset X_1^\rho$. Then for any $c \geq 0$ we have $\Pr[\rho \geq T \log T + cT] \leq e^{-c}$.

For any function Φ that maps sequences from $\mathcal{X}^*$ to $\mathbb{R}$, we say Φ is monotone if for any $\mathbf{x}^T \subset \mathbf{z}^{T_1}$ we have $\Phi(\mathbf{x}^T) \leq \Phi(\mathbf{z}^{T_1})$, where $\mathbf{x}^T \subset \mathbf{z}^{T_1}$ means that for any $\mathbf{s} \in \mathcal{X}$, the number of $\mathbf{s}$ appearances in $\mathbf{x}^T$ is no more than the number of appearances of $\mathbf{s}$ in $\mathbf{z}^{T_1}$. We also assume a regularity condition for the loss ℓ such that for all $\hat{y}_1, \hat{y}_2 \in \mathcal{Y}$ there exists $y \in \mathcal{Y}$ with $\ell(\hat{y}_1, y) \geq \ell(\hat{y}_2, y)$. We also recall that $r_T^*(\mathcal{H}) = \sup_{\mathbf{x}^T} r^*(\mathcal{H}|\mathbf{x}^T)$.

Theorem 7.20. Let $\mathcal{H}$ be any $[0,1]$-valued class. If the fixed design regret $r_T^*(\mathcal{H}) = \max_{\mathbf{x}^T} r_T^*(\mathcal{H}|\mathbf{x}^T)$, where $r_T^*(\mathcal{H} \mid \mathbf{x}^T)$ defined in (4.5), is monotone over $\mathbf{x}^T$ and ℓ satisfies the above regularity condition, then:

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \geq (1 - O(1/\log T)) r^*_{\kappa^{-1}(T)}(\mathcal{H}) \geq (1 - O(1/\log T)) r^*_{(T/\log T)}(\mathcal{H}),$$

where P is the class of all *i.i.d.* distributions over $\mathcal{X}^T$ and $\kappa(T) = T\log T + T \log\log T$.

Proof. Let $\tilde{\mathbf{x}}^T$ be the feature that achieves the maximum of $r_T^*(\mathcal{H} \mid \tilde{\mathbf{x}}^T)$ (i.e., $r_T^*(\mathcal{H})$). We define the distribution ν to be the uniform distribution over $\{\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_T\}$ (with possibly repeated elements). Let $T_1 = T\log T + T\log\log T$. We have

$$\begin{aligned}
\tilde{r}_{T_1}(\mathcal{H}, \mathsf{P}) &= \inf_{\phi^{T_1}} \sup_{\boldsymbol{\nu}^T \in \mathsf{P}} \mathbb{E}_{\mathbf{x}^{T_1} \sim \boldsymbol{\nu}^T} \left[\sup_{y^{T_1}} \left(\sum_{t=1}^{T_1} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T_1} \ell(h(\mathbf{x}_t), y_t) \right) \right] \\
&\geq \inf_{\phi^{T_1}} \mathbb{E}_{\mathbf{x}^{T_1} \sim \nu^{T_1}} \left[\sup_{y^{T_1}} \left(\sum_{t=1}^{T_1} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T_1} \ell(h(\mathbf{x}_t), y_t) \right) \right] \\
&\overset{(a)}{\geq} \inf_{\phi^{T_1}} \Pr[\tilde{\mathbf{x}}^T \subset \mathbf{x}^{T_1}] \cdot \mathbb{E} \left[\sup_{y^{T_1}} \left(\sum_{t=1}^{T_1} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T_1} \ell(h(\mathbf{x}_t), y_t) \right) \mid \tilde{\mathbf{x}}^T \subset \mathbf{x}^{T_1} \right] \\
&\overset{(b)}{\geq} \Pr[\tilde{\mathbf{x}}^T \subset \mathbf{x}^{T_1}] \cdot \mathbb{E} \left[\inf_{\phi^{T_1}} \sup_{y^{T_1}} \left(\sum_{t=1}^{T_1} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T_1} \ell(h(\mathbf{x}_t), y_t) \right) \mid \tilde{x}^T \subset \mathbf{x}^{T_1} \right] \\
&= \Pr[\tilde{\mathbf{x}}^T \subset \mathbf{x}^{T_1}] \cdot \mathbb{E} \left[r^*_{T_1}(\mathcal{H} \mid \mathbf{x}^{T_1}) \mid \tilde{\mathbf{x}}^T \subset \mathbf{x}^{T_1} \right] \\
&\overset{(c)}{\geq} \Pr[\tilde{\mathbf{x}}^T \subset \mathbf{x}^{T_1}] r_T^*(\mathcal{H} \mid \tilde{\mathbf{x}}^T) \overset{(d)}{\geq} \left(1 - \frac{1}{\log T}\right) r_T^*(\mathcal{H}),
\end{aligned}$$

where (a) follows by conditioning on the event $\{\tilde{\mathbf{x}}^T \subset \mathbf{x}^{T_1}\}$ and observing that the regret is positive for all $\mathbf{x}^{T_1}$; (b) follows by $\inf \mathbb{E} \geq \mathbb{E} \inf$; (c) follows from the fact that $r^*_{T_1}(\mathcal{H} \mid \mathbf{x}^{T_1}) \geq r^*_T(\mathcal{H} \mid \tilde{\mathbf{x}}^T)$ which further follows from the monotonicity of $r_T^*(\mathcal{H} \mid \mathbf{x}^T)$; (d) follows by Lemma 7.19. To complete, we notice that $T = \kappa^{-1}(T_1)$ and $\kappa^{-1}(T_1) \geq \frac{T_1}{\log T_1}$. □

The following lemma shows the monotonicity for Log-loss:

Lemma 7.21. For Log-loss, we have $r^*_{T_1}(\mathcal{H} \mid \mathbf{x}^{T_1}) \geq r^*_T(\mathcal{H} \mid \tilde{\mathbf{x}}^T)$, so long as $\tilde{\mathbf{x}}^T \subset \mathbf{x}^{T_1}$.

Proof. Note that for any $\mathbf{x}^T$, we have by Theorem 4.3 that:

$$r_T^*(\mathcal{H} \mid \mathbf{x}^T) = \log \sum_{y^T} \sup_{h \in \mathcal{H}} \prod_{t=1}^{T} h(\mathbf{x}_t)^{y_1}(1 - h(\mathbf{x}_t))^{1-y_t}.$$

Therefore, any permutation over $\mathbf{x}^T$ does not change the value r_T^*. Now, suppose $\tilde{\mathbf{x}}^T \subset \mathbf{x}^{T_1}$; we can permute $\mathbf{x}^{T_1}$ so that the first T samples match with $\tilde{\mathbf{x}}^T$. The result follows from the fact that playing more rounds does not decrease the regret. To see this, we let $h \in \mathcal{H}$ to be the hypothesis that achieves minimal accumulated loss in the first T rounds, we then select the label y_t for which $\ell(\hat{y}_t, y_t) \geq \ell(h(\mathbf{x}_t), y_t)$ for the following steps $t > T$, which ensures non-decreasing regret. □

Finally, we apply the above general lower bound to the expected worst case minimax regret.

Corollary 7.22. Assume ℓ is the Log-loss. If $r_T^*(\mathcal{H}) \geq C \log^\alpha T$ then

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \geq C \log^\alpha T - o(\log^\alpha T),$$

where P is the class of *i.i.d.* distributions. If $r_T^*(\mathcal{H}) \geq CT^\alpha$, then

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \geq \frac{CT^\alpha}{\log^\alpha T} - o(T^\alpha / \log^\alpha T).$$

Remark 7.2. A question arises whether the $\log T$ factor in Corollary 7.22 can be eliminated. We do not have a complete answer for this question at this point; however, it is easy to show that there exists a class $\mathcal{H}$ such that $\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq (1 - e^{-1}) r_T^*(\mathcal{H})$, where P is the class of all *i.i.d.* processes. Meaning that the reduction as in Corollary 7.22 will necessarily introduce a factor < 1 for polynomial regrets $r_T^*(\mathcal{H})$.

We refer to Chapter 5 for the lower bounds on $r_T^*(\mathcal{H})$ of various classes $\mathcal{H}$ under Log-loss. In particular, the following lower bound is a complement to Corollary 7.18.

Proposition 7.3. For any $l \geq 1$, there exists a $[0, 1]$-valued class $\mathcal{H}$ with α-fat-shattering number of order $O(\alpha^{-l})$ and P is the class of all *i.i.d.* distributions over $\mathcal{X}^T$, such that

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \geq \tilde{\Omega}(T^{l/(l+1)}),$$

under logarithmic loss.

Proof. Let $\mathcal{X} = [T]$; we define $\mathcal{H} = \{h \in [0,1]^{\mathcal{X}} : \sum_{t=1}^{T} h(t)^l \leq 1\}$. We claim that the α-fat shattering number of $\mathcal{H}$ is upper bounded by α^{-l}. To see this, we assume there exist d points $\mathbf{x}^d \in [T]$ such that $d > \alpha^{-l}$ and $\mathbf{x}^d$ is α-fat shattered by $\mathcal{H}$. By definition of α-fat shattering, there exist two functions $h_1, h_2 \in \mathcal{H}$ such that $\forall i \in [d]$, $|h_1(\mathbf{x}_i) - h_2(\mathbf{x}_i)| \geq 2\alpha$. This implies that $\sum_{t=1}^{T} |h_1(t) - h_2(t)|^l \geq d \cdot (2\alpha)^l > 2^l$, i.e., $||h_1 - h_2||_l > 2$. However, this contradicts the fact that $||h_1 - h_2||_l \leq ||h_1||_l + ||h_2||_l \leq 2$ by the triangle inequality of L_l norm. By Theorem 5.16, we have $r_T^*(\mathcal{H}) \geq \Omega(T^{l/(l+1)})$. Invoking Corollary 7.22, the result follows. □

Note that Proposition 7.3 only shows that the lower bound $\tilde{\Omega}(T^{l/(l+1)})$ holds for certain *hard* classes. We prove in the following proposition a lower bound that holds for *all* classes.

Proposition 7.4. Let $l \geq 1$, $\mathcal{H}$ be any $[0,1]$-valued class with α-fat-shattering number of order $\Omega(\alpha^{-l})$ and P is the class of all *i.i.d.* distributions over $\mathcal{X}^T$. Then

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \geq \tilde{\Omega}(T^{(l-1)/l}),$$

under logarithmic loss.

Proof. Let $\mathbf{x}^T$ be samples that are α-fat-shattered by $\mathcal{H}$ and witnessed by s^T, where $\alpha \geq \Omega(T^{-1/l})$. We now describe an adversary strategy that achieves the $\Omega(T^{(l-1)/l})$ lower bound for the fixed design regret $r_T^*(\mathcal{H} \mid \mathbf{x}^T)$. To see this, for any $t \in [T]$, if the predictor predicts $\hat{y}_t \geq s_t$, we set $y_t = 0$, else, we set $y_t = 1$. By definition of α-fat shattering, there exists $h \in \mathcal{H}$ such that $\forall t \in [T], |h(\mathbf{x}_t) - \hat{y}_t| \geq \alpha$ and $\ell(\hat{y}_t, y_t) \geq \ell(h(\mathbf{x}_t), y_t)$. We assume without loss of generality, $y_t = 1$. By definition of Log-loss, we have:

$$\ell(\hat{y}_t, y_t) - \ell(h(\mathbf{x}_t), y_t) = \log(h(\mathbf{x}_t)/\hat{y}_t) \geq \log((\hat{y}_t + \alpha)/\hat{y}_t) \geq \alpha/2,$$

The last inequality follows by $\log(1 + x) \geq x/(x + 1)$. Therefore, we have $r_T^*(\mathcal{H} \mid \mathbf{x}^T) \geq T\alpha/2 \geq \Omega(T^{(l-1)/l})$. The proposition now follows by Corollary 7.22. □

Note that when $l \geq 2$ the lower bound in Proposition 7.4 is achieved by Logistic regression [65, Example 2]. Therefore, the lower bound is not *universally* improvable (this is similar to Corollary 7.18).

Example 7.8 (Well-specified v.s. worst case y^T)**.** In this example we demonstrate that the expected worst case regret $\tilde{r}_T$ can be substantially different than the well-specified average case regret $\bar{r}_T$ as in [50]. This will explain why our Theorem 7.20 is a necessary technique for establishing lower bonds for $\tilde{r}_T$. To see this, for any $\mathcal{X}$ with $|\mathcal{X}| \geq T$ we define:

$$\mathcal{H} = \left\{ h_b(\mathbf{x}) = \frac{1}{2} + \frac{b(\mathbf{x})}{\sqrt{T}} : b \in [-1,1]^{\mathcal{X}} \right\}.$$

This class admits an $O(1)$ uniform KL-cover at scale $O(1/T)$, and therefore by [50], the well-specified regret is of order $O(1)$. However, by [51, Page 6], the fixed design regret $r_T^*(\mathcal{H}) \geq 2(1/\sqrt{T})T \geq \Omega(\sqrt{T})$. Invoking Corollary 7.22, this implies an $\tilde{\Omega}(\sqrt{T})$ lower bound for $\tilde{r}_T$. This also demonstrates that the KL-cover (or equivalently the Hellinger cover) as in [50] cannot capture the behaviour of $\tilde{r}_T$ under Log-loss even with values bounded away from 0.

7.6 Bibliographical Notes

The *hybrid* scenario was introduced by [60], where it is assumed that the features are generated by an unknown i.i.d. source, but the labels are still presented adversarially. In particular, [60] showed that for finite VC-dimensional classes and under absolute loss, regret grows as $O(\sqrt{\mathsf{VC}(\mathcal{H})T\log T})$. One of the core techniques in this work is an epoch-based approach that reduces the infinite class to a finite class case using successive coverings. However, their upper bound is dominated by a $\sqrt{T}$ term from the approximation error of the covering, which is too loose for many loss functions, e.g., logarithmic loss. Indeed, the same epoch-based approach (and its approximation error analysis) was used in [56] for logarithmic loss, resulting in an $O(\sqrt{T})$ regret bound. In [50], the authors showed that for logarithmic loss and finite VC-dimensional classes, the *risk* grows as $O((\mathsf{VC}(\mathcal{H})\log^2 T)/T)$. However, their proof applies only to the *average-case* minimax regret (see Section 7.1) and in the *realizable* (i.e., *well-specified*) case.

In [25], the authors considered a scenario where, at each time step, Nature adversarially selects a distribution to sample from a restricted class of distributions determined by previously generated samples (*not* previously selected distributions). This is characterized by the concept of *distribution-dependent* Rademacher complexity, using a similar minimax approach as discussed above. However, their result holds only for the *distribution non-blind* case (i.e., when the distribution is known in advance); see [25, Section 7].

This chapter builds upon the results in [66].

8

Online Leaning under General Unknown Processes

We have studied in Chapter 7 the online learning setting where features are generated by an unknown i.i.d. process, and we characterized tight performance guarantees using the notion of expected worst-case regret. In this chapter, we turn to the stochastic setting under more general feature generation processes, which may be highly flexible and, in particular, *non-stationary*.

8.1 Stochastic Modeling of Data Processes

Let $\mathcal{X}$ be a feature (instance) space, $\hat{\mathcal{Y}} = [0,1]$ be the prediction space, and $\mathcal{Y} = \{0,1\}$ be the true label space. As always, we write $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ for a class of functions $\mathcal{X} \to \hat{\mathcal{Y}}$. For any time horizon T, we consider a class P of random processes over $\mathcal{X}^T$. We are interested in the *expected worst case* minimax regret $\tilde{r}_T(\mathcal{H}, \mathsf{P})$ as defined in (7.1) under a general convex loss ℓ. This includes, for instance, the absolute loss $\ell(\hat{y}, y) = |\hat{y} - y|$ (which can be interpreted as $\mathbb{E}_{\hat{b} \sim \text{Bernoulli}(\hat{y})}[1\{\hat{b} \neq y\}]$) and the logarithmic loss $\ell(\hat{y}, y) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$. Using the minimax inequality (cf. Section 2.3), it is easy to observe that

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \geq \sup_{\boldsymbol{\xi}^{2T}} \inf_{\phi^T} \mathbb{E}_{(\mathbf{x}^T, y^T) \sim \boldsymbol{\xi}^{2T}} \left[\sum_{t=1}^{T} \ell(\phi_t(\mathbf{x}^t, y^{t-1}), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right],$$

where $\boldsymbol{\xi}^{2T}$ is a joint distribution over $\mathcal{X}^T \times \mathcal{Y}^T$ such that the marginal distribution of $\boldsymbol{\xi}^{2T}$ restricted on $\mathcal{X}^T$ is in P. We will use such a relation to derive *lower bounds* for $\tilde{r}_T$.

In this chapter we assume that $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ is binary valued [1] and has finite VC-dimension. We specifically study here how the structure of the *random process class* P impacts expected worst case regret. This is unlike most of the results in our prior chapters that focus on the impact of the structure of $\mathcal{H}$ on regret. We now provide several examples of P that demonstrate how previously considered setups in the literature fit into our framework.

Example 8.1. If P is the class of all singleton distributions over $\mathcal{X}^T$, our setup recovers the adversary setting, as discussed in Chapters 5 and 6, as well as in [27]. If P is the class of all *i.i.d.* processes over $\mathcal{X}^T$, our setup recovers those of [60] discussed in Chapter 7.

Example 8.2 (The smooth adversary setting)**.** The smooth adversary setting is an intermediate setting between the full adversary and the *i.i.d.* case. In this setting, one assumes that there is some (known) underlying reference measure μ over $\mathcal{X}$, such that at each time step t an adversary selects some σ-smooth distribution ν_t w.r.t. μ that generates sample $\mathbf{x}_t$. Formally, we say a distribution ν is σ-smooth (with $\sigma \leq 1$) w.r.t. to μ if ν is absolutely continuous w.r.t. μ and has density $v(\mathbf{x}) = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ such that $\mu\left(\{\mathbf{x} : v(\mathbf{x}) \leq 1/\sigma\}\right) = 1$. We denote by $\mathcal{S}^\sigma(\mu)$ the class of all σ-smooth distributions w.r.t. μ. We say a process $\boldsymbol{\nu}^T$ over $\mathcal{X}^T$ is σ-smooth w.r.t. μ if for all $t \leq T$ the conditional distribution $\nu_t(X_t \mid X^{t-1})$ of X_t conditioning on X^{t-1} is in $\mathcal{S}^\sigma(\mu)$ almost surely. Using a standard *skolemization* argument as discussed in Lemma 2.2, the minimax regret for any class $\mathcal{H}$ w.r.t. smooth adversaries, as in [63], [67], [68], is simply $\tilde{r}_T(\mathcal{H}, \mathsf{S}^\sigma(\mu))$.

[1] We assume $\mathcal{H}$ to be binary valued for the clarity of presentation. However, our results can be extended to real valued functions as well.

Note that the smooth adversary setting in Example 8.2 requires the reference distribution μ to be *known* and fixed in advance. However, in many practical scenarios, one may have *no* prior knowledge of the underlying reference measures. To address this limitation, we introduce the following general concepts.

Let $\mu_1, \ldots, \mu_K$ be probability measures over $\mathcal{X}$, and let $\mathcal{S}^\sigma(\mu_k)$ denote the set of all σ-smooth distributions over $\mathcal{X}$ with respect to reference measure μ_k. Let $\nu_t(X_t \mid X^{t-1})$ denote the conditional distribution of X_t given the past X^{t-1}. Then a random process X^T over $\mathcal{X}^T$, with joint distribution $\boldsymbol{\nu}^T$, is said to be a (K, σ)-*smooth process* if

$$\Pr\left[\exists\, \mu_1, \ldots, \mu_K \text{ such that } \forall t \in [T],\ \nu_t(X_t \mid X^{t-1}) \in \bigcup_{k \in [K]} \mathcal{S}^\sigma(\mu_k)\right] = 1. \tag{8.1}$$

We denote by U_K^σ the class of *all* (K, σ)-smooth processes.

Furthermore, we denote by $\mathsf{S}^\sigma(\mu_1, \ldots, \mu_K)$ the class of all σ-smooth stochastic processes with (known) reference measures $\mu_1, \ldots, \mu_K$; that is, for any $\boldsymbol{\nu}^T \in \mathsf{S}^\sigma(\mu_1, \ldots, \mu_K)$, we have for all $t \in [T]$,

$$\nu_t(X_t \mid X^{t-1}) \in \bigcup_{k \in [K]} \mathcal{S}^\sigma(\mu_k) \quad \text{almost surely.}$$

Note that the processes in Example 8.2 correspond to the special case $\mathsf{S}^\sigma(\mu)$, for a single known reference measure μ.

This leads to our next definition:

$$\widetilde{\mathsf{U}}_K^\sigma := \bigcup_{\mu_1, \ldots, \mu_K} \mathsf{S}^\sigma(\mu_1, \ldots, \mu_K),$$

where the union is taken over all K-tuples of distributions over $\mathcal{X}$. It is straightforward to show (see Propositions 8.1 and 8.2 below) that:

$$\mathsf{U}_K^\sigma \subset \mathsf{U}_1^{\sigma/K}, \qquad \widetilde{\mathsf{U}}_K^\sigma \subset \widetilde{\mathsf{U}}_1^{\sigma/K}, \qquad \mathsf{S}^\sigma(\mu_1, \ldots, \mu_K) \subsetneq \widetilde{\mathsf{U}}_K^\sigma \subsetneq \mathsf{U}_K^\sigma,$$

where the last two inclusions are *strict*.

Proposition 8.1. Let $\mu_1, \cdots, \mu_K$ be K *arbitrary* distributions over the same domain $\mathcal{X}$. Then for all $k \in [K]$ the measure μ_k is $1/K$-smooth w.r.t. $\bar{\mu}$, where $\bar{\mu} = \frac{1}{K} \sum_{k=1}^{K} \mu_k$.

Proof. Note that $\bar{\mu}$ is interpreted as follows: for any measurable event $A \subset \mathcal{X}$, we have $\bar{\mu}(A) = \frac{1}{K}\sum_{t=1}^{K} \mu_k(A)$. It is easy to verify that $\bar{\mu}$ is a probability measure over $\mathcal{X}$. We now show that, for all $k \in [K]$, μ_k is $1/K$-smooth w.r.t. $\bar{\mu}$. To see this, we observe that μ_k is absolutely continuous w.r.t. $\bar{\mu}$. By Radon–Nikodym theorem, there is a density $u_k(\mathbf{x}) = \frac{\mathrm{d}\mu_k}{\mathrm{d}\bar{\mu}}$ of μ_k w.r.t. $\bar{\mu}$. Let $\mathcal{E}_k = \{\mathbf{x} : u_k(\mathbf{x}) > K\}$. We have $\mu_k(\mathcal{E}_k)/K > \bar{\mu}(\mathcal{E}_k)$ provided $\bar{\mu}(\mathcal{E}_k) > 0$. However, by definition of $\bar{\mu}$, we also have $\mu_k(\mathcal{E}_k)/K \leq \bar{\mu}(\mathcal{E}_k)$. This implies that $\bar{\mu}(\mathcal{E}_k) = 0$. □

Proposition 8.2. Let μ_1, μ_2, μ_3 be distributions over $\mathcal{X}$ such that μ_1 is σ_1-smooth w.r.t. μ_2 and μ_2 is σ_2-smooth w.r.t. μ_3. Then μ_1 is $\sigma_1\sigma_2$-smooth w.r.t. μ_3.

We emphasize that the class U_K^σ is a very broad family of processes, encompassing many interesting and natural settings. A full characterization of this class is left for future work. Instead, we concentrate on the subclasses U_K^1 and $\widetilde{\mathsf{U}}_1^\sigma$, which are rich enough to convey our core insights. Intuitively, one can view U_K^1 as the collection of all *dynamically changing processes* with cost K: that is, along any sample path, the number of distinct conditional distributions of the process is bounded above by K.

Notations. Throughout this chapter, we use lower case Greek letters μ, ν to denote a probability measure over $\mathcal{X}$. For any two measures μ_1, μ_2, we use $\mu_1 \cdot \mu_2$ to denote the product distribution of μ_1, μ_2 and $\mu^{\otimes T}$ to denote the *i.i.d.* measure of μ over $\mathcal{X}^T$. We use boldface Greek letters $\boldsymbol{\nu}^T$ to denote *general* distributions over $\mathcal{X}^T$. We use Math Sans Serif font P to denote classes of distributions over $\mathcal{X}^T$. For any random process X^T over $\mathcal{X}^T$, $t \leq [T]$ and $\mathbf{x}^{t-1}$, we use $\nu_t(X_1 \mid \mathbf{x}^{t-1})$ to denote the conditional distribution of X_t conditioning on $\mathbf{x}^{t-1}$. We also use $\boldsymbol{\nu}^T$ to denote the joint distribution of X^T over $\mathcal{X}^T$. Sometimes, we write $\nu_t = \nu_t(X_t \mid \mathbf{x}^{t-1})$ to simplify the notation when the conditioning context $\mathbf{x}^{t-1}$ is clear. We should emphasize that all parameters appearing in our bounds are *non-asymptotic*, meaning that one should *not* view them as constants. We will often provide ranges of the parameters for our bounds to hold.

8.2 Known Distribution

We start by examining the simpler case where the reference measures are *known*, i.e., the class $\mathsf{S}^\sigma(\mu_1, \ldots, \mu_K)$. By Propositions 8.1 and 8.2, the analysis of the smooth adversary setting with multiple (known) reference measures can be reduced to the case with a single reference measure. It is therefore sufficient to consider the setup from Example 8.2, where a single reference distribution μ is given.

The following proposition, due to Haghtalab *et al.* [68] and substantially simplified in Block *et al.* [63], plays a key role in our analysis.

Proposition 8.3. For any σ-smooth random process X^T with reference measure μ, there exists a (coupled) random processes V^{mT} with *i.i.d.* distribution $\mu^{\otimes mT}$ such that w.p. $\geq 1 - Te^{-\sigma m}$ (over the joint distribution of X^T, V^{mT}), we have

$$\forall t \in [T],\ X_t \in \{V_{m(t-1)+1}, \cdots, V_{mt}\}.$$

Proof. We first sample V^{mT} according to the *i.i.d.* distribution $\mu^{\otimes mT}$. We then construct X_t recursively in the following manner. After generating $X_1, \cdots, X_{t-1}$, the conditional distribution of $\nu(X_t \mid X^{t-1})$ is determined. Let S_t be a random set such that each $Z_i \in \{V_{m(t-1)+1}, \cdots, V_{mt}\}$ is included into S_t independently w.p. $\sigma v_t(Z_i)$ (i.e., w.p. $1 - \sigma v_t(Z_i)$ we do not include it), where v_t is the density of $\nu(X_t \mid X^{t-1})$ w.r.t. μ (see Example 8.2). We then generate X_t by sampling uniformly from S_t if S_t is non-empty and sampling independently from ν_t if S_t is empty. It is easy to verify that X^T is distributed exactly according to $\boldsymbol{\nu}^T$, and w.p. $\geq 1 - (1-\sigma)^m$, we have $X_t \in \{V_{m(t-1)+1}, \cdots, V_{mt}\}$. The result then follows by union bound on $[T]$. □

A set $A \subset \mathcal{X}^\infty$ is monotone if for any $\mathbf{x}^T \subset \mathbf{z}^{T'}$, we have $\mathbf{x}^T \in A \Rightarrow \mathbf{z}^{T'} \in A$, where $\mathbf{x}^T \subset \mathbf{z}^{T'}$ means $\mathbf{x}^T$ is a *sub-sequence* of $\mathbf{z}^{T'}$ and $\mathbf{x}^T \in A$ means any infinite sequence with *prefix* $\mathbf{x}^T$ is in A. We have the following lemma.

Lemma 8.1. Let X^T and V^{mT} be the coupling as in Proposition 8.3 and $A \subset \mathcal{X}^\infty$ be an arbitrary monotone set, then

$$\Pr[X^T \in A] \leq Te^{-\sigma m} + \Pr[V^{mT} \in A].$$

Proof. By Proposition 8.3, we have w.p. $\geq 1 - Te^{-\sigma m}$ that $X^T \subset V^{mT}$. Denote B to be such an event. Since A is monotone, we have

$$\mathbb{E}[1\{\{X^T \in A\} \wedge B\} - 1\{\{V^{mT} \in A\} \wedge B\}] \leq 0.$$

This implies

$$\Pr[\{X^T \in A\} \wedge B] \leq \Pr[\{V^{mT} \in A\} \wedge B] \leq \Pr[V^{mT} \in A].$$

Our result follows by observing that:

$$\begin{aligned} &\Pr[X^T \in A] = \Pr[\{X^T \in A\} \wedge B] + \Pr[\{X^T \in A\} \wedge \bar{B}] \\ &\leq \Pr[\{X^T \in A\} \wedge B] + \Pr[\bar{B}] \leq \Pr[\{X^T \in A\} \wedge B] + Te^{-\sigma m} \end{aligned}$$

which completes the proof. □

Note that unions and intersections of *any* collection of monotone sets are monotone. For any two functions $h_1, h_2 : \mathcal{X} \to \{0,1\}$, the set $A_N = \{\mathbf{x}^\infty \in \mathcal{X}^\infty : \sum_{t=1}^{\infty} 1\{h_1(\mathbf{x}_t) \neq h_2(\mathbf{x}_t)\} \geq N\}$ is monotone for all $N \in \mathbb{N}$.

We now present one of our key technical lemmas, which establishes a high-probability covering property for finite VC classes.

Lemma 8.2. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any class with finite VC-dimension and μ be an arbitrary probability measure over $\mathcal{X}$. If $\mathcal{F}_\epsilon$ is an ϵ-cover of $\mathcal{H}$ w.r.t. μ, i.e.,

$$\sup_{h \in \mathcal{H}} \inf_{f \in \mathcal{F}_\epsilon} \Pr_{\mathbf{x} \sim \mu}[h(\mathbf{x}) \neq f(\mathbf{x})] \leq \epsilon, \tag{8.2}$$

with $\epsilon = \frac{1}{2M^2}$, then for all $n \in \mathbb{N}$ and $M \geq 2$ we have:

$$\Pr_{\mathbf{x}^M \sim \mu^{\otimes M}}\left[\sup_{h \in \mathcal{H}} \inf_{f \in \mathcal{F}_\epsilon} \sum_{t=1}^{M} 1\{h(\mathbf{x}_t) \neq f(\mathbf{x}_t)\} \geq 3\mathsf{VC}(\mathcal{H}) + n\right] \leq \frac{2}{M^n}.$$

Proof. For any $h \in \mathcal{H}$, we denote by $\hat{f}_h = \arg\min_{f \in \mathcal{F}_\epsilon} \Pr_{\mathbf{x} \sim \mu}[h(\mathbf{x}) \neq f(\mathbf{x})]$. Let S^0 and S^1 be *i.i.d.* samples of μ with size M and M^2, respectively. For any $N \leq M$, we define two events:

$$A_1^N = \left\{\exists h \in \mathcal{H} \ s.t. \sum_{s \in S^0} 1\{h(s) \neq \hat{f}_h(s)\} \geq N\right\},$$

and

$$A_2^N = \left\{ \exists h \in \mathcal{H} \ s.t. \ \sum_{s \in S^0} 1\{h(s) \neq \hat{f}_h(s)\} \geq N \ \text{ and } \ \sum_{s \in S^1} 1\{h(s) \neq \hat{f}_h(s)\} = 0 \right\}.$$

We now claim that $\Pr[A_2^N \mid A_1^N] \geq \frac{1}{2}$. To see this, conditioning on A_1^N, there exists some h for A_1^N to happen. For such function h, we can select $\epsilon = 1/(2M^2)$ in (8.2) such that (since $|S^1| = M^2$):

$$\mathbb{E}\left[\sum_{s \in S^1} 1\{h(s) \neq \hat{f}_h(s)\}\right] \leq \frac{1}{2}.$$

By the First Moment method we know that $1 - \Pr[X = 0] = \Pr[X \geq 1] \leq \mathbb{E}[X] \leq 1/2$ for any random variable X supported on $\mathbb{N}$ with $\mathbb{E}[X] \leq 1/2$. Thus $\Pr[A_2^N \mid A_1^N] \geq \frac{1}{2}$. This implies that $\Pr[A_1^N] \leq 2\Pr[A_1^N \cap A_2^N] \leq 2\Pr[A_2^N]$.

We now upper bound $\Pr[A_2^N]$. By symmetries of *i.i.d.* distribution, we have $\Pr[A_2^N(S^0 \cup S^1)] = \mathbb{E}_\pi \Pr[A_2^N(\pi(S^0 \cup S^1))] \leq \sup_{S^0 \cup S^1} \Pr_\pi[A_2^N(\pi(S^0 \cup S^1))]$, where π is uniform random permutation over $S^0 \cup S^1$. We now fix any $S^0 \cup S^1$ and perform a uniform random permutation π. Let $h \in \mathcal{H}$ be any function such that there exist at least N elements in $S^0 \cup S^1$ for which $\hat{f}_h(s) \neq h(s)$ (otherwise $\Pr_\pi[A_2^N] = 0$). Note that, in order for A_2^N to happen under π, none of the elements $s \in S^0$ for which $\hat{f}_h(s) \neq h(s)$ should be permuted to S^1. Denote such an event to be B. We have

$$\Pr_\pi[B] = \frac{\binom{M}{N}}{\binom{M^2+M}{N}} \leq \frac{1}{M^N},$$

where we have used the fact that $\frac{a}{b} \geq \frac{a-i}{b-i}$ for all $b \geq a \geq i > 0$. Since there are at most $(M^2 + M)^{\mathsf{VC}(\mathcal{H})}$ functions restricted on $S^0 \cup S^1$, we have by union bound that

$$\Pr_\pi[A_2^N] \leq \frac{(M^2 + M)^{\mathsf{VC}(\mathcal{H})}}{M^N} \leq M^{3\mathsf{VC}(\mathcal{H}) - N},$$

where we used the fact that $M \geq 2$. The result follows by taking $N := 3\mathsf{VC}(\mathcal{H}) + n$ in the above expression and noting that $\Pr[A_1^N] \leq 2\Pr[A_2^N]$. □

Lemma 8.2 implies the following important approximating bound for σ-smooth processes.

Proposition 8.4. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a class with finite VC-dimension, μ be an arbitrary distribution over $\mathcal{X}$ and X^T be any σ-smooth random process w.r.t. μ. If we take $\epsilon = \frac{\sigma^2}{2T^2 \log^2(T/\beta)}$ for some $\beta > 0$ and $\mathcal{F}_\epsilon$ to be the ϵ-covering set of $\mathcal{H}$ w.r.t. μ as in Lemma 8.2, then

$$\Pr\left[\sup_{h\in\mathcal{H}} \inf_{f\in\mathcal{F}_\epsilon} \sum_{t=1}^{T} 1\{h(X_t) \neq f(X_t)\} \geq 3\mathsf{VC}(\mathcal{H}) + n\right] \leq \beta + \frac{2}{T^n}.$$

Proof. Taking $m = \frac{\log(T/\beta)}{\sigma}$ as in Proposition 8.3 one can make the error probability upper bounded by β. Let $M = mT$ as in Lemma 8.2, we have by setting $\epsilon = \frac{1}{2M^2} = \frac{\sigma^2}{2T^2\log^2(T/\beta)}$ the probability as in Lemma 8.2 is upper bounded by $\frac{2}{T^n}$ since $M \geq T$. The theorem follows by Lemma 8.1 by noticing that the event of the proposition is monotone (see the discussion follows Lemma 8.1 by noticing that $\sup\inf \equiv \cup\cap$) and we apply Lemma 8.2 over the process V^{mT}. □

Corollary 8.3. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary valued class with finite VC-dimension, and μ be arbitrary distributions over $\mathcal{X}$. For any convex and bounded loss, we have

$$\tilde{r}_T(\mathcal{H}, \mathsf{S}^\sigma(\mu)) \leq O\left(\sqrt{T \cdot \mathsf{VC}(\mathcal{H}) \log(T/\sigma)} + \mathsf{VC}(\mathcal{H})\right).$$

For logarithmic loss we have

$$\tilde{r}_T(\mathcal{H}, \mathsf{S}^\sigma(\mu)) \leq O(\mathsf{VC}(\mathcal{H}) \log(T/\sigma)).$$

Proof. Let $\mathcal{F}_\epsilon$ be as in Proposition 8.4, and let $\beta = \frac{1}{2T}$. Taking $n = 2$, the tail probability in Proposition 8.4 is upper bounded by $\frac{1}{T}$.

Applying the EWA algorithm to $\mathcal{F}_\epsilon$, we obtain the following regret bound for bounded convex losses (cf. Theorem 7.1):

$$\sqrt{(T/2)\log|\mathcal{F}_\epsilon|} + 3\mathsf{VC}(\mathcal{H}) + O(1) = O\left(\sqrt{T\mathsf{VC}(\mathcal{H})\log(T/\sigma)} + \mathsf{VC}(\mathcal{H})\right),$$

where we have used the standard bound on the covering number: $\log|\mathcal{F}_\epsilon| \leq O(\mathsf{VC}(\mathcal{H})\log(1/\epsilon))$ [69].

Applying the Smooth Truncated Bayesian algorithm (Algorithm 5.2) to $\mathcal{F}_\epsilon$ with truncation parameter $\frac{1}{T}$, we obtain the following regret bound for the log-loss:

$$\log|\mathcal{F}_\epsilon| + 3\mathsf{VC}(\mathcal{H})\log T + O(1) = O(\mathsf{VC}(\mathcal{H})\log(T/\sigma)). \qquad \square$$

Remark 8.1. Note that the first bound in Corollary 8.3 recovers the bound in [68], while the second bound is new and improves a $\log T$ factor for Log-loss if we use the $\mathsf{VC}(\mathcal{H})\log T$ approximation bound of [68, Lemma B.2] instead of our Proposition 8.4.

Corollary 8.4. Let $\mathcal{H}$ be a class of finite VC-dimension and $\mathsf{S}^\sigma(\mu_1, \cdots, \mu_K)$ be the smooth process with multiple (known) reference measures $\mu_1, \cdots, \mu_K$. Then

$$\tilde{r}_T(\mathcal{H}, \mathsf{S}^\sigma(\mu_1, \cdots, \mu_K)) \le O(\sqrt{T\mathsf{VC}(\mathcal{H})\log(KT/\sigma)})$$

under bounded convex losses, and

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \le O(\mathsf{VC}(\mathcal{H})\log(KT/\sigma))$$

under logarithmic loss and bounded mixable losses.

Proof. This follows directly from Corollary 8.3 and Proposition 8.1 and 8.2. $\square$

8.3 Regret for Unknown Distributions: Universal Case

We now analyze the minimax regret for the *universal* smooth processes. In Section 8.3.1, we examine *dynamically changing processes* of cost K, i.e., the class U_K^1, and establish tight upper and lower bounds for hypothesis classes with finite VC dimension under the absolute loss. These bounds are further refined for specific loss functions, such as the logarithmic loss. In Section 8.3.2, we investigate the class $\widetilde{\mathsf{U}}_1^\sigma$ (i.e., smooth processes with a fixed but unknown reference measure) by establishing a key connection between $\widetilde{\mathsf{U}}_1^\sigma$ and the *adversarial K-selection* process. We demonstrate our approach by proving sublinear regret bounds for one-dimensional threshold functions.

8.3.1 The class U_K^1 with finite VC class

Before we analyze the class U_K^1, we note that the processes in U_K^1 are highly *non-stationary*. Our first main technical ingredient is the following decoupling of the random processes in U_K^1 into K (conditional) *i.i.d.* processes.

Let X^T be an arbitrary process in U_K^1. We can *extend* X^T into another process V^{KT} in the following manner. The first T samples of V^{KT} equal X^T. For any *conditional* marginal ν_k of X^T with $k \in [K]$, we extend the sample X^T by sampling *i.i.d.* from ν_k such that ν_k is used *exactly* T times in the sample V^{KT} for each $k \in [K]$. Now, we denote $V^{(k)} = V_{k_1}, \cdots, V_{k_T}$ as the subsequence in V^{KT} that corresponds to ν_k, where k_ts are *random* indices.

Proposition 8.5. Conditioning on k_1 and V^{k_1-1}, the sample $V^{(k)}$ is an *i.i.d.* process of length T for all $k \in [K]$ (the $V^{(k)}$s are not necessarily independent for different k).

Proof. Note that conditioning on k_1 and V^{k_1-1}, the distribution ν_k is determined. By definition of the conditional distribution for any events $A \subset \mathcal{X}^{T-1}$ and $B \subset \mathcal{X}$, we have

$$\begin{aligned}
&\Pr[V_{k_1}^{k_{T-1}} \in A,\ V_{k_T} \in B \mid V^{k_1-1}] \\
&= \Pr[V_{k_1}^{k_{T-1}} \in A \mid V^{k_1-1}] \cdot \Pr[V_{k_T} \in B \mid V_{k_1}^{k_{T-1}} \in A,\ V^{k_1-1}] \\
&= \Pr[V_{k_1}^{k_{T-1}} \in A \mid V^{k_1-1}] \cdot \nu_k(B),
\end{aligned}$$

where $V_{k_1}^{k_{T-1}} = \{V_{k_1}, V_{k_2}, \cdots, V_{k_{T-1}}\}$. The proposition follows by induction on T. □

It is important to point out that the extension of X^T to V^{KT} is required for the decoupling to work. Otherwise, the constructed process $V^{(k)}$ is *not* necessarily *i.i.d.* (instead it is a *random* prefix of an *i.i.d.* process). Now, to analyze the performance of a predictor Φ on the process X^T, it is sufficient to study Φ on each of the sub-sequences $V^{(k)}$. Note that, this is generally a non-trivial task, since the predictor can only access to each of $V^{(k)}$s *obliviously*, i.e., it never exactly knows the decoupling. The technical challenge is to ensure that the presence of other $V^{(k)}$s do not affect the performance of the predictor on each individual $V^{(k)}$.

The adaptive epoch-EWA algorithm. The *epoch* approach [60] is a common way for dealing with *distribution blind* (i.e., universal) cases. The algorithm proceeds as follows: we partition the time horizon into $\lceil \log T \rceil$ epochs, where each epoch s ranges from time steps $2^s - 1, \cdots, 2^{s+1}$. In epoch s, we perform *Exponential Weighted Average* (EWA) Algorithm 3.1 on a finite expert class by selecting one function from each equivalent class of $\mathcal{H}$ that agrees on the samples of the previous epochs. The rationale behind this approach is that as we obtain more and more samples, we can *learn* the underlying hypothesis and then use the learned hypothesis to make prediction for the next epoch. However, this heavily relies on the assumption that the distributions are *stationary* (i.e., the samples should have similar statistics among different epochs). This does not hold even for U_2^1.

Example 8.3 (Failure of epoch approach). Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2\}$ be the instance space and $\mathcal{H} = \{h_1, h_2\}$ be the hypothesis class with $h_1(\mathbf{x}_1) = h_2(\mathbf{x}_1) = 1$, $h_1(\mathbf{x}_2) = 0$ and $h_2(\mathbf{x}_2) = 1$. We define distributions ν_1, ν_2 to be the singleton distributions on $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. We assume that the time horizon is $T = 2^{s+1} - 1$. For the first $s - 1$ epochs, we use ν_1 to generate samples and use ν_2 for the last epoch. Now, after $s - 1$ epochs, the algorithm, as in [60], will choose the expert to be any one of h_1, h_2 (since they agree on the previous samples). It is easy to see that the algorithm must incur at least $T/2$ regrets (the adversary simply labels the following samples using h_i that differs from the algorithm's selection) .

It can be shown that *any predefined set of* epochs cannot provide bounds better than $\Omega(T^{2/3})$, even for the simple class of Example 8.3 (see Example 8.4 in Section 8.3.3). Our main idea for resolving this issue is the *adaptive* epoch approach, presented in Algorithm 8.1. Note that the "adaptive" in Algorithm 8.1 has two different meanings. First, we select the learning rate η_r adaptively, and second, the error bound E is computed adaptively (i.e., we change the epochs according to the

Algorithm 8.1 Adaptive epoch-EWA algorithm

```
Input: Reference class $\mathcal{H}$ and update threshold $N$
Initialize $s \leftarrow 0$, $E \leftarrow 0$, and $\mathcal{H}^0 \leftarrow \{h\}$, where $h \in \mathcal{H}$ is arbitrary.
for $t \leftarrow 1$ to $T$ do
    Set $t_s \leftarrow t$, $r \leftarrow 1$, $m \leftarrow |\mathcal{H}^s|$, and $W^r \leftarrow \{1, \dots, 1\} \in \mathbb{R}^m$.
    while $E \leq N$ do
        Set learning rate $\eta_r \leftarrow \sqrt{\frac{8 \log m}{r}}$.
        Receive $\mathbf{x}_t$.
        Make prediction:
            $\hat{y}_t = \frac{\sum_{i=1}^m h_i^s(\mathbf{x}_t) \cdot W_i^r}{\sum_{i=1}^m W_i^r}, \quad h_i^s \in \mathcal{H}^s.$
        Receive $y_t$.
        Update weights:
            $W_i^{r+1} \leftarrow W_i^r e^{-\eta_r \ell(h_i^s(\mathbf{x}_t), y_t)} \quad \forall i \leq m.$
        Update $E$:
            $E \leftarrow \max_{h \in \mathcal{H},\, h^s \in \mathcal{H}^s} \left\{ \sum_{e=0}^{r-1} \mathbf{1}\{h(\mathbf{x}_{t_s+e}) \neq h^s(\mathbf{x}_{t_s+e})\} : \forall j < t_s,\, h(\mathbf{x}_j) = h^s(\mathbf{x}_j) \right\}.$
        Increment $t \leftarrow t + 1$ and $r \leftarrow r + 1$.
    end while
    Increment $s \leftarrow s + 1$, reset $t \leftarrow t - 1$, and $E \leftarrow 0$.
    Define equivalence $h_1 \sim_s h_2$ if $\forall j \leq t$, $h_1(\mathbf{x}_j) = h_2(\mathbf{x}_j)$, where $h_1, h_2 \in \mathcal{H}$.
    Let $\mathcal{H}^s$ be the class that selects exactly one element from each equivalence class under $\sim_s$.
end for
```

samples we observe). Our main result for this section is the following performance bound of Algorithm 8.1.

Theorem 8.5. Assume that the loss ℓ is convex in the first argument and upper bounded by 1, and $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ is a class of finite VC-dimension. If $\hat{y}_t$ is the prediction rule of Algorithm 8.3 that takes input $\mathcal{H}$ and $N = \sqrt{(T \cdot \mathsf{VC}(\mathcal{H}) \log T)/K}$, we have for all $\epsilon > 0$ if $K^3 \cdot \mathsf{VC}(\mathcal{H}) \leq O(T^{1-\epsilon}/\log T)$

$$\tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \leq O(\sqrt{KT \cdot \mathsf{VC}(\mathcal{H}) \log T}),$$

where O hides a constant that depends only linearly on $1/\epsilon$. Furthermore, for any numbers $d, K \geq 1$ with $(8Kd) \cdot \log(2Kd) \leq T$, we have

$$\sup_{\mathcal{H}, \mathsf{VC}(\mathcal{H}) \leq d} \tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \geq \sqrt{KdT/64},$$

under the absolute loss. For any $K \leq T$ the bound $\Omega(\sqrt{KT})$ holds for threshold functions.

Sketch of Proof. We only sketch the main idea here and refer to Section 8.3.3 for detailed proof. At a high level, our goal is to bound the number of epochs (i.e., the number of times we reenter the while loop). Note that, we are exiting the while loop only when the approximation error E of current expert class $\mathcal{H}^s$ is larger than the threshold N. Suppose we can upper bound the number of epochs by S. We denote $T_1, \cdots, T_S$ to be the length of each epoch. Note that for each epoch $s \leq S$, the regret can be split into two parts: the regret against expert class $\mathcal{H}^s$ and the error of approximating $\mathcal{H}$ by $\mathcal{H}^s$. For the first term, we have by standard result (see Chapter 3 that the regret is upper bounded by $\sqrt{2T_s|\mathcal{H}^s|} \leq \sqrt{2T_s \cdot \mathsf{VC}(\mathcal{H}) \log T}$, the last inequality follows from $|\mathcal{H}^s| \leq T^{\mathsf{VC}(\mathcal{H})}$. The second term is trivially upper bounded by N, since we change epochs once the approximation error is larger than N. Therefore the regret is upper bounded by $\sum_{s=1}^{S}(\sqrt{2T_s \cdot \mathsf{VC}(\mathcal{H}) \log T} + N) \leq SN + \sqrt{2ST \cdot \mathsf{VC}(\mathcal{H}) \log T}$, where the inequality follows from Cauchy–Schwarz inequality $\sum_{s=1}^{S} \sqrt{T_s} \leq \sqrt{S \sum_{s=1}^{n} T_s} = \sqrt{ST}$. The key technical challenge is to show that if we choose $N = \sqrt{(T \cdot \mathsf{VC}(\mathcal{H}) \log T)/K}$, we can ensure that $S \leq O(K)$

w.h.p. under *any* process in U_K^1, provided $K^3 \cdot \mathsf{VC}(\mathcal{H}) \leq O(T^{1-\epsilon}/\log T)$. This is achieved using the decoupling of U_K^1, together with a symmetric argument for bounding the approximation errors on each of the decoupled sub-sequences, see Lemma 8.15 and 8.16 in Section 8.3.3.

To prove the lower bound, we use a *hard* hypothesis class similar to [68], together with a *mixed* adversary-*i.i.d.* process based on the concept of *Littlestone forests* that achieves the tightest dependency $\Omega(\sqrt{K \cdot \mathsf{VC}(\mathcal{H})T})$. We note that a reduction to the Littlestone dimension as in [68] can only provide an $\Omega(\sqrt{KT})$ bound. Our technical contribution is to obtain a tight dependency on both $\mathsf{VC}(\mathcal{H})$ and K. See Section 8.3.3 for detailed proof. □

Remark 8.2. Note that, for $K = 1$, Theorem 8.5 recovers the upper bound in [60] with lower computational cost (we only run $O(1)$ epochs for $K = 1$, while [60] runs $O(\log T)$ epochs). We believe the condition $K^3 \cdot \mathsf{VC}(\mathcal{H}) \ll T^{1-\epsilon}/\log T$ is an artifact of our analysis and could be eliminated via a further refined approach. We will establish a tighter dependency on K for the full range $K \leq T$ in the next section with a slightly worse $\log^3 T$ factor. Furthermore, Algorithm 8.3 can be made adaptive to K as well, see Remark 8.5 (in Section 8.3.3). Theorem 8.5 also establishes a fundamental distinction between the universal and distribution aware case, as in Corollary 8.4 w.r.t dependency of K, i.e., K vs $\log K$.

The adaptive epoch approach proposed in the previous section results in tight bounds for the absolute loss and general convex bounded losses. For some special losses such as the logarithmic loss and general mixable losses, we provide tighter bounds on regret. We note that our results in this section also provide tighter bounds for bounded convex losses with parameters beyond the ranges of Theorem 8.5. We start with the following generic upper bounding technique:

A generic upper bounding technique: A crucial part of establishing regret bounds when the reference distribution is known (e.g., [68]), is to apply the EWA algorithm over a *uniform* cover of $\mathcal{H}$ (see Corollary 8.3). This, unfortunately, is not available for our *universal* case, since we do not know the reference measure μ in advance. A general methodology for dealing with such cases is via the so called *stochastic sequential cover* as in Definition 7.1 of Chapter 7. This definition immediately implies the following regret bounds by the standard expert algorithms (e.g., EWA), as in Theorem 7.1 and 7.2.

Proposition 8.6. Let $\mathcal{G}$ be a stochastic sequential cover of $\mathcal{H}$ w.r.t P at scale $\alpha = 0$ and confidence $\beta = \frac{1}{T}$. Then $\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq O(\sqrt{T \log |\mathcal{G}|})$ under bounded convex losses and $\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq \log |\mathcal{G}|$ under logarithmic loss.

The above results lead us to the following general approach for upper bounding $\tilde{r}_T$ through stochastic sequential cover. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ and P be *arbitrary* classes as defined above. We first find a prediction rule $\Phi : (\mathcal{X} \times \{0,1\})^* \times \mathcal{X} \to \{0,1\}$ such that:

$$\forall \boldsymbol{\nu}^T \in \mathsf{P},\ \Pr_{\mathbf{x}^T \sim \boldsymbol{\nu}^T}\left[\sup_{h \in \mathcal{H}} \mathsf{err}(\Phi, h, \mathbf{x}^T) \geq B(T, \beta)\right] \leq \beta, \tag{8.3}$$

where $\mathsf{err}(\Phi, h, \mathbf{x}^T) = \sum_{t=1}^T 1\{\Phi(\mathbf{x}^t, h(\mathbf{x}_1), \cdots, h(\mathbf{x}_{t-1})) \neq h(\mathbf{x}_t)\}$ is the cumulative error of Φ under the *realizable* sample of h on $\mathbf{x}^T$ and $B(T, \beta)$ is an error bound depending on the confidence parameter β and the time horizon T. For any such prediction rule Φ, we can then bound the stochastic sequential cover size using the following lemma as shown in Lemma 7.6.

Lemma 8.6. Let $\mathcal{H}$ and P be arbitrary classes and Φ be a predictor satisfying (8.3). Then there exists a stochastic sequential cover $\mathcal{G}$ of $\mathcal{H}$ w.r.t. P at scale $\alpha = 0$ and confidence β such that

$$\log |\mathcal{G}| \leq O((B(T, \beta) + 1) \cdot \log T).$$

The upper bound on $\tilde{r}_T(\mathcal{H}, \mathsf{P})$ then follows from Proposition 8.6. We remark that a crucial part for applying this approach is finding the predictor Φ and the upper bound $B(T, \beta)$, which is generally non-trivial if the processes in P are *non-stationary* due to the requirement of finding a bound on the form $\Pr[\sup_h]$.

The product distributions: We first consider a simpler random process class and illustrate how our technique works. We say a distribution $\boldsymbol{\nu}^T$ over $\mathcal{X}^T$ is a product distribution of type K if there exist distributions $\nu_1, \cdots, \nu_K$ over $\mathcal{X}$ such that $\boldsymbol{\nu}^T = \prod_{t=1}^T \nu_t$, where $\nu_t \in \{\nu_1, \cdots, \nu_K\}$. Note that distributions ν_ks and the configuration of the marginals of $\boldsymbol{\nu}^T$ need not be fixed and are *unknown* to the learner (e.g., the processes in Example 8.3 are product distributions of type 2). We prove the following upper bound for the stochastic sequential covering for such distributions:

Theorem 8.7. Let $\mathcal{H}$ be a binary valued class with finite VC-dimension, and P be the class of all product distributions over $\mathcal{X}^T$ with type K. Then, there exists a global sequential covering set $\mathcal{G}$ of $\mathcal{H}$ at scale $\alpha = 0$ and confidence β such that $\log |\mathcal{G}| \le O(K \cdot \mathsf{VC}(\mathcal{H}) \log^2 T + \log T \log(1/\beta))$.

Proof. We start with the following technical lemma.

Lemma 8.8. Let $I_1, \cdots, I_T$ be random variables over $\{0,1\}^T$ such that there exists a number $C > 0$ and partition $J_1, \cdots J_K \subset [T]$ of $[T]$ such that for all $k \in [K]$ and $k_t \in J_k$

$$\mathbb{E}[I_{k_t} \mid I^{k_t - 1}] \le \frac{C}{t},$$

where k_t is the tth element in J_k. Then for all $\beta > 0$, we have

$$\Pr\left[\sum_{t=1}^T I_t \ge 3CK \log(T/K) + 7CK + \log(1/\beta)\right] \le \beta.$$

Proof. Let $I'_t = I_t - \mathbb{E}[I_t \mid I^{t-1}]$, we have I'_t form martingale differences. We now analyze the conditional variance of I'_t, i.e., $\sum_{t=1}^T \mathbb{E}[{I'}_t^2 \mid I^{t-1}]$. We compute the variance for each partition J_k. For any $k_t \in J_k$, we have $|I'_{k_t}| \le 1$ w.p. p_t and $|I'_t| \le p_t$ w.p. $1 - p_t$, where $p_t \le \min\{\frac{C}{t}, 1\}$. Therefore, we have $\sum_{t=1}^{|J_k|} \mathbb{E}[{I'}_{k_t}^2 \mid I^{k_t - 1}] \le \sum_{t=1}^{|J_k|} p_t + p_t^2 \le C \log |J_k| + 3C$. Here, we have used the fact that $\sum_{t=1}^{\infty} p_t^2 \le 2C$ and $\sum_{t=1}^{|J_k|} \frac{C}{t} \le C \log |J_k| + C$. The second inequality is straightforward; we prove the first inequality. We split the summation into $\sum_{t=1}^{C} p_t^2 + \sum_{t=C}^{\infty} p_t^2 \le C + \sum_{t=C}^{\infty} \frac{C^2}{t^2} \le 2C$, where the first inequality follows by $p_t \le \min\{\frac{C}{t}, 1\}$.

Now, the lemma follows by a simple application of the Bernstein's inequality Lemma 2.10 and noting that $\sum_{k=1}^{K} \log |J_k| \leq K \log(T/K)$ since $\sum_{k=1}^{K} |J_k| = T$. □

Proof of Theorem 8.7. Our proof exploits the symmetries of the product distributions of type K. At a high level, we will show that there exists an algorithm, i.e., the *1-inclusion graph* algorithm [58], that achieves $O(K \log T + \log(1/\beta))$ cumulative error bound w.p. $\geq 1-\beta$ if the features $\mathbf{x}^T$ are sampling from a product distribution of type K and the labels y^T are realized by some $h \in \mathcal{H}$. Suppose this holds, then one will be able to derive the covering size bound through Lemma 8.6.

We now establish the realizable cumulative error bound. Let Φ be the *1-inclusion graph* algorithm, as in [58], and $\boldsymbol{\nu}^T$ be an arbitrary product distribution of type K. We partition the index set $[T]$ into K groups $J_1, \cdots, J_K$ such that for any indices i, j belonging to the same group J_k, we have $\nu_i = \nu_j$. Note that such a partition will only be used in our analysis and it is *unknown* to the algorithm Φ. Denote by π a random permutation such that the restriction of π to any J_k with $k \in [K]$ is uniform random permutation over J_k and is independent for different k. Let A be an arbitrary event over $\mathbf{x}^T$. We have by symmetries of the product distribution that:

$$\Pr_{\mathbf{x}^T \sim \boldsymbol{\nu}^T}[A(\mathbf{x}^T)] = \mathbb{E}_{\pi} \Pr_{\mathbf{x}^T \sim \boldsymbol{\nu}^T}[A(\mathbf{x}^{\pi(T)})] \leq \sup_{\mathbf{x}^T} \Pr_{\pi}[A(\mathbf{x}^{\pi(T)})].$$

It is therefore sufficient to fix the features $\mathbf{x}^T$ and prove the cumulative error bound under permutation π. For any $h \in \mathcal{H}$, we denote I_t^h to be the indicator of the event

$$\Phi(\mathbf{x}^{\pi(t)}, \{h(\mathbf{x}_{\pi(1)}), \cdots, h(\mathbf{x}_{\pi(t-1)})\}) \neq h(\mathbf{x}_{\pi(t)}),$$

i.e., the predictor Φ makes an error at time t for the realizable sample of h. We claim that

$$\mathbb{E}_{\pi}[I_t^h \mid \mathbf{x}_{\pi(t+1)}, \cdots, \mathbf{x}_{\pi(T)}] \leq \frac{\mathsf{VC}(\mathcal{H})}{t_{k_t}},$$

where t_{k_t} is the position of t in J_{k_t} and $k_t \in [K]$ is the index such that $t \in J_{k_t}$. To see this, we have by [58, Theorem 2.3(ii)] that for any realization $\mathbf{x}^t$, there are at most $\mathsf{VC}(\mathcal{H})$ positions $j \in [t]$ such that

$\Phi(\mathbf{x}_{-j}^t, h(\{\mathbf{x}_{-j}^t\})) \neq h(\mathbf{x}_j)$, where $\mathbf{x}_{-j}^t$ is the sample of $\mathbf{x}^t$ by removing $\mathbf{x}_j$ and $h(\{\mathbf{x}_{-j}^t\}) = \{h(\mathbf{x}_1), \cdots, h(\mathbf{x}_{j-1}), h(\mathbf{x}_{j+1}), \cdots, h(\mathbf{x}_t)\}$ be the labels of h on $\mathbf{x}_{-j}^t$. Thus, there are at most $\mathsf{VC}(\mathcal{H})$ such indices in J_{k_t}; by restricting π on $\mathbf{x}^t$, we have $I_t^h = 1$ only if such indices are switched to $\mathbf{x}_t$ under π, which happens w.p. $\leq \frac{\mathsf{VC}(\mathcal{H})}{t_{k_t}}$. Now, by the *permutation invariance* of 1-inclusion graph predictor, we have that I_t^h is completely determined by $\mathbf{x}_{\pi(t)}, \cdots, \mathbf{x}_{\pi(T)}$. Therefore, we have

$$\mathbb{E}_\pi[I_t^h \mid I_{t+1}^h, \cdots, I_T^h] = \mathbb{E}_\pi[I_t^h \mid \mathbf{x}_{\pi(t+1)}, \cdots, \mathbf{x}_{\pi(T)}] \leq \frac{\mathsf{VC}(\mathcal{H})}{t_{k_t}}.$$

This implies that $I_1^h, \cdots, I_T^h$ form the *reversed* sequence as in Lemma 8.8. Invoking Lemma 8.8 with $C = \mathsf{VC}(\mathcal{H})$, we have

$$\Pr_\pi\left[\sum_{t=1}^T I_t^h \geq O(K \cdot \mathsf{VC}(\mathcal{H}) \log(T/K) + \log(1/\beta))\right] \leq \beta.$$

Since there are only $T^{\mathsf{VC}(\mathcal{H})}$ functions restricted on any $\mathbf{x}^T$ by Sauer's lemma, we have by union bound

$$\Pr_\pi\left[\sup_{h\in\mathcal{H}} \sum_{t=1}^T I_t^h \geq O(K \cdot \mathsf{VC}(\mathcal{H}) \log(T/K) + \log(T^{\mathsf{VC}(\mathcal{H})}/\beta))\right] \leq \beta.$$

The upper bound on the stochastic sequential covering number now follows by Lemma 8.6. □

□

Thus upper bounds on the regret follow from Theorem 8.7 and Proposition 8.6.

Corollary 8.9. Let $\mathcal{H}$ be a binary valued class of finite VC-dimension and P be the class of all production distributions of type K. For any $K, T \geq 1$ we have $\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq O(\sqrt{KT \cdot \mathsf{VC}(\mathcal{H}) \log^2 T})$ under bounded convex losses and $\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq O(K \cdot \mathsf{VC}(\mathcal{H}) \log^2 T)$ under log-loss.

The class U_K^1: The 1-inclusion graph algorithm for *product* processes in the previous part relies heavily on symmetries in the product distribution. This, unfortunately, does not hold for general processes in U_K^1. Our main technique to deal with this issue is to replace the 1-inclusion graph predictor with the ERM rule, together with a *perturbation* argument for establishing a realizable cumulative error bound, as in (8.3). This allows us to establish the following stochastic sequential covering bound:

Theorem 8.10. Let $\mathcal{H}$ be a binary valued class of finite VC-dimension. Then there exists a stochastic sequential covering set $\mathcal{G}$ of $\mathcal{H}$ w.r.t. U_K^1 at scale $\alpha = 0$ and confidence $\beta > 0$ such that

$$\log|\mathcal{G}| \leq O(K(\mathsf{VC}(\mathcal{H})\log^3 T \\ + \log^2 T\log(K/\beta))\log(\mathsf{VC}(\mathcal{H})\log T\log(K/\beta))),$$

where O hides an absolute constant independent of $K, \mathsf{VC}(\mathcal{H}), T, \beta$.

Sketch of Proof. We sketch only the high-level idea here and refer to Section 8.3.4 for the full proof. We show that for any process in U_K^1 and the *ERM predictor* Φ, the realizable cumulative error (see Equation (8.3)) is upper bounded by $B(T,\beta) \leq O(K(\mathsf{VC}(\mathcal{H})\log^2 T + \log T\log(K/\beta))\cdot\Delta)$, where $\Delta = \log(\mathsf{VC}(\mathcal{H})\log T\log(K/\beta))$. To achieve this, we first decouple the process in U_K^1 into K *conditional i.i.d.* processes (Proposition 8.5). We then establish the realizable cumulative error bound on each of the decoupled sub-sequences (which are *conditional i.i.d.*). The key technical justification that allows us to do so is that an ERM rule with additional *realizable* samples is still an ERM rule. This allows us to bound the cumulative error for each decoupled sub-sequence *independently* even though we can only access them *obliviously*. We emphasize that to bound the realizable cumulative error for ERM rule even for *i.i.d.* process is still a non-trivial task, since we require a $\Pr[\sup_h]$ type bound for Lemma 8.6 to apply. To resolve this issue we introduce a novel *perturbation* argument, as presented in Lemma 8.18, which provides a generic way of converting a $\sup_h \Pr$ bound to a $\Pr[\sup_h]$ bound for any finite VC class with *i.i.d.* sampling. □

We now have the following regret bounds for VC-class. See Section 8.3.4 for detailed proof.

Corollary 8.11. For VC class $\mathcal{H}$ we have

$$\tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \leq O\left(\sqrt{\Delta\cdot KT\cdot\mathsf{VC}(\mathcal{H})\log^3 T}\right)$$

under bounded convex losses and $\tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \leq O(\Delta \cdot K \cdot \mathsf{VC}(\mathcal{H}) \log^3 T)$ under log-loss and bounded mixable losses, where $\Delta = \log(\mathsf{VC}(\mathcal{H}) \log(KT))$. Moreover, for $Kd \ll T/\log d$, we have $\sup_{\mathcal{H}, \mathsf{VC}(\mathcal{H}) \leq d} \tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \geq d \max\{K, \log(T/d)\}$ under log-loss.

8.3.2 The class $\tilde{\mathsf{U}}_1^\sigma$ with threshold functions

We now study the universal smooth process $\tilde{\mathsf{U}}_1^\sigma$ with fixed (but unknown) reference measure, where $\sigma \in (0,1]$ is any positive real [2]. We start with the following reduction. Let μ be an arbitrary distribution over $\mathcal{X}$. We say a random variable X is K-selection w.r.t. μ if there exists a *deterministic* function f such that $X = f(V^K) \in \{V_1, \cdots, V_K\}$, where $V^K \sim \mu^{\otimes K}$. We say a *random process* over $\tilde{X}^T$ is *adversary* K*-selection* w.r.t. μ if for all $t \leq T$ the *conditional* marginals $\nu_t(X_t \mid X^{t-1})$ are K-selection w.r.t. μ almost surely. In Section 8.3.5, we prove the following key lemma that relates the class $\tilde{\mathsf{U}}_1^\sigma$ to the adversary K-selection processes.

Lemma 8.12. Let $A \subset \mathcal{X}^T$ be any event. If for all adversary K-selection process $\tilde{X}^T$ we have $\Pr[\tilde{X}^T \in A] \geq 1 - \beta$, then for any σ-smooth process $X^T \in \tilde{\mathsf{U}}_1^\sigma$ we have $\Pr[X^T \in A] \geq 1 - 2\beta$, provided $K \geq \frac{\log(T/\beta)}{\sigma}$.

Lemma 8.12 shows that to bound the prediction performance for $\tilde{\mathsf{U}}_1^\sigma$ it is sufficient to bound the performance of the adversary K-selection processes. Perhaps surprisingly, this reduction essentially loses no information, since the adversary K-selection processes are also $\in \tilde{\mathsf{U}}_1^{1/K}$. This follows from the fact that for any event A we have $\Pr[f(V^K) \in A] \leq 1 - (1 - \Pr_{V \sim \mu}[V \in A])^K \leq K \Pr_{V \sim \mu}[V \in A]$, i.e., the conditional marginals ν_t must be $1/K$-smooth w.r.t. μ.

Our main result of this section is the following stochastic sequential covering bound for the threshold functions w.r.t. adversary K-selection processes. See Section 8.3.5 for a detailed proof.

Theorem 8.13. Let $\mathcal{H} = \{h_a(x) = 1\{x \geq a\} : x, a \in [0,1]\}$ be the class of 1-dimension threshold functions and P be the class of all adversary

[2] Note that, the classes U_K^1 and $\tilde{\mathsf{U}}_1^\sigma$ *do not* include each other, for all $\sigma \in (0,1)$.

K-selection processes. Then there exists a stochastic sequential covering set $\mathcal{G}$ w.r.t. P at scale $\alpha = 0$ and confidence $\beta > 0$ such that

$$\log |\mathcal{G}| \leq O(\sqrt{KT \log(2KT^2/\beta)}).$$

Sketch of Proof. We sketch the main idea here and refer to Section 8.3.5 for a detailed proof. We stress that even though the threshold functions may be simple from a classical learning theory perspective, the proof of Theorem 8.13 is not. This is due to the complex structure of adversary K-selection processes. Our proof follows a similar path as in Theorem 7.11 but with a substantially more sophisticated analysis. To do so, we maintain a *realization* tree, with each node of the tree labeled by a subset of $\mathcal{H}$. We expand the leaves of the tree every time we receive a sample $\tilde{X}_t$ by splitting the associated subset of $\mathcal{H}$ according to the labels on $\tilde{X}_t$. Our main technical contribution is to bound the *maximum* depth of the realization tree to be $O(\sqrt{KT \log(2KT^2/\beta})$ w.p. $\geq 1 - \beta$. This relies on a careful analysis of the splitting process. The bound for the stochastic sequential covering will then follow from a similar construction as in Theorem 7.11. □

We complete this section with the following bounds for the regret.

Corollary 8.14. Let $\mathcal{H} = \{h_a(x) = 1\{x \geq a\} : x, a \in [0, 1]\}$, then

$$\tilde{r}_T(\mathcal{H}, \tilde{\mathsf{U}}_1^\sigma) \leq O\left(\sqrt{(T/\sigma) \log^2(T/\sigma)}\right),$$

under bounded mixable losses and logarithmic loss. For bounded convex losses, we have

$$\tilde{r}_T(\mathcal{H}, \tilde{\mathsf{U}}_1^\sigma) \leq O\left(\sqrt{\frac{T^{3/2} \log(T/\sigma)}{\sigma^{1/2}}}\right).$$

Proof. This follows directly by Theorem 8.13, Lemma 8.12 and Proposition 8.6. □

Remark 8.3. Corollary 8.14 establishes sublinear regret as long as $\sigma^{-1} \ll T/\log^2 T$. However, it should be noted that the regrets presented here are not quantitatively optimal. For recent developments that characterize the exact minimax regret for $\tilde{\mathsf{U}}_1^\sigma$ and general finite VC classes $\mathcal{H}$, we refer the reader to [70].

8.3.3 Proof of Theorem 8.5

Before we present a formal proof of Theorem 8.5, we first develop some technical concepts that are necessary for our proof. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary valued class. For any $i < j \leq M$ and $\mathbf{x}^M \in \mathcal{X}^M$, we define the *agreed-mismatch* number of $\mathcal{H}$ on discrete interval $[i,j] := \{i, i{+}1, \cdots, j\}$ to be

$$\mathsf{AM}(\mathcal{H}, i, j, \mathbf{x}^M) = \sup_{h_1,h_2\in\mathcal{H}} \left\{\sum_{t=i}^{j} 1\{h_1(\mathbf{x}_t) \neq h_2(\mathbf{x}_t)\} : \forall t < i,\ h_1(\mathbf{x}_t) = h_2(\mathbf{x}_t)\right\}.$$

Note that the error bound E in Algorithm 8.3 at the end of each epoch is always a *lower bound* for the agreed-mismatch number at that epoch (with i, j being the start and end of the epoch, respectively). We have the following key lemmas for bounding the *agreed-mismatch* number:

Lemma 8.15. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a class of finite VC-dimension and μ be an arbitrary distribution over $\mathcal{X}$. Then for any $i < j \leq M \in \mathbb{N}^+$, we have for all $E \geq 0$

$$\Pr_{\mathbf{x}^M \sim \mu^{\otimes M}}\left[\mathsf{AM}(\mathcal{H}, i, j, \mathbf{x}^M) \geq E\right] \leq e^{2\mathsf{VC}(\mathcal{H})\log j - (i\cdot E)/j}.$$

Proof. We use a symmetric argument as in the proof of Lemma 8.2. The event $\mathsf{AM}(\mathcal{H}, i, j, \mathbf{x}^M) \geq E$ is equivalent to

$$A = \left\{\exists h_1, h_2 \in \mathcal{H}\ s.t.\ \forall t < i,\ h_1(\mathbf{x}_t) = h_1(\mathbf{x}_t) \text{ and } \sum_{t=i}^{j} 1\{h_1(\mathbf{x}_t) \neq h_2(\mathbf{x}_t)\} \geq E\right\}.$$

By symmetries of $i.i.d.$ samples, we can fix $\mathbf{x}^j$ and perform a uniform random permutation π over $[j]$. Now, for the event A to happen, there must be some $h_1, h_2 \in \mathcal{H}$ that differ on at least E positions in $\mathbf{x}^j$. Denote $B \geq E$ to be the number of mismatches of h_1, h_2 on $\mathbf{x}^j$. In order for the event A to happen, one must not switch any $t \in [i,j]$ for which $h_1(\mathbf{x}_t) \neq h_2(\mathbf{x}_t)$ to $[1, i-1]$ under permutation π. This happens with probability upper bounded by (using a simple combinatorial argument):

$$\frac{\binom{j-i}{B}}{\binom{j}{B}} \leq \left(1 - \frac{i}{j}\right)^B \leq e^{-(i\cdot B)/j} \leq e^{-(i\cdot E)/j},$$

where we have used the fact that $\frac{a}{b} \geq \frac{a-t}{b-t}$ for all $b \geq a \geq t$ and $e^{-(i\cdot B)/j}$ is decreasing on B.

The lemma follows by applying a union bound on all the pairs (h_1, h_2) restricted on $\mathbf{x}^j$ and an application of Sauers's lemma (Lemma 2.1), and $\Pr_{\mathbf{x}^j}[A(\mathbf{x}^j)] \leq \sup_{\mathbf{x}^j} \Pr_\pi[A(\mathbf{x}^{\pi(j)})]$ due to symmetries of *i.i.d.* samples. □

The following lemma is the key element of our proof.

Lemma 8.16. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a class of finite VC-dimension and μ be an arbitrary distribution over $\mathcal{X}$. For any $E \leq M \in \mathbb{N}^+$ and $\mathbf{x}^M \in \mathcal{X}^M$, we denote by A the event that there exists

$$n > \frac{\log M}{\log(E/(2\mathsf{VC}(\mathcal{H})\log M + \log(M^2/\beta)))}$$

and $1 = i_1 < i_2 < \cdots < i_{n+1} = M$ such that

$$\forall j \leq n,\ \mathsf{AM}(\mathcal{H}, i_j, i_{j+1}, \mathbf{x}^M) \geq E.$$

Then

$$\Pr_{\mathbf{x}^M \sim \mu^{\otimes M}}[A] \leq \beta.$$

Proof. Let $B_{i,j}$ be the event that $\{\mathsf{AM}(\mathcal{H}, i, j, \mathbf{x}^M) \geq E$ and $j \leq (E \cdot i)/(2\mathsf{VC}(\mathcal{H})\log M + \log(M^2/\beta))\}$. By Lemma 8.15, we have for all i, j and $\beta > 0$

$$\Pr[B_{i,j}] \leq \frac{\beta}{M^2}.$$

Using the union on all the pairs (i, j), we have

$$\Pr[\exists i, j,\ B_{i,j}] \leq \beta.$$

Let $B = \bigcap_{i,j} \neg B_{i,j}$. Then $\Pr[B] \geq 1 - \beta$. Note that the event $\neg B_{i,j}$ implies that if $\mathsf{AM}(\mathcal{H}, i, j, \mathbf{x}^M) \geq E$ then

$$j \geq \frac{E \cdot i}{2\mathsf{VC}(\mathcal{H})\log M + \log(M^2/\beta)}.$$

Conditioning on the event B happening, we have, if event A (defined in the statement of Lemma 8.16) happens then

$$\forall j \leq n,\ i_{j+1} \geq (E \cdot i_j)/(2\mathsf{VC}(\mathcal{H})\log M + \log(M^2/\beta)),$$

since event A implies $\mathsf{AM}(\mathcal{H}, i_{j+1}, i_j, \mathbf{x}^M) \geq E$ for all $j \leq n$. Note that $i_2 \geq E$, hence by induction

$$i_{n+1} \geq \left(\frac{E}{2\mathsf{VC}(\mathcal{H})\log M + \log(M^2/\beta)}\right)^n.$$

However, since we also have $i_{n+1} \leq M$, we must have

$$n \leq \frac{\log M}{\log(E/(2\mathsf{VC}(\mathcal{H})\log M + \log(M^2/\beta)))}.$$

This contradicts the definition of A (the event A requires number n to be *larger* than the above quantity) and implies that conditioning on event B, event A cannot happen. Therefore, we have $\Pr[A \mid B] = 0$, i.e., $\Pr[A \cap B] = 0$. This implies

$$\Pr[A] = \Pr[A \cap B] + \Pr[A \cap \neg B] \leq \Pr[A \cap \neg B] \leq \Pr[\neg B] \leq \beta$$

as needed. □

Remark 8.4. We remark that the results in both Lemma 8.15 and 8.16 hold for a general exchangeable process as well. Note that these two results cannot be applied directly on the processes in U_K^1 since they require the underlying process to be exchangeable. Our key approach, as in Proposition 8.5, is to decouple the process in U_K^1 into *conditional i.i.d.* processes.

We now prove the upper bound of Theorem 8.5.

Proof of Theorem 8.5 (Upper Bound). Let $\boldsymbol{\nu}^T \in \mathsf{U}_K^1$ be an arbitrary dynamic changing process with cost K. We denote by X^T the random process generated by $\boldsymbol{\nu}^T$. Note that the main difficulty here is to deal with the dependency among the samples in X^T. Our key idea is to *extend* the sample X^T into a coupled sample V^{KT} such that the first T samples in V^{KT} match X^T and each *conditional* distribution selected for generating X^T contributes *exactly* T samples in V^{KT}. We denote $V^{(k)} = V_{k_1}, \cdots, V_{k_T}$ to be the samples generated by the kth conditional distribution (that is used to generate X^T), where $k \leq K$. We also denote by $X^{(k)}$ the truncated sample of $V^{(k)}$ on V^T. By Proposition 8.5, $V^{(k)}$ is a conditional *i.i.d.* process, conditioning on V^{k_1-1}. Therefore, the unconditioned process $V^{(k)}$ is a *mixture* of *i.i.d.* processes, thus *exchangeable*. Note that the truncated process $X^{(k)}$ need not be exchangeable.

Taking $N = \sqrt{(T \cdot \mathsf{VC}(\mathcal{H}) \log T)/K}$ in Algorithm 8.1, we show that the claimed regret upper bound holds. Let $E = N/K$ and

$$n = \frac{\log T}{\log(E/(2\mathsf{VC}(\mathcal{H}) \log T + \log(T^2K/\beta)))} + 1.$$

We show that w.p. $\geq 1-\beta$, the parameter s in Algorithm 8.1 is upper bounded by nK. Suppose otherwise, we have the algorithm reenter the while loop at least nK times. Denote $i_1 < i_2 < \cdots < i_{Kn}$ to be the time steps of reentering the while loop. Note that by construction of Algorithm 8.3, we exit the while loop only if the *agreed-mismatch* number at current phase is larger than N. Therefore, we have, for each of the phases $i_{l+1} - i_l$, there must be some $k \leq K$ such that $X^{(k)}$ contributes at least N/K mismatches. This implies that there exists some $k \leq K$ and indexes $t_1, \cdots, t_n$ (which is a sub-sequence of $i_1, \cdots, i_{nK}$) such that $X^{(k)}$ contributes at least N/K mismatches in all the phases $t_{j+1} - t_j$ with $j \leq n$ (note that here the phase $t_{j+1} - t_j$ may combine multiple phases of form $i_{l+1} - i_l$). Therefore, the *agreed-mismatch* number restricted only on $X^{(k)}$ at each phase $t_{j+1} - t_j$ is larger than N/K. This is because the phase $t_{j+1} - t_j$ includes a sub-phase $i_{l+1} - i_l$ such that the *agreed-mismatch* number restricted on $X^{(k)}$ for the sub-phase is larger than N/K. Taking h_1, h_2 to be the functions that whiteness such a agree-mismatch number, we have h_1, h_2 also agrees on $\mathbf{x}^{t_j-1}$ and differs on at least N/K positions on $t_{j+1} - t_j$. Hence the agree-mismatch number restricted on $X^{(k)}$ on phase $t_{j+1} - t_j$ is also larger than N/K. Since $X^{(k)}$ is a prefix of $V^{(k)}$, this implies the event of Lemma 8.16 restricted on $V^{(k)}$ happens. By Lemma 8.16 and exchangability of $V^{(k)}$, we have the event A in Lemma 8.16 with the selected n happens w.p. $\leq \beta/K$ for each $V^{(k)}$. Using a union bound on all the $V^{(k)}$s we have the assumed event (i.e., $s > nK$) happens w.p. $\leq \beta$.

Taking $\beta = \frac{1}{T}$ and conditioning on the event $s \leq nK$, we now split the regret into two parts – one that is incurred by the mismatches and the other incurred by the adaptive EWA algorithm. Let $T_1, \cdots, T_s$ be the lengths of the the epochs. We have, by standard results of Chapter 3, that the regret contributed by EWA algorithm is upper bounded by

$$\begin{aligned}\sum_{a=1}^{s}\sqrt{4T_a\cdot \mathsf{VC}(\mathcal{H})\log T} &\leq \sqrt{4sT\cdot \mathsf{VC}(\mathcal{H})\log T}\\ &\leq O(\sqrt{KT\cdot \mathsf{VC}(\mathcal{H})\log T}).\end{aligned}$$

where the first inequality follows from Cauchy–Schwartz and $\sum_a T_a = T$, while the second inequality follows from $s \leq nK$ and $n = O(1/\epsilon)$ provided $K \ll (T^{(1-\epsilon)}/(\mathsf{VC}(\mathcal{H})\log T))^{1/3}$. For the number of mismatches, each epoch contributes at most N mismatches and there are at most s epochs, therefore the number of mismatches is upper bounded by

$$sN \leq O(\sqrt{KT\cdot \mathsf{VC}(\mathcal{H})\log T}).$$

Finally, the bad event $s > nK$ contributes at most $O(1)$ regret, since the loss is bounded by 1 and the event happens with probability $\leq \frac{1}{T}$. □

Remark 8.5. Note that the upper bound in Theorem 8.5 can be made adaptive to K (i.e., without knowing K) as well via a simple doubling trick. To see this, we set $K = 1$ initially and run Algorithm 8.1 as in the proof above. Once the algorithm has updated for more than nK epochs, we update K being $2K$ and rerun the algorithm with the new K. Taking $\beta = \frac{1}{T^2}$, we have by union bound (on the updates of K) w.p. $\geq 1 - \frac{1}{T}$ there can be at most $\lceil \log K\rceil$ updates if the process is in U_K^1. Therefore, the regret is upper bounded by

$$\sum_{k=1}^{\lceil \log K\rceil} O\left(\sqrt{2^kT\cdot \mathsf{VC}(\mathcal{H})\log T}\right) = O\left(\sqrt{KT\cdot \mathsf{VC}(\mathcal{H})\log T}\right),$$

as needed.

We now prove the lower bound of Theorem 8.5.

Proof of Theorem 8.5 (Lower Bound). Let $\mathcal{X} = [0,1]\times\{1,2,\cdots,d\}$. We construct the following class of *product threshold* functions

$$\mathcal{H} = \{h_{\mathbf{a}}(x,b) = 1\{x \geq a_b\} : \mathbf{a}\in[0,1]^d,\ (x,b)\in[0,1]\times[d]\}.$$

It is easy to see that $\mathsf{VC}(\mathcal{H}) = d$, since the set $(0.5, 1), \cdots, (0.5, d)$ is shattered by $\mathcal{H}$, and any $d+1$ points must have two points with the same index in $[d]$, which cannot be shattered by $\mathcal{H}$.

We now describe a strategy for selecting y^T and $\{\nu_1, \cdots, \nu_K\}$ that achieve the claimed lower bound for any prediction rule (possibly randomize) under absolute loss. Let τ be a Littlestone tree for threshold functions $\{h_a(x) = 1\{x \geq a\} : a, x \in [0,1]\}$ of depth K, which is a $[0,1]$-valued full binary tree such that each path can be realized by a threshold function (see e.g., [24]). This must exist since threshold functions have infinite Littlestone dimension. We take d *copies* $\{\tau_1, \cdots, \tau_d\}$ of τ (i.e., the Littlestone forest). We select y^T uniformly from $\{0,1\}^T$ and select the ν_ks in the following manner: let $I_1, \cdots, I_d$ be d *pointers* such that each I_b points to a node in τ_b for all $b \in [d]$; initially all the I_bs point to the roots of τ_bs, respectively. We partition the time horizon into K epochs, each of length T/K. At the beginning of the kth epoch, we define the distribution

$$\nu_k = \mathsf{Uniform}\{(\mathsf{V}(I_1), 1), (\mathsf{V}(I_2), 2) \cdots, (\mathsf{V}(I_d), d)\},$$

where $\mathsf{V}(I_b) \in [0,1]$ denotes the value of the node in τ_b pointed to by index I_b. After the epoch k, we update the indices I_bs in the following manner: for any $b \in [d]$, if the number of 0s is more than the number of 1s for the labels in y^T corresponding to sample $(\mathsf{V}(I_b), b)$ during epoch k, we move I_b to its left child, and move to its right child otherwise.

We now show that the strategy described above achieves a regret lower bound $\Omega(\sqrt{KdT})$ for any prediction rule provided $\frac{T}{8Kd} \geq \log(2Kd)$. To see this, we note that by the selection of y^T, any prediction rule must incur $T/2$ *actual* expected cumulative loss. For any $k \in [K]$ and $b \in [d]$, we denote $n_{k,b}$ to be the number of appearances of $(\mathsf{V}(I_b), b)$ during epoch k. We have by the *multiplicative* Chernoff bound [71, Theorem 4.5(2)] that

$$\Pr\left[n_{k,b} \geq \frac{T}{2Kd}\right] \geq 1 - e^{-T/(8Kd)}.$$

Assuming $\frac{T}{8Kd} \geq \log(2Kd)$, then by union bound on all pairs (k, b), w.p. $\geq \frac{1}{2}$, $n_{k,b} \geq \frac{T}{2Kd}$ for all $k \in [K]$ and $b \in [d]$. We now condition on that such an event happens, which is independent of y^T. By the Khinchine's inequality, discussed in Chapter 2, the expected number of 1s of the labels corresponding to $(\mathsf{V}(I_b), b)$ in epoch k is bounded away from $\frac{n_{k,b}}{2}$ by $\sqrt{n_{k,b}/8} \geq \sqrt{T/(16Kd)}$. By our selection of ν_ks, we know that there must be some $h \in \mathcal{H}$ such that the difference of the expected (over randomness of y^T) cumulative losses incurred by the predictor and by h is lower bounded by:

$$\sum_{k=1}^{K}\sum_{b=1}^{d}\sqrt{T/(16Kd)} \geq \sqrt{KdT/16}.$$

This implies that there must *exist* some y^T such that the regret against the predictor is lower bounded by $\sqrt{KdT/16}$. Since our conditioning event on $\mathbf{x}^T$ happens w.p. $\geq 1/2$, the expected worst case regret is lower bounded by $\sqrt{KdT/64}$.

Finally, to see the unconditional $\Omega(\sqrt{KT})$ lower bound, we can replicate the argument above with $b = 1$ and note that $n_{k,1} = T/K$ holds always without invoking the multiplicative Chernoff bound. □

We now provide a supplement to Example 8.3 that demonstrates the failure of the epoch approach with *any predefined* epochs.

Example 8.4. Let $\mathcal{H}$, ν_1, ν_2 be as in Example 8.3. Now, for any predefined epochs and number M, there are two cases: (i) there exists an epoch of length larger than M; (ii) all of the epochs have lengths less than M. For case (i), we can replicate the argument as in Example 8.3 to obtain an $\Omega(M)$ lower bound. For case (ii), we use ν_2 to generate samples for all the T steps. Since the EWA algorithm is deterministic for absolute loss (though it can be interpreted as a randomized algorithm for miss-classification loss), by standard lower bounds (e.g., [24, Lemma 14]) for any $n \in [T]$, there must be some y^n and $h_i \in \{h_1, h_2\}$ such that the regret of EWA on y^n against h_i is lower bounded by $\Omega(\sqrt{n})$. Denote $n_1, \cdots, n_L$ to be the length of all epochs such that $n_l \leq M$ for all $l \in [L]$. We claim that:

$$\sum_{l=1}^{L}\sqrt{n_l} \geq (T-M)/\sqrt{M}. \tag{8.4}$$

This follows from the inequality $\sqrt{a+1}+\sqrt{b-1} \leq \sqrt{a}+\sqrt{b}$ for $a \geq b$ (since the function $\sqrt{x}-\sqrt{x-1}$ is monotone decreasing). Therefore, one can "merge" the n_ls with as many components equal to M as possible, yet the LHS of (8.4) does not increase. Since there are at least $(T-M)/M$ such components after the "merge", (8.4) holds. By the above discussion, each epoch l corresponding to some y^{n_l} and h_{i_l} with regret of EWA against h_i is lower bounded by $\Omega(\sqrt{n_l})$. Therefore, there must be a subset $A \subset [L]$ corresponding to the same h_i such that $\sum_{l\in A}\sqrt{n_l} \geq (T-M)/(2\sqrt{M})$. We choose the label y^{n_l} at epoch l for $l \in A$ and the label $h_i(\mathbf{x}_2)$ for all other epochs. This yields the lower bound $\Omega((T-M)/\sqrt{M})$ hence also $\Omega(\max\{M,(T-M)/\sqrt{M}\}) \geq \Omega(T^{2/3})$, where the minimum is attained when $M = T^{2/3}$ leading to $\tilde{r}_T \geq \Omega(T^{2/3})$.

8.3.4 Proof of Theorem 8.10 and Corollary 8.11

For any hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$, the ERM rule is any function $\mathsf{ERM} : (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{H}$ such that for all $t \geq 1$ and $(\mathbf{x}^t, y^t) \in (\mathcal{X}\times\mathcal{Y})^t$, we have

$$\sum_{i=1}^{t} 1\{\mathsf{ERM}(\mathbf{x}^t, y^t)[\mathbf{x}_i] \neq y_i\} = \inf_{h\in\mathcal{H}} \sum_{i=1}^{t} 1\{h(\mathbf{x}_i) \neq y_i\}.$$

Let $\Phi : (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$ be a prediction rule, $h \in \mathcal{H}$ and $\mathbf{x}^T \in \mathcal{X}^T$, we denote the *cumulative* error of Φ under the realizable sample of h on $\mathbf{x}^T$ as (recall the definition in (8.3)):

$$\mathsf{err}(\Phi, h, \mathbf{x}^T) = \sum_{t=1}^{T} 1\{\Phi(\mathbf{x}^{t-1}, \{h(\mathbf{x}_1), \cdots, h(\mathbf{x}_{t-1})\})[\mathbf{x}_t] \neq h(\mathbf{x}_t)\}.$$

We begin with the following high probability *cumulative* error bound for the ERM rule under realizable *i.i.d.* sampling:

Lemma 8.17. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any class with finite VC-dimension and ERM be an arbitrary ERM rule of $\mathcal{H}$. Then for any distribution μ over $\mathcal{X}$ and $\beta > 0$ we have w.p. $\geq 1-\beta$ over $\mathbf{x}^T \sim \mu^{\otimes T}$

$$\sup_{h\in\mathcal{H}} \mathsf{err}(\mathsf{ERM}, h, \mathbf{x}^T) \leq O((\mathsf{VC}(\mathcal{H})\log^2 T + \log(1/\beta)\log T)\cdot\Delta)$$

where $\Delta = \log(\mathsf{VC}(\mathcal{H})\log T\log(1/\beta))$ and O hides absolute constant independent of $\mathsf{VC}(\mathcal{H}), T, \beta$.

Note that even though the samples $\mathbf{x}^T$ in Lemma 8.17 are *i.i.d.*, the *predictions* made by ERM rule are *not* independent, which is the main technical difficulty in proving Lemma 8.17. To resolve this issue, we first establish the following key lemma which provides a general approach for converting a $\sup_h \mathbb{E}$ type bound to a $\mathbb{E}\sup_h$ bound. Our main proof technique is a *perturbation* argument, which is the main technical contribution of this section. For any prediction rule $\Phi : (\mathcal{X} \times \{0,1\})^* \to \{0,1\}^{\mathcal{X}}$ and $I \subset [T]$, we define a perturbed function Φ^I such that for all $\mathbf{x}^t, y^t$ we have

$$\Phi^I(\mathbf{x}^t, y^t) = \Phi(\mathbf{x}^t, \tilde{y}^t),$$

where $\tilde{y}_t = y_t$ if $t \notin I$ and $\tilde{y}_t = 1 - y_t$ if $t \in I$.

Lemma 8.18. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a class of finite VC-dimension, μ be a distribution over $\mathcal{X}$, and $\mathcal{F}_\epsilon$ is an ϵ-cover of $\mathcal{H}$ w.r.t. μ (see Lemma 8.2), where $\epsilon = \frac{1}{2T^2}$. Then for any prediction rule $\Phi : (\mathcal{X} \times \{0,1\})^* \to \{0,1\}^{\mathcal{X}}$ we have for all $m, n \in \mathbb{N}^+$

$$\begin{aligned}\Pr_{\mathbf{x}^T \sim \mu^{\otimes T}} &\left[\sup_{h \in \mathcal{H}} \mathsf{err}(\Phi, h, \mathbf{x}^T) \geq m + 3\mathsf{VC}(\mathcal{H}) + n\right] \\ &\leq \Pr_{\mathbf{x}^T \sim \mu^{\otimes T}} \left[\sup_{f \in \mathcal{F}_\epsilon} \sup_{I \subset [T], |I| \leq 3\mathsf{VC}(\mathcal{H}) + n} \mathsf{err}(\Phi^I, f, \mathbf{x}^T) \geq m\right] + \frac{1}{T^n}.\end{aligned}$$

Proof. Let A be the event that

$$A = \left\{\mathbf{x}^T : \sup_{h \in \mathcal{H}} \inf_{f \in \mathcal{F}_\epsilon} \sum_{t=1}^{T} 1\{h(\mathbf{x}_t) \neq f(\mathbf{x}_t)\} \leq 3\mathsf{VC}(\mathcal{H}) + n\right\}.$$

We have by Lemma 8.2 that $\Pr[A] \geq 1 - \frac{1}{T^n}$ (taking $M = T$ in the lemma). Conditioning on the event A happening, we have for all $h \in \mathcal{H}$, there exists $f \in \mathcal{F}_\epsilon$ such that there are at most $3\mathsf{VC}(\mathcal{H}) + n$ positions $t \in [T]$ such that $h(\mathbf{x}_t) \neq f(\mathbf{x}_t)$. Denote $I \subset [T]$ to be the set of such positions. Then Φ and Φ^I have the same outputs on $\mathbf{x}^t$ with labeling of h for all $t \in [T]$; meaning that

$$\mathsf{err}(\Phi, h, \mathbf{x}^T) - \mathsf{err}(\Phi^I, f, \mathbf{x}^T) \leq 3\mathsf{VC}(\mathcal{H}) + n,$$

since only the positions for which $h(\mathbf{x}_t) \neq f(\mathbf{x}_t)$ contribute 1 to the difference of errors. This implies

$$\sup_{h \in \mathcal{H}} \inf_{f \in \mathcal{F}_\epsilon} \inf_{I} \mathsf{err}(\Phi, h, \mathbf{x}^T) - \mathsf{err}(\Phi^I, f, \mathbf{x}^T) \leq 3\mathsf{VC}(\mathcal{H}) + n.$$

The result follows by noting that

$$\sup_{h\in\mathcal{H}} \inf_{f\in\mathcal{F}_\epsilon} \inf_{I} [\mathsf{err}(\Phi, h, \mathbf{x}^T) - \mathsf{err}(\Phi^I, f, \mathbf{x}^T)] =$$

$$= \sup_{h\in\mathcal{H}} \mathsf{err}(\Phi, h, \mathbf{x}^T) - \sup_{f\in\mathcal{F}_\epsilon} \sup_{I} (\Phi^I, f, \mathbf{x}^T),$$

and removing the conditioning on A by a union bound. □

Lemma 8.18 is interesting since it reduces an event of form $\sup_h$ with infinite $\mathcal{H}$ to an event of form $\sup_{f,I}$ with *finite* $\mathcal{F}_\epsilon$ and $\{I \subset [T] : |I| \leq 3\mathsf{VC}(\mathcal{H}) + n\}$. The latter can be handled using union bounds if we are able to obtain a high probability error bound for Φ^I for any such f and I. The following lemma establish such a result for ERM rule with *i.i.d.* sampling.

Lemma 8.19. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a class of finite VC-dimension, μ be a distribution over $\mathcal{X}$. For any $h \in \mathcal{H}$ and $I \subset [T]$ with $|I| \leq e$ for some integer $e \geq 1$, we have for all $\beta > 0$

$$\Pr_{\mathbf{x}^T\sim\mu^{\otimes T}}\left[\mathsf{err}(\mathsf{ERM}^I, h, \mathbf{x}^T) \geq O(\log T(\mathsf{VC}(\mathcal{H})\log T + e + \log(1/\beta))\cdot\Delta)\right] \leq \beta,$$

where $\Delta = \log(e\mathsf{VC}(\mathcal{H})\log T\log(1/\beta))$, ERM is *any* ERM rule, and O hides absolute constant independent of $e, \mathsf{VC}(\mathcal{H}), T, \beta$.

Proof. Fix any $h \in \mathcal{H}$ and $I \subset [T]$ with $|I| \leq e$. We denote by ERM_t^I the function generated by ERM^I using samples $\mathbf{x}^t, y^t$. Let $\mathsf{err}_t = \Pr_{\mathbf{x}\sim\mu}[\mathsf{ERM}_t^I(\mathbf{x}) \neq h(\mathbf{x})]$. We now claim that for all $t \in [T]$ we have

$$\Pr_{\mathbf{x}^t\sim\mu^{\otimes t}}\left[\mathsf{err}_t \geq O\left(\frac{(\mathsf{VC}(\mathcal{H})\log t + e + \log(1/\beta))\log(e\mathsf{VC}(\mathcal{H})\log t\log(1/\beta))}{t}\right)\right] \leq \beta. \quad (8.5)$$

To see this, we use a symmetric argument. Let S_1, S_2 be two *i.i.d.* samples of μ both of length t. For any $h_1, h_2 \in \mathcal{H}$, we define distance $d(h_1, h_2) = \Pr_{\mathbf{x}\sim\mu}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]$. We define two events

$$A_1^h = \left\{\exists h' \in \mathcal{H},\ d(h, h') \geq \epsilon \text{ and } \sum_{s\in S_1} 1\{h'(s) \neq h(s)\} \leq e\right\},$$

and

$$A_2^h = \left\{ \exists h' \in \mathcal{H},\ d(h,h') \geq \epsilon \text{ and } \sum_{s \in S_1} 1\{h'(s) \neq h(s)\} \leq e \text{ but} \right.$$

$$\left. \sum_{s \in S_2} 1\{h'(s) \neq h(s)\} \leq\leq \epsilon t/2 \right\}.$$

Using the same argument as in Lemma 8.2, we have $\Pr[A_1^h] \leq 2\Pr[A_2^h]$. By symmetries of *i.i.d.* distributions we can fix $S_1 \cup S_2$ and perform a random permutation π that switches coordinate i of S_1 and S_2 w.p. $\frac{1}{2}$ and independent of different $i \in [t]$. In order for the event A_2^h to happen we cannot switch more than e elements for which $\mathsf{ERM}_t^I(s) \neq h(s)$ with $s \in S_2$ to S_1. This happens with probability upper bounded by

$$\frac{1}{2^{\epsilon t/2}} \sum_{i=0}^{e} \binom{\epsilon t/2}{i} \leq 2^{-\epsilon t/2 + (e+1)\log(\epsilon t/2)}.$$

Using a union bound on functions of $\mathcal{H}$ restricted on $S_1 \cup S_2$, we have

$$\Pr[A_2^h] \leq 2^{\mathsf{VC}(\mathcal{H})\log t - \epsilon t/2 + (e+1)\log(\epsilon t/2)}.$$

Taking

$$\epsilon = c \cdot \left(\frac{(\mathsf{VC}(\mathcal{H})\log t + e + \log(2/\beta))\log(e\mathsf{VC}(\mathcal{H})\log t \log(2/\beta))}{t} \right)$$

one can make $\Pr[A_2^h]$ upper bounded by $\beta/2$ for some absolute constant $c > 0$. The Claim (8.5) follows by noting that $\mathsf{err}_t \geq \epsilon$ implies event A_1^h happens by construction of ERM^I.

We now upper bound the *cumulative errors* of ERM^I. Let event

$$G_t = \left\{ \mathsf{err}_t \leq c \cdot \left(\frac{(\mathsf{VC}(\mathcal{H})\log t + e + \log(4T/\beta))\log(e\mathsf{VC}(\mathcal{H})\log t \log(4T/\beta))}{t} \right) \right\},$$

and indicator

$$I_t = \{\mathsf{ERM}_{t-1}^I(\mathbf{x}_t) \neq h(\mathbf{x}_t) \text{ and } G_{t-1}\}.$$

We have $\Pr[G_t] \geq 1 - \beta/(2T)$ for all $t \leq T$. Note that G_{t-1} is independent of $\mathbf{x}_t$, thus we have (since $I_t = 1$ happens only when G_{t-1} happens *and*

$\mathsf{ERM}^I_{t-1}(\mathbf{x}_t) \neq h(\mathbf{x}_t))$

$$\mathbb{E}[I_t \mid I_1, \cdots, I_{t-1}] \leq c \cdot \left(\frac{(\mathsf{VC}(\mathcal{H}) \log t + e + \log(4T/\beta)) \log(e\mathsf{VC}(\mathcal{H}) \log t \log(4T/\beta))}{t} \right). \tag{8.6}$$

By Lemma 8.8 with $K = 1$, C being the numerator of Equation (8.6) and upper bound $\log t$ by $\log T$, we have for sufficiently large T that

$$\Pr\left[\sum_{t=1}^{T} I_t \geq 4c \cdot \log T(\mathsf{VC}(\mathcal{H}) \log T + e + \log(4T/\beta)) \cdot \Delta + \log(2/\beta)\right] \leq \beta/2,$$

where $\Delta = \log(e\mathsf{VC}(\mathcal{H}) \log T \log(4T/\beta))$. Note that, the events $G = \cap_{t\in[T]} G_{t-1}$ and $\mathsf{ERM}^I_{t-1}[\mathbf{x}_t] \neq h(\mathbf{x}_t)$ together imply that $I_t = 1$. Therefore, using the fact that $\Pr[A] \leq \Pr[A \cap G] + \Pr[\neg G] \leq \Pr[A \cap G] + \beta/2$ for any event A, we conclude

$$\Pr_{\mathbf{x}^T \sim \mu^{\otimes T}} \left[\mathsf{err}(\mathsf{ERM}^I, h, \mathbf{x}^T) \geq O(\log T(\mathsf{VC}(\mathcal{H}) \log T + e + \log(1/\beta)) \cdot \Delta)\right] \leq \beta.$$

This completes the proof. □

Proof of Lemma 8.17. By Lemma 8.18, it is sufficient to upper bound

$$\Pr_{\mathbf{x}^T \sim \mu^{\otimes T}} \left[\sup_{f \in \mathcal{F}_\epsilon} \sup_{I \subset [T], |I| \leq 3\mathsf{VC}(\mathcal{H}) + n} \mathsf{err}(\Phi^I, f, \mathbf{x}^T) \geq m\right]. \tag{8.7}$$

We now take $n = \log(2/\beta)/\log T$ in Lemma 8.18, i.e., $\frac{1}{T^n} = \beta/2$. By Lemma 8.19 with $e = 3\mathsf{VC}(\mathcal{H}) + n$ together with a union bound on $\mathcal{F}_\epsilon$ and $\{I \subset [T] : |I| \leq 3\mathsf{VC}(\mathcal{H}) + n\}$ and letting

$$m = O(\log T(\mathsf{VC}(\mathcal{H}) \log T + e + \log(2B/\beta)) \cdot \Delta)$$

where $\Delta = \log(e\mathsf{VC}(\mathcal{H}) \log T \log(2B/\beta))$ and $B = |\mathcal{F}_\epsilon| \cdot |\{I \subset [T] : |I| \leq 3\mathsf{VC}(\mathcal{H}) + n\}|$, one can make the error probability (8.7) upper bounded by $\beta/2$. We now observe that $\log |\mathcal{F}_\epsilon| \leq O(\mathsf{VC}(\mathcal{H}) \log T)$ and $\log |\{I \subset [T] : |I| \leq 3\mathsf{VC}(\mathcal{H}) + n\}| \leq O(\mathsf{VC}(\mathcal{H}) \log T + \log(1/\beta))$. Putting everything together and simplifying the expression, we have w.p. $\geq 1 - \beta$ over $\mathbf{x}^T \sim \mu^{\otimes T}$

$$\sup_{h \in \mathcal{H}} \mathsf{err}(\mathsf{ERM}, h, \mathbf{x}^T) \leq O(\log T(\mathsf{VC}(\mathcal{H}) \log T + \log(1/\beta)) \log(\mathsf{VC}(\mathcal{H}) \log T \log(1/\beta))).$$

This completes the proof. □

The following lemma is the key element in our proof.

Lemma 8.20. For any random process $X^T \in \mathsf{U}_K^1$, we denote V^{KT} and $V^{(k)} = V_{k_1}, \cdots, V_{k_T}$ for $k \in [K]$ as in Proposition 8.5. We have for all $k \in [K]$ w.p. $\geq 1 - \beta$ over V^{KT}

$$\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} 1\{\mathsf{ERM}(V^{k_t-1}, \{h(V_1), \cdots, h(V_{k_t-1})\})[V_{k_t}] \neq h(V_{k_t})\} \leq$$
$$O((\mathsf{VC}(\mathcal{H}) \log^2 T + \log T \log(1/\beta)) \log(\mathsf{VC}(\mathcal{H}) \log T \log(1/\beta))),$$

where ERM is any ERM rule and O hides absolute constant independent of $\mathsf{VC}(\mathcal{H})$, T and $\log(1/\beta)$.

Proof. By Proposition 8.5, we have $V^{(k)}$ is an *i.i.d.* process conditioning on V^{k_1-1}. The key observation is that the ERM rule over V^{KT} restricted on $V^{(k)}$ is still an (randomized) ERM rule, since we have assumed that the samples are *realizable*. Conditioning on any V^{k_1-1}, the upper bound then follows by Lemma 8.17 since it only requires that the ERM rule at each time step k_t is independent of V_{k_t} and it does not depend on how the ERM functions are selected (even if the selections are randomized). To remove the conditioning on V^{k_1-1}, we use the following law of total probability: for any event $A \subset V^{KT}$ we have

$$\Pr[A] = \mathbb{E}_{V_{k_1-1}}\left[\Pr[A \mid V^{k_1-1}]\right] \leq \sup_{k_1, V_{k_1-1}} \Pr[A \mid V^{k_1-1}].$$

The lemma now follows by taking A to be the event in the statement of the lemma. □

We now ready to prove Theorem 8.10.

Proof of Theorem 8.10. We first observe that for any prediction rule the cumulative error on X^T is less than the cumulative error on V^{KT}. Using Lemma 8.20 and a union bound on all the K subsequences $V^{(k)}$, we have for any ERM rule ERM, w.p. $\geq 1-\beta$ over V^{KT}, the cumulative error

$$\sup_{h \in \mathcal{H}} \mathsf{err}(\mathsf{ERM}, h, V^{KT}) \leq O(K(\mathsf{VC}(\mathcal{H}) \log^2 T + \log T \log(K/\beta)) \cdot \Delta),$$

where $\Delta = \log(\mathsf{VC}(\mathcal{H}) \log T \log(K/\beta))$. Since we have the following $\mathsf{err}(\mathsf{ERM}, h, X^T) \leq \mathsf{err}(\mathsf{ERM}, h, V^{KT})$, the sequential covering size then follows by Lemma 8.6. □

Finally, we prove Corollary 8.11.

Proof of Corollary 8.11. The upper bounds follow directly by Proposition 8.6 and Theorem 8.10 by taking $\beta = \frac{1}{T}$. We only need to prove the lower bound for log-loss. For the $\Omega(Kd)$ lower bound, we consider the same hard class $\mathcal{H}$ as in the lower bound proof of Theorem 8.5 and the *Littlestone forests* $\tau_1, \cdots, \tau_d$ with pointers I_bs. We partition the time steps into K epochs. At each epoch k, we use the same ν_k as in the lower bound proof of Theorem 8.5 to generate samples. We move to the next epoch if all elements in the support of ν_k (which is a uniform distribution) have appeared at least once in the sample. We then change the pointers I_b of each tree τ_b in the following manner: if the prediction made by the predictor on the first appearance of $(\mathsf{V}(I_b), b)$ is $\geq \frac{1}{2}$, we update I_b to its left child, and update to right child if the prediction is $< \frac{1}{2}$. It is easy to verify that the expected regret is lower bounded by $\Omega(Kd)$, provided $Kd \ll T/\log d$ by the coupon collector problem. The lower bound for $\Omega(d \log(T/d))$ as discussed in Chapter 7.5. □

8.3.5 Proof of Lemma 8.12 and Theorem 8.13

Proof of Lemma 8.12. The proof is an operational interpretation of the coupling argument as in Proposition 8.3. Let μ be the reference measure that defines the σ-smooth process X^T (with $\boldsymbol{\nu}^T$ being the joint distribution of X^T). For any $m \in \mathbb{N}$, we denote V^{mT} to be an *i.i.d.* process with marginal μ and I^{mT} to be an *i.i.d.* process with marginal of uniform distribution over $[0, 1]$ that is independent of V^{mT}. We now construct a coupling between X^T and V^{mT}, I^{mT}. Suppose we have constructed X^{t-1}, we have that the conditional density $\nu_t = \nu_t(X_t \mid X^{t-1})$ is determined and we denote the density $v_t(\mathbf{x}) = \frac{\mathrm{d}\nu_t}{\mathrm{d}\mu}$. To construct X_t, we define the random set S_t as in Proposition 8.3 in the following manner: for any $V_{m(t-1)+i}$ with $i \in [m]$, if $\sigma v_t(V_{m(t-1)+i}) \geq I_{m(t-1)+i}$, we include $V_{m(t-1)+i}$ to S_t (and do not include otherwise). If S_t is not empty, we select the first element in S_t as X_t, else we sample a fresh independent sample $X'_t \sim \nu_t$ and let $X_t = X'_t$. It is easy to verify that the constructed process is distributed w.r.t $\boldsymbol{\nu}^T$. Note that the main difference with the proof of Proposition 8.3 is that we used the random variables I^{mT} on the selection of S_t instead of the Bernoulli$(\sigma v_t(V_{m(t-1)+i}))$ random variables (it is easy to check these two construction results in the same distribution of S_t).

We now denote $R_t = \{I_{m(t-1)+1}, \cdots, I_{mt}\}$, where R_t is independent of $\{V_{m(t-1)+1}, \cdots, V_{mt}\}$. The above coupling process can be expressed as $X_t = f(R_t, X'_t, \{V_{m(t-1)+1}, \cdots, V_{mt}\})$, where f is a *deterministic* function, such that w.p. $\geq 1 - Te^{-m\sigma}$ over R^T, X'^T, V^{mT}

$$\forall t \in [T],\ f(R_t, X'_t, \{V_{m(t-1)+1}, \cdots, V_{mt}\}) \in \{V_{m(t-1)+1}, \cdots, V_{mt}\}.$$

Let $\tilde{f}$ be the truncated function of f such that if $\forall i \in [m], \sigma v_t(V_{m(t-1)+i}) < I_{m(t-1)+i}$ we set

$$\tilde{f}(R_t, \{V_{m(t-1)+1}, \cdots, V_{mt}\}) = V_{m(t-1)+1}$$

and set $\tilde{f}(R_t, \{V_{m(t-1)+1}, \cdots, V_{mt}\}) = f(R_t, \{V_{m(t-1)+1}, \cdots, V_{mt}\})$ otherwise. We write

$$\tilde{X}_t = \tilde{f}(R_t, \{V_{m(t-1)+1}, \cdots, V_{mt}\}).$$

It is easy to see that w.p. $\geq 1 - Te^{-m\sigma}$ over the joint distribution $(X^T, \tilde{X}^T)$ that $\forall t \in [T]$, $X_t = \tilde{X}_t$. We now observe that conditioning on R^T, $\tilde{X}^T$ is an adversary m-selection process (since I^{mT} is independent of V^{mT} and $\tilde{X}^T$ is independent of X'^T). Therefore, we have by conditioning on R^T that

$$\Pr\left[\tilde{X}^T \in A\right] = \mathbb{E}\left[\Pr\left[\tilde{X}^T \in A \mid R^T\right]\right] \geq 1 - \beta.$$

Using a union bound we have

$$\Pr\left[X^T \in A\right] \geq 1 - \beta - Te^{-m\sigma}.$$

Taking $m = K$ and the assumption that $K \geq \frac{\log(T/\beta)}{\sigma}$, one finishes the proof. □

Proof of Theorem 8.13. Let $\tilde{X}^T$ be an adversary K-selection process with reference measure μ over $[0, 1]$. We assume that for any $x \in$

$[0,1]$, $\mu(\{x\}) = 0$. This assumption can be eliminated with a more tedious argument. However, we make the assumption here for clarity of presentation.

We consider the following *random* partitions of interval $[0,1]$. Initially, the partition $\mathcal{I}_0$ consists of only the interval $[0,1]$. At each time step t, we denote $\mathcal{I}_{t-1}$ to be the current partition. Let $J_t \in \mathcal{I}_{t-1}$ be the interval for which $\tilde{X}_t \in J_t$, we split J_t into two parts with values $< \tilde{X}_t$ and $\geq \tilde{X}_t$ respectively (if $\tilde{X}_t$ is the end point of J_t, we do not split and remain on the same J_t). We then replace J_t with the newly split intervals in $\mathcal{I}_{t-1}$ to form the partition $\mathcal{I}_t$. Note that one may view this partitioning process as expanding a binary tree with each node labeled by the intervals in $\mathcal{I}_t$ and expanding a leaf when the corresponding interval is split into two parts. Such a tree can be viewed as the (compressed) *realization* tree in Theorem 7.11 if we view the $\mathcal{I}_t$ as subsets of $\mathcal{H}$. Our goal is to bound the *maximum* depth of the tree.

For any time step t, we denote $J_t = [a_t, b_t]$ to be the interval for which $\tilde{X}_t \in J_t$ and

$$\lambda_t = \frac{\max\{\mu([a_t, X_t]), \mu([X_t, b_t])\}}{\mu([a_t, b_t])}$$

to be the *splitting ratio* of J_t. We claim that for any $\alpha > 0$,

$$\Pr\left[\lambda_t \geq 1 - \alpha \mid \tilde{X}^{t-1}\right] \leq 2\alpha K. \tag{8.8}$$

To see this, we denote $\mathcal{I}_{t-1} = \{J^1, \cdots, J^{n_t}\}$ to be the partition at time t before receiving $\tilde{X}_t$, where $J_t \in \mathcal{I}_{t-1}$ and $n_t \leq t$. For any interval $J^i = [a_i, b_i] \in \mathcal{I}_{t-1}$, we define the α-margin of J^i w.r.t. μ to be the intervals $[a_i, c_i]$ and $[d_i, b_i]$ such that:

$$\begin{aligned} c_i &= \sup\{x \in [a_i, b_i] : \mu([a_i, x]) \leq \alpha\mu([a_i, b_i])\} \\ d_i &= \inf\{x \in [a_i, b_i] : \mu([x, b_i]) \leq \alpha\mu(a_i, b_i)\}. \end{aligned}$$

Let $V_1^t, \cdots, V_K^t$ be the *i.i.d.* samples of μ that is used to generate $\tilde{X}_t$ and $B_t(\alpha)$ be the event that there exists some V_k^t and $J^i \in \mathcal{I}_{t-1}$ such that V_k^t is in the α-margin of interval J^i. Note that for any given V_k^t, the probability that V_k^t is in the α-margin of some interval in $\mathcal{I}_{t-1}$ is upper bounded by 2α. We have by independence of V_k^ts that

$$\Pr[B_t(\alpha)] \leq 1 - (1 - 2\alpha)^K \leq 2\alpha K.$$

By definition of adversary K-selection, we have the conditional event $\{\lambda_t \geq 1-\alpha \mid \tilde{X}^{t-1}\}$ implying that the event $B_t(\alpha)$ happens, i.e., the Equation (8.8) follows.

Let $I_t = 1\{\lambda_t \geq 1-\alpha\}$. Then $\mathbb{E}[I_t \mid I^{t-1}] \leq 2\alpha K$ and $I'_t = I_t - \mathbb{E}[I_t \mid I^{t-1}]$ form martingale differences. Using Azuma's inequality [9, Lemma A.7] for all $\alpha > 0$

$$\Pr\left[\sum_{t=1}^{T} I_t \geq 2\alpha KT + x\right] \leq \Pr\left[\sum_{t=1}^{T} I'_t \geq x\right] \leq e^{-2x^2/T}. \tag{8.9}$$

Taking $x \geq \sqrt{T\log(2T/\beta)}$, one can make the above probability less than $\beta/(2T)$. This implies that for any $n \leq T$ and $\alpha = \frac{n-\sqrt{T\log(2T/\beta)}}{4KT}$, w.p. $\geq 1-\beta/(2T)$, for any $\lambda_{t_1}, \cdots, \lambda_{t_n}$, we have

$$\sum_{i=1}^{n}(1-\lambda_{t_i}) \geq \left(n - \left(2\alpha KT + \sqrt{T\log(2T/\beta)}\right)\right)\alpha \geq \frac{\left(n-\sqrt{T\log(2T/\beta)}\right)^2}{8KT}, \tag{8.10}$$

where the first inequality follows by the fact that $I_t = 1$ implies $1-\lambda_t \leq \alpha$. Using a union bound on all $n \leq T$, we have w.p. $\geq 1-\beta/2$ that for any $n \leq T$ and $\lambda_{t_1}, \cdots, \lambda_{t_n}$, we have:

$$\sum_{i=1}^{n}(1-\lambda_{t_i}) \geq \frac{\left(n-\sqrt{T\log(2T/\beta)}\right)^2}{8KT}. \tag{8.11}$$

We now claim that w.p. $\geq 1-\beta/2$, for any interval J_t either $\mu(J_t) \geq \frac{\beta}{2KT^2}$ or J_t is in the final partition. To see this, we note that for any interval J_t at time step t, once $\mu(J_t) \leq \frac{\beta}{2KT^2}$, the probability it will be split at any following time step is upper bounded by (using the same argument for bounding the event $B_t(\alpha)$)

$$T\left(1-\left(1-\frac{\beta}{2KT^2}\right)^K\right) \leq \frac{\beta}{2T}.$$

Using a union bound on all the T intervals, w.p. $\geq 1-\beta/2$, all J_ts will either satisfy $\mu(J_t) \geq \frac{\beta}{2KT^2}$ or that J_t is in the final partition. By union bound, w.p. $\geq 1-\beta$, this happens simultaneously with the event of Equation (8.11). Conditioning on such a joint event, suppose now there exists a decreasing chain $J_{t_1} \supsetneq J_{t_2} \cdots \supsetneq J_{t_n}$, hence

$$\mu(J_{t_n}) \leq \prod_{i=1}^{n} \lambda_{t_i} \leq e^{-\sum_{i=1}^{n}(1-\lambda_{t_i})}.$$

This implies that if

$$n > \sqrt{8KT\log(2KT^2/\beta)} + \sqrt{T\log(2T/\beta)},$$

then $\mu(J_{t_n}) < \frac{\beta}{2KT^2}$ and therefore the chain must terminate.

Combining all of the above results, we conclude w.p. $\geq 1-\beta$ that there is no decreasing chain of length greater than

$$\sqrt{8KT\log(2KT^2/\beta)} + \sqrt{T\log(2T/\beta)} + 1$$

i.e., the *realization* tree has maximum depth upper bounded by the following $O(\sqrt{KT}\log(2KT^2/\beta))$. The bound on the stochastic sequential covering now follows by the same argument as in Theorem 7.11.

For the reader's convenience, we outline the argument in the following discussion. We construct a *sequential* function set $\mathcal{G}$ with *fixed* index set $\mathcal{W}$ of size $|\mathcal{W}| = 2^{\lceil\sqrt{15KT\log(2KT^2/\beta)}\rceil}$, i.e., for each $w \in \mathcal{W}$, we construct a *sequential* function $g_w : \mathcal{X}^* \to \{0,1\}$. To do so, we maintain for each node in the realization tree a *subset* of $\mathcal{W}$. We initially associate $\mathcal{W}$ to the root. At each time step after receiving $\tilde{X}_t$, for each node v in the realization tree, if v splits at the current step, we split the associated subset $\mathcal{W}_v \subset \mathcal{W}$ into two disjoint subsets of *equal* size and associate them to the newly split nodes, respectively. For any $w \in \mathcal{W}_v$, we assign the value $g_w(\tilde{X}^t) = 0$ if w is in the subset associated to the new left child and $g_w(\tilde{X}^t) = 1$ otherwise. If the node v does not split, we assign $g_w(\tilde{X}^t)$ to be the value on the agreed label (of the subset of $\mathcal{H}$ associated to v, see construction of realization tree at the beginning of the proof) on $\tilde{X}_t$. The process is said to have failed, if at some step a node v splits but the associated set $|\mathcal{W}_v| \leq 1$. Clearly, if the process does not fail until time T, the constructed set $\mathcal{G}$ sequentially covers $\mathcal{H}$ on $\tilde{X}^T$. Now, the key observation is that, from the discussion above, w.p. $\geq 1-\beta$ on $\tilde{X}^T$, any node is constructed after at most $\sqrt{8KT\log(2KT^2/\beta)} + \sqrt{T\log(2T/\beta)} + 1 \leq \lceil\sqrt{15KT\log(2KT^2/\beta)}\rceil$ splits. Since any split will decrease the associated subset of $\mathcal{W}$ by exactly $\frac{1}{2}$, we know that the process does not fail w.p. $\geq 1-\beta$ since

$|\mathcal{W}| = 2^{\lceil \sqrt{15KT\log(2KT^2/\beta)}\rceil}$. Therefore, the constructed set $\mathcal{G}$ stochastic sequential covers $\mathcal{H}$ at scale 0 and confidence β by Definition 7.1. □

8.4 Bibliographical Notes

The concept of the *smoothed adversary* was first introduced by [25] and later developed by [63], [67], [68]. Notably, the key technical ingredient—the *coupling argument*—was introduced by [68]. However, all of these works assumed that the reference measure of the smoothed adversary's samples must be known in advance. The treatment of unknown reference measures, and the results developed in this chapter, are based on [72]. We also refer to [70] for recent developments on minimax regret under the unknown distribution setting.

9

Constructive Algorithms for Minimax Regret

We have demonstrated in Chapters 7 and 8 general approaches for deriving tight minimax regret for the *hybrid* online setting via the concept of stochastic sequential covering. However, a major limitation of this approach is that the covering size can be exponentially large, rendering it computationally intractable.

This chapter introduces algorithmic primitives that address this gap through a reduction to the Empirical Risk Minimization (ERM) oracle, while still achieving sublinear regret.

9.1 Prelimilaries

We consider a slightly different formulation of the *expected worst-case regret* that is more suitable for our algorithmic treatment. Let $\mathcal{X}$ be the instance (feature) space, and let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be a function class mapping $\mathcal{X} \to [0,1]$. Recall the following *hybrid* online learning scenario: Nature selects an (unknown) distribution μ over $\mathcal{X}$ at the start of the game. At each time step t, Nature independently samples $\mathbf{x}_t \sim \mu$ and then adversarially selects $y_t \in [0,1]$, but only reveals $\mathbf{x}_t$ to the learner. A predictor then (randomly) generates $\hat{y}_t \in [0,1]$ based on the past inputs and outcomes, i.e., $\mathbf{x}^t = \{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$ and $y^{t-1} = \{y_1, \ldots, y_{t-1}\}$.

Nature then reveals y_t, and the predictor incurs a loss $\ell(\hat{y}_t, y_t)$ for a predefined loss function $\ell : [0,1]^2 \to \mathbb{R}^+$. We assume that the loss ℓ is *convex* in its first argument and L-Lipschitz in *both* arguments; for example, the *absolute loss* $\ell(\hat{y}, y) = |\hat{y} - y|$ satisfies these properties.

A prediction rule is a function Φ that takes inputs from $(\mathcal{X} \times [0,1])^* \times \mathcal{X}$ and outputs a *distribution* over $[0,1]$. For any prediction rule Φ and function class $\mathcal{H}$, we define the hybrid minimax regret [1] as:

$$\tilde{r}_T(\mathcal{H}, \Phi) = \sup_{\mu} \mathbb{E}_{\mathbf{x}_1} \sup_{y_1 \in [0,1]} \mathbb{E}_{\hat{y}_1} \cdots \mathbb{E}_{\mathbf{x}_T} \sup_{y_T \in [0,1]} \mathbb{E}_{\hat{y}_T} \left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right], \tag{9.1}$$

where $\mathbf{x}_t \sim \mu$ and $\hat{y}_t \sim \Phi(\mathbf{x}^t, y^{t-1})$ for all $t \in [T]$.

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathcal{X} \times [0,1]$, $C \in \mathbb{R}^+$, $\epsilon^n \in \{-1, +1\}^n$, and $\tilde{\mathbf{x}}^n \in \mathcal{X}^n$. The *mixed-ERM* oracle is the task of solving the following optimization problem:

$$\inf_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i) + C \sum_{j=1}^{n} \epsilon_j h(\tilde{\mathbf{x}}_j) \right\}.$$

This problem can typically be solved efficiently using gradient-based methods, even for complex function classes $\mathcal{H}$ (such as neural networks).

We say a predictor Φ is *oracle-efficient* if the running time of computing $\hat{y}_t \sim \Phi(\mathbf{x}^t, y^{t-1})$ is polynomial with respect to t by accessing a mixe-ERM oracle (with each oracle call treated as unit time) for any $\mathbf{x}^t, y^{t-1}$. Our goal is to design an *oracle-efficient* prediction rule Φ that minimizes the hybrid minimax regret $\tilde{r}_T(\mathcal{H}, \Phi)$ defined in (9.1).

9.2 Oracle-Efficient Hybrid Minimax Regrets

We first recall the following *standard* notion of Rademacher complexity:

Definition 9.1. Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be a function class and $T \in \mathbb{N}^+$. The Rademacher complexity of $\mathcal{H}$ at horizon T is defined to be

$$\mathsf{Rad}_T(\mathcal{H}) = \sup_{\mathbf{x}^T \in \mathcal{X}^T} \mathbb{E}_{\epsilon^T} \left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} \epsilon_t h(\mathbf{x}_t) \right],$$

[1] This concept is always upper bounded by the expected worst-case regret in (7.1). However, we conjecture that the two notions are, in fact, equivalent.

where ϵ_t is i.i.d. sampled from the uniform distribution over $\{\pm 1\}$.

The main result of this chapter is stated as follows:

Theorem 9.1. Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be a class with $\mathsf{Rad}_T(\mathcal{H}) \leq O(T^q)$ for some $q \in [\frac{1}{2}, 1]$, and let ℓ be a L-Lipschitz loss that is convex in its first argument. Then there exists an *oracle-efficient* prediction rule Φ with at most $O(L\sqrt{T}\log T)$ calls to the ERM oracle per round, such that

$$\tilde{r}_T(\mathcal{H}, \Phi) \leq O\left(L\sqrt{\log(LT)} \cdot T^{\frac{2-q}{3-2q}}\right).$$

In particular, for a binary-valued class with finite VC-dimension, we have

$$\tilde{r}_T(\mathcal{H}, \Phi) \leq O(L\sqrt{\mathsf{VC}(\mathcal{H})\log(LT)} \cdot T^{\frac{3}{4}}),$$

and for a real-valued class $\mathcal{H}$ with an α-fat shattering dimension of order α^{-p} for $p > 0$ [62], we have

$$\tilde{r}_T(\mathcal{H}, \Phi) \leq \tilde{O}(LT^{\max\{\frac{3}{4}, \frac{p+1}{p+2}\}}).$$

9.2.1 Efficient Predictor with Side-Information

To establish Theorem 9.1, we first consider a *hypothetical* scenario where we assume the predictor has access to some *side-information* $\mathbf{x}^0_{-N+1}$ sampled *i.i.d.* from the same distribution μ. It is crucial to note that this information is *known* to the adversarial as well, i.e., the adversary's strategy could also depend on $\mathbf{x}^0_{-N+1}$, which turns out to be the main obstacle in our analysis.

Formally, we consider the following learning game proceeds over a horizon of length M:

1. At the start of the game, Nature selects an unknown distribution μ over $\mathcal{X}$, samples an *i.i.d.* sample $\mathbf{x}^0_{-N+1}$ of size N from μ and reveals $\mathbf{x}^0_{-N+1}$ to a predictor;

2. At each time step $j \in [M]$, Nature samples $\mathbf{x}_j \sim \mu$ and selects *adversarially* $y_j \in [0,1]$ (depends on $\mathbf{x}^j_{-N+1}$ and $\hat{y}^{t-1}$) but reveals only $\mathbf{x}_j$;

3. The predictor then (randomly) generates $\hat{y}_j \in [0,1]$ based on $\mathbf{x}_{-N+1}^{j}$ and y^{j-1};

4. Nature reveals y_j and the predictor incurs loss $\ell(\hat{y}_j, y_j)$, for some predefined convex and L-Lipschitz loss.

Predictor via *surrogate* relaxation. Let $\hat{\mu}_N$ be the *empirical* distribution $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathbf{x}_{-N+i}}$ based on $\mathbf{x}_{-N=1}^{0}$, where $\delta_{\mathbf{x}}$ is the Dirac measure on $\mathbf{x}$. For any time step $j \in [M]$ and horizon M satisfying $M \leq N/2$, we construct the following *randomized* prediction rule:

1. Sample (internally) the *dummy samples* $\tilde{\mathbf{x}}_{j+1}, \cdots, \tilde{\mathbf{x}}_M$ from $\hat{\mu}_N$ *without replacement* [2] and $\epsilon_{j+1}, \cdots, \epsilon_M$ *i.i.d.* from the uniform distribution over $\{-1, +1\}$;

2. Make prediction

$$\hat{y}_j = \arg\min_{\hat{y}\in[0,1]} \sup_{y\in[0,1]} \left\{ \ell(\hat{y}, y) + \sup_{h\in\mathcal{H}} \left[2L \sum_{i=j+1}^{M} \epsilon_i h(\tilde{\mathbf{x}}_i) \right.\right.$$
$$\left.\left. -\ell(h(\mathbf{x}_j), y) - \sum_{i=1}^{j-1} \ell(h(\mathbf{x}_i), y_i) \right] \right\}. \tag{9.2}$$

Note that the dummy samples are generated from $\hat{\mu}_N$ instead of μ. Crucially, our sampling is performed *without replacement* (not *i.i.d.*), which is essential for our following analysis (Lemma 9.6). More generally, one may also replace the estimation $\hat{\mu}_N$ with other estimation rules instead of the empirical distribution we used here. This could provide tighter bounds if the distribution μ is well-structured.

The following lemma shows that the predictor $\hat{y}_j$ can be computed *efficiently* by accessing to a mixed-ERM oracle.

Lemma 9.2. The predictor $\hat{y}_j$ can be computed upto error $\pm\frac{1}{L\sqrt{M}}$ by making at most $O(L\sqrt{M}\log M)$ mixed-ERM oracle calls. Moreover, for

[2] For technical reasons, we assume here that $\tilde{\mathbf{x}}_{j+1}^{M}$ is sampled from $\hat{\mu}_N$ *without replacement*. Equivalently, $\tilde{\mathbf{x}}_{j+1}^{M}$ is sampled uniformly from all (permuted) *subseqeunces* of $\mathbf{x}_{-N+1}^{0}$ of length $M - j$.

binary valued class $\mathcal{H}$ with $y \in \{0,1\}$ and *absolute* loss, we need only 2 (regular) ERM orcale calls to compute $\hat{y}_j$ exactly.

Proof. Clearly, a naive approach for discretizing both $\hat{y}$ and y with scale $\frac{1}{L\sqrt{M}}$ yields an algorithm with L^2M oracle calls. The $O(L\sqrt{M}\log M)$ bound follows from [63, Thm 7] leveraging the convexity on $\hat{y}$. The second part follows from the relation $\epsilon h(\tilde{\mathbf{x}}) = |h(\tilde{\mathbf{x}}) - \frac{(1-\epsilon)}{2}| - \frac{1-\epsilon}{2}$ for $\epsilon \in \{+1,-1\}$ and the second assertion of [63, Thm 7]. □

Analysis of the regret. Denote by Φ the prediction rule derived from (9.2). We consider the following analogous hybrid minimax regret with the additional *side-information*:

$$\tilde{r}^{\mathsf{side}}_{M,N}(\mathcal{H},\Phi) = \sup_{\mu} \mathbb{E}_{\mathbf{x}^0_{-N+1}} \mathbb{E}_{\mathbf{x}_1} \sup_{y_1} \mathbb{E}_{\hat{y}_1} \cdots \mathbb{E}_{\mathbf{x}_M} \sup_{y_M} \mathbb{E}_{\hat{y}_M} \left[\sum_{j=1}^{M} \ell(\hat{y}_j, y_j) - \inf_{h\in\mathcal{H}} \sum_{j=1}^{M} \ell(h(\mathbf{x}_j), y_j)\right], \tag{9.3}$$

where the randomness of $\hat{y}_j$s is over the $\tilde{\mathbf{x}}$'s and ϵ's as in (9.2), while $\mathbf{x}_j$s are sampled *i.i.d.* from μ.

To proceed, we first introduce the following key concept. Let now $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_M, y_M) \in \mathcal{X} \times [0,1]$ be any realization of the feature-label pairs. We write $L^h_j = \sum_{i=1}^{j} \ell(h(\mathbf{x}_i), y_i)$ to simplify our discussion. The *surrogate* relaxation is defined as [3]

$$R_j = \mathbb{E}_{\tilde{\mathbf{x}},\epsilon}\left[\sup_{h\in\mathcal{H}} 2L \sum_{i=j+1}^{M} \epsilon_i h(\tilde{\mathbf{x}}_i) - L^h_j\right], \tag{9.4}$$

where $\tilde{\mathbf{x}}_i$s and ϵ_is are generated the same way as in (9.2). We also define the following variation that replaces the single $\tilde{\mathbf{x}}_{j+1}$ with a sample $\mathbf{x} \sim \mu$:

$$\tilde{R}_j = \mathbb{E}_{\mathbf{x}\sim\mu}\mathbb{E}_{\tilde{\mathbf{x}},\epsilon}\left[\sup_{h\in\mathcal{H}} 2L\epsilon_{j+1}h(\mathbf{x}) + 2L\sum_{i=j+2}^{M} \epsilon_i h(\tilde{\mathbf{x}}_i) - L^h_j\right]. \tag{9.5}$$

[3] Throughout the paper, we use the convention $\mathbb{E}_{\tilde{\mathbf{x}},\epsilon} \equiv \mathbb{E}_{\tilde{\mathbf{x}}^M_{j+1},\epsilon^M_{j+1}}$ to simplify notation.

Note that the main technique for proving the relaxation based regret bounds, such as [26], is through the concept of *admissibility*, which essentially asserts that

$$\mathbb{E}_{\mathbf{x}_j} \sup_{y_j} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}} \left[\ell(\hat{y}_j, y_j) + R_j\right] \leq R_{j-1}.$$

However, a major technical step for establishing such an result is based on the so-called *randomness matching* argument by leveraging the fact that the *dummy samples* used to define the relaxation are the same as the actual feature generating process. This, unfortunately, is not true in our case since the empirical distribution $\hat{\mu}_N$ can deviate arbitrarily from μ under total variation, regardless of how large the sample size N is. We instead establish the following *approx-admissibility* of our surrogate relaxation, with the proof deferred to Section 9.2.4.

Lemma 9.3 (Approx-Admissibility)**.** Let $\hat{y}_j$ be as in (9.2), then for all $j \in [M]$ we have:

$$\mathbb{E}_{\mathbf{x}_j} \sup_{y_j} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}} \left[\ell(\hat{y}_j, y_j) + R_j\right] \leq \tilde{R}_{j-1}. \tag{9.6}$$

We are now ready to state our first main technical lemma of this section, which follows from Lemma 9.3 by a "backward tracing" argument.

Lemma 9.4 (Regret Bound via Approx-Admissibility)**.** Let Φ be the predictor as in (9.2). Then for any class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ with a convex and L-Lipschitz loss ℓ, we have

$$\tilde{r}^{\mathsf{side}}_{M,N}(\mathcal{H}, \Phi) \leq \mathbb{E}_{\mathbf{x}^0_{-N+1}} \left[\tilde{R}_0 + \sum_{j=1}^{M-1} \mathbb{E}_{\mathbf{x}^j} \sup_{y^j} (\tilde{R}_j - R_j)\right], \tag{9.7}$$

where $\mathbf{x}^M_{-N+1}$ are sampled *i.i.d.* from μ and R_j, $\tilde{R}_j$ are defined as in (9.4) and (9.5).

Proof. Denote $\mathbb{Q}_j \equiv \mathbb{E}_{\mathbf{x}^0_{-N+1}} \mathbb{E}_{\mathbf{x}_1} \sup_{y_1} \mathbb{E}_{\hat{y}_1} \cdots \mathbb{E}_{\mathbf{x}_j} \sup_{y_j} \mathbb{E}_{\hat{y}_j}$ for notation convenience. We have

$$
\begin{aligned}
&\tilde{r}^{\mathsf{side}}_{M,N}(\mathcal{H},\Phi)\\
&\quad\overset{(a)}{=} \mathbb{Q}_M\left[\sum_{j=1}^{M}\ell(\hat{y}_j,y_j)+R_M\right]\\
&\quad\overset{(b)}{=} \mathbb{Q}_{M-1}\left[\sum_{j=1}^{M-1}\ell(\hat{y}_j,y_j)+\mathbb{E}_{\mathbf{x}_M}\sup_{y_M}\mathbb{E}_{\hat{y}_M}\left[\ell(\hat{y}_M,y_M)+R_M\right]\right]\\
&\quad\overset{(c)}{\le} \mathbb{Q}_{M-1}\left[\sum_{j=1}^{M-1}\ell(\hat{y}_j,y_j)+\tilde{R}_{M-1}\right]\\
&\quad= \mathbb{Q}_{M-1}\left[\sum_{j=1}^{M-1}\ell(\hat{y}_j,y_j)+R_{M-1}+(\tilde{R}_{M-1}-R_{M-1})\right]\\
&\quad\overset{(d)}{\le} \mathbb{Q}_{M-1}\left[\sum_{j=1}^{M-1}\ell(\hat{y}_j,y_j)+R_{M-1}\right]+\mathbb{E}_{\mathbf{x}^0_{-N+1}}\mathbb{E}_{\mathbf{x}^{M-1}}\sup_{y^{M-1}}(\tilde{R}_{M-1}-R_{M-1})\\
&\quad\overset{(e)}{\le} \mathbb{E}_{\mathbf{x}^0_{-N+1}}[\tilde{R}_0]+\sum_{j=1}^{M-1}\mathbb{E}_{\mathbf{x}^0_{-N+1}}\mathbb{E}_{\mathbf{x}^j}\sup_{y^j}(\tilde{R}_j-R_j),
\end{aligned}
$$

where (a) follows by definition of R_M; (b) follows by extracting the last layer of $\mathbb{Q}_M$; (c) follows by Lemma 9.3 and noticing that $\hat{y}_j$ has the same randomness as R_j; (d) follows by the the facts that $\sup(A+B)\le\sup A+\sup B$, $\sup\mathbb{E}\le\mathbb{E}\sup$, the linearity of expectation and $\tilde{R}_{M-1}-R_{M-1}$ is independent of $\hat{y}_j$ for all $j\le M-1$; (e) follows by repeating the same arguments for another $M-1$ steps. This completes the proof. □

Remark 9.1. Note that the decomposition presented in Lemma 9.4 holds whenever the approx-admissibility condition of Lemma 9.3 is satisfied. We believe this could be applicable to a broader set of problems and is of independent interest.

Bounding the relaxations. By Lemma 9.4, we know that the regret $\tilde{r}^{\mathsf{side}}_{M,N}(\mathcal{H},\Phi)$ can be upper bounded by $\tilde{R}_0$ and the discrepancies between R_j and $\tilde{R}_j$. Clearly, by the definition of $\tilde{R}_j$, we have $\tilde{R}_0\le 2L\mathsf{Rad}_M(\mathcal{H})$, where $\mathsf{Rad}_M(\mathcal{H})$ is the Rademacher complexity of $\mathcal{H}$ as in Definition 9.1.

To bound the discrepancies, for any $j \in [M-1]$, $\mathbf{x}^j, \tilde{\mathbf{x}}_{j+2}^M \in \mathcal{X}^*$, $\epsilon_{j+1}^M \in \{\pm 1\}^*$ and $y^j \in [0,1]^j$, we define the following function:

$$f_{\mathbf{x}^j,\tilde{\mathbf{x}}_{j+2}^M,\epsilon_{j+1}^M,y^j}(\mathbf{x}) = \sup_{h\in\mathcal{H}} \left\{ 2L\epsilon_{j+1}h(\mathbf{x}) + 2L\sum_{i=j+2}^{M} \epsilon_i h(\tilde{\mathbf{x}}_i) - L_j^h \right\}. \tag{9.8}$$

The following fact is a consequence of our definitions.

Lemma 9.5. We have $R_j = \mathbb{E}_{\tilde{\mathbf{x}},\epsilon}\left[f_{\mathbf{x}^j,\tilde{\mathbf{x}}_{j+2}^M,\epsilon_{j+1}^M,y^j}(\tilde{\mathbf{x}}_{j+1})\right]$ and $\tilde{R}_j = \mathbb{E}_{\mathbf{x}\sim\mu}\mathbb{E}_{\tilde{\mathbf{x}},\epsilon}\left[f_{\mathbf{x}^j,\tilde{\mathbf{x}}_{j+2}^M,\epsilon_{j+1}^M,y^j}(\mathbf{x})\right]$.

Let $\mathbf{z}_j = (\mathbf{x}^j, \tilde{\mathbf{x}}_{j+2}^M, \epsilon_{j+1}^M)$. We now observe the following key properties of the functions $f_{\mathbf{z}_j,y^j}(\mathbf{x})$, which demonstrates that $f_{\mathbf{z}_j,y^j}(\mathbf{x})$ has *sensitivity* upper bounded by $4L$ and is Lipschitz on y^j.

Proposition 9.1. For any $\mathbf{z}_j$ and y^j, we have $\sup_{\mathbf{x},\mathbf{x}'} |f_{\mathbf{z}_j,y^j}(\mathbf{x}) - f_{\mathbf{z}_j,y^j}(\mathbf{x}')| \leq 4L$. Moreover, for all $\mathbf{z}_j$, $\mathbf{x}$ and $y^j, y'^j \in [0,1]^j$, we have $|f_{\mathbf{z}_j,y^j}(\mathbf{x}) - f_{\mathbf{z}_j,y'^j}(\mathbf{x})| \leq jL||y^j - y'^j||_\infty$.

Proof. Denote

$$F(h) = 2L\sum_{i=j+2}^{M} \epsilon_i h(\tilde{\mathbf{x}}_i) - L_j^h.$$

Let $\hat{h} = \arg\max_{h\in\mathcal{H}} F(h)$ (find an approximation if necessary). We claim that for any $\mathbf{x} \in \mathcal{X}$,

$$F(\hat{h}) - 2L \leq \sup_{h\in\mathcal{H}} \{2\epsilon_{j+1}Lh(\mathbf{x}) + F(h)\} \leq F(\hat{h}) + 2L.$$

This will complete the proof of the first part. To see the upper bound, we have

$$\sup_{h\in\mathcal{H}} \{2\epsilon_{j+1}Lh(\mathbf{x}) + F(h)\} \leq \sup_{h}\{2\epsilon_{j+1}Lh(\mathbf{x})\} + \sup_{h} F(h) \leq 2L + F(\hat{h}),$$

since $h(\mathbf{x}) \in [0,1]$. For the lower bound, we have

$$\sup_{h\in\mathcal{H}}\{2\epsilon_{j+1}Lh(\mathbf{x}) + F(h)\} \geq 2\epsilon_{j+1}L\hat{h}(\mathbf{x}) + F(\hat{h}) \geq F(\hat{h}) - 2L,$$

since sup do not increase by replacing h with any specific $\hat{h}$ and $\hat{h}(\mathbf{x}) \in [0,1]$.

To prove the second part, for any given $h \in \mathcal{H}$, we denote

$$g_h(y^j) = 2L\epsilon_{j+1}h(\mathbf{x}) + 2L\sum_{i=j+2}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i) - L_j^h.$$

Note that, y^j only appears in the L_j^h term. By definition of L_j^h and L-Lipschitz property of the loss ℓ, we have

$$\forall h \in \mathcal{H},\ |g_h(y^j) - g_h(y'^j)| \leq jL||y^j - y'^j||_\infty.$$

Let $\hat{h} = \arg\max_h g_h(y^j)$, we have

$$\sup_h g_h(y^j) - \sup_h g_h(y'^j) \leq g_{\hat{h}}(y^j) - g_{\hat{h}}(y'^j) \leq jL||y^j - y'^j||_\infty.$$

Let $\hat{h}' = \arg\max_h g_h(y'^j)$, we have

$$\sup_h g_h(y^j) - \sup_h g_h(y'^j) \geq g_{\hat{h}'}(y^j) - g_{\hat{h}'}(y'^j) \geq -jL||y^j - y'^j||_\infty.$$

The proposition follows by noticing that $f_{\mathbf{z}_j,y^j}(\mathbf{x}) = \sup_h g_h(y^j)$. □

Note that Proposition 9.1 and Lemma 9.5 immediately imply that $\tilde{R}_j - R_j \leq 4L||\mu - \hat{\mu}_N||_{\mathsf{TV}}$ [4]. Unfortunately, we are unable to bound the total variation distance $||\mu - \hat{\mu}_N||_{\mathsf{TV}}$ due to the lack of any structure we impose on μ. We instead establish the following key technical result, which bounds the discrepancies via a Rademacher sum of the functions $f_{\mathbf{z}_j,y^j}$. This result constitutes the main technical ingredient in our next analysis.

Lemma 9.6. For all $j \in [M-1]$, $M \leq N/2$ and $B = N - M + j + 1$, we find

$$\mathbb{E}_{\mathbf{x}_{-N+1}^0}\mathbb{E}_{\mathbf{x}^j}\sup_{y^j}(\tilde{R}_j - R_j) \leq \sup_{\mathbf{x}_{-N+1}^{-N+B},\mathbf{x}'^B,\mathbf{z}_j} \mathbb{E}_{\epsilon'^B}\left[\sup_{y^j}\frac{1}{B}\sum_{i=1}^{B}\epsilon_i'(f_{\mathbf{z}_j,y^j}(\mathbf{x}_i') - f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i}))\right],$$

where $\mathbf{x}_{-N+1}^{-N+B}, \mathbf{x}'^B, \mathbf{z}_j$ run over all possible values and ϵ'^B is distributed uniformly over $\{\pm 1\}^B$.

[4] Using the fact that $\mathbb{E}_{\mathbf{x}\sim\mu}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\nu}[f(\mathbf{x})] \leq \sup_{\mathbf{x},\mathbf{x}'}|f(\mathbf{x}) - f(\mathbf{x}')| \cdot ||\mu - \nu||_{\mathsf{TV}}$.

Sketch of Proof. We highlight only the main idea here and refer to Section 9.2.5 for the complete proof. By Lemma 9.5, we can upper bound the discrepancies by $\mathbb{E}_{\mathbf{x}^0_{-N+1}}\mathbb{E}_{\mathbf{z}_j}\sup_{y^j}[\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})]-\mathbb{E}_{\tilde{\mathbf{x}}_{j+1}}[f_{\mathbf{z}_j,y^j}(\tilde{\mathbf{x}}_{j+1})]]$, where $\mathbf{z}_j=(\mathbf{x}^j,\tilde{\mathbf{x}}^M_{j+2},\epsilon^M_{j+1})$. Note that $\tilde{\mathbf{x}}^M_{j+1}$ is sampled uniformly from $\mathbf{x}^0_{-N+1}$ *without replacement* as in (9.2). Therefore, the randomness of $\tilde{\mathbf{x}}^M_{j+1}$ can be described as follows: we first sample $\tilde{\mathbf{x}}^M_{j+2}$ from $\mathbf{x}^0_{-N+1}$ and then sample $\tilde{\mathbf{x}}_{j+1}$ *uniformly* from the remaining samples in $\mathbf{x}^0_{-N+1}$. Now, the key observation is that, by symmetries of $\mathbf{x}^0_{-N+1}$ (which are *i.i.d.*), we can *fix* $\tilde{\mathbf{x}}^M_{j+2}$ being the last $M-j-1$ samples in $\mathbf{x}^0_{-N+1}$. Therefore, we have $\mathbb{E}_{\tilde{\mathbf{x}}_{j+1}}[f_{\mathbf{z}_j,y^j}(\tilde{\mathbf{x}}_{j+1})]=\frac{1}{B}\sum_{i=1}^B f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i})$, where $B=N-M+j+1$. Since $\mathbf{z}_j$ is *decoupled* from $\mathbf{x}^{-N+B}_{-N+1}$ by our construction, we obtain the upper bound $\mathbb{E}_{\mathbf{z}_j}\mathbb{E}_{\mathbf{x}^{-N+B}_{-N+1}}\sup_{y^j}[\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})]-\frac{1}{B}\sum_{i=1}^B f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i})]$. The lemma then follows by *symmetrization* with $\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})]$ (see Section 9.2.5). □

For any $j\in[M-1]$ and $\mathbf{z}_j$ as above, we define the following function class [5]:

$$\mathcal{G}_{\mathbf{z}_j}=\{g_{\mathbf{z}_j,y^j}(\mathbf{x},\mathbf{x}')\overset{\text{def}}{=}f_{\mathbf{z}_j,y^j}(\mathbf{x}')-f_{\mathbf{z}_j,y^j}(\mathbf{x}):y^j\in[0,1]^j,(\mathbf{x},\mathbf{x}')\in\mathcal{X}^2\}. \tag{9.9}$$

Lemma 9.6 essentially states that the discrepancy between R_j and $\tilde{R}_j$ is upper bounded by the Rademacher complexity of the class $\mathcal{G}_{\mathbf{z}_j}$ as $\mathbb{E}_{\mathbf{x}^0_{-N+1}}\mathbb{E}_{\mathbf{x}^j}\sup_{y^j}(\tilde{R}_j-R_j)\le\sup_{\mathbf{z}_j}\frac{1}{B}\mathsf{Rad}_B(\mathcal{G}_{\mathbf{z}_j})$.

The following lemma provides a useful bound on such Rademacher complexities.

Lemma 9.7. Let $\mathcal{G}_{\mathbf{z}_j}$ be as in (9.9), $M\le N/2$ and $B=N-M+j+1$. Then

$$\sup_{\mathbf{z}_j}\frac{1}{B}\mathsf{Rad}_B(\mathcal{G}_{\mathbf{z}_j})\le O\left(\sqrt{\frac{jL^2\log(jLB)}{B}}\right)\le O\left(\sqrt{\frac{2jL^2\log(jLN/2)}{N}}\right). \tag{9.10}$$

Proof. Let $\mathcal{C}\subset[0,1]^j$ be a covering of $[0,1]^j$ with norm L_∞ radius $\frac{1}{jLB}$. We have $|\mathcal{C}|\le(jLB)^j$. By the second part of Proposition 9.1,

[5] Note that the "complexity" of $\mathcal{G}_{\mathbf{z}_j}$ arises from the $y^j\in[0,1]^j$.

we know that the class $\mathcal{G}'_{\mathbf{z}_j} \overset{\text{def}}{=} \{g_{\mathbf{z}_j, y^j} : y^j \in \mathcal{C}\}$ forms a uniform L_∞-covering of $\mathcal{G}_{\mathbf{z}_j}$ with radius $\frac{2}{B}$. Therefore, $\frac{1}{B}\mathsf{Rad}_B(\mathcal{G}_{\mathbf{z}_j}) \leq \frac{1}{B}\mathsf{Rad}_B(\mathcal{G}'_{\mathbf{z}_j}) + \frac{2}{B}$. The first inequality then follows by a simple application of Massart's lemma [17, Lemma 26.8] over $\mathcal{G}'_{\mathbf{z}_j}$, since $|\mathcal{G}'_{\mathbf{z}_j}| \leq |\mathcal{C}| \leq (jLB)^j$ and $\sup_{(\mathbf{x},\mathbf{x}')\in\mathcal{X}^2}\{g_{\mathbf{z}_j,y^j}(\mathbf{x},\mathbf{x}')\} \leq 4L$ for all $g_{\mathbf{z}_j,y^j} \in \mathcal{G}_{\mathbf{z}_j}$ due to the first part of Proposition 9.1. The second inequality is implied by that $B \geq N/2$ and the fact that the function $\frac{\log B}{B}$ is monotone decreasing. □

Putting everything together, we arrive at:

Theorem 9.8. Let Φ be the predictor as in (9.2) and $M \leq N/2$. Then for any class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ with a convex and L-Lipschitz loss ℓ, the predictor Φ can be computed efficiently with access to at most $O(L\sqrt{M}\log M)$ mixed-ERM oracle calls per round such that

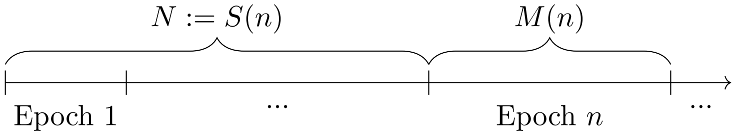
$$\tilde{r}^{\mathsf{side}}_{M,N}(\mathcal{H},\Phi) \leq 2L\mathsf{Rad}_M(\mathcal{H}) + \sqrt{M} + O\left(\sqrt{\frac{M^3L^2\log(MLN)}{N}}\right). \quad (9.11)$$

Proof. The regret bound follows directly from Lemma 9.2, Lemma 9.4 and Lemma 9.6. We then invoke Lemma 9.7 to bound the discrepancies by noticing that $j \leq M$. □

Remark 9.2. Note that Theorem 9.8 shows that if $N \gg M^2 \log M$ then the regret with side-information is reduced to the Rademacher complexities of $\mathcal{H}$, and thus matches the case when the distribution is known in advance. However, in reality such side-information is not available for the unknown distribution case, which can only be obtained from prior samples.

9.2.2 Proof of Theorem 9.1 of the Orcale Efficient Predictor

We are now equipped with all the technical tools to prove Theorem 9.1, with the only missing ingredient of constructing the *side-information*. For this purpose, we employ an *epoch-based* approach, resembling those used in Chapter 8.3.1, but in a completely different context. We partition the time horizon into epochs, with epoch n of length $M(n)$. Let $S(n) = \sum_{i=1}^{n-1} M(i)$ be the total time steps after $n-1$ epochs. We will use the features observed upto time $S(n)$ as the side-information introduces in Section 9.2.1 (denoted there as $\mathbf{x}^0_{N+1}$) and apply the predictor constructed in (9.2) to make the prediction during the nth epoch.



To this end, our main technical part is to *optimize* the epoch length $M(n)$ that balances the trade-off in (9.11) and achieving the minimal total regret. Let Φ be the predictor derived from (9.2), which we write as $\Phi(\mathbf{x}^0_{-N+1}, \mathbf{x}^j, y^{j-1})$ for the side-information $\mathbf{x}^0_{-N+1}$, features $\mathbf{x}^j$ and labels y^{j-1} observed thus far. We define the following *epoch* predictor Ψ: for any epoch n and time step j during such epoch, we set

$$\Psi(\mathbf{x}^{S(n)+j}, y^{S(n)+j-1}) = \Phi\left(\mathbf{x}^{S(n)}, \mathbf{x}^{S(n)+j}_{S(n)+1}, y^{S(n)+j-1}_{S(n)+1}\right). \tag{9.12}$$

Let $S^{-1}(T)$ be the largest number n such that $S(n) < T$. The following lemma upper bounds the hybrid minimax regret of Ψ using the regrets with side information (9.3) incurred by Φ. Note that this is *not* immediately obvious since we have *reused* the side-information among different epochs.

Lemma 9.9. For any $\mathcal{H}$ and convex L-Lipschitz loss ℓ, we have

$$\tilde{r}_T(\mathcal{H}, \Psi) \leq \sum_{n=1}^{S^{-1}(T)} \tilde{r}^{\mathsf{side}}_{M(n),S(n)}(\mathcal{H}, \Phi).$$

Proof. Define the operator $\mathbb{Q}_i^j \equiv \mathbb{E}_{\mathbf{x}_i} \sup_{y_i} \mathbb{E}_{\hat{y}_i} \cdots \mathbb{E}_{\mathbf{x}_j} \sup_{y_j} \mathbb{E}_{\hat{y}_j}$, where $\hat{y}_t \sim \Psi(\mathbf{x}^t, y^{t-1})$ for all $t \in [T]$. We have (truncate the last $S(n+1)$

above T if necessary):

$$\begin{aligned}\tilde{r}_T(\mathcal{H},\Psi) &= \mathbb{Q}_1^T \sup_{h\in\mathcal{H}}\left[\sum_{n=1}^{S^{-1}(T)}\sum_{j=S(n)+1}^{S(n+1)} \ell(\hat{y}_j,y_j)-\ell(h(\mathbf{x}_j),y_j)\right]\\ &\overset{(a)}{\le} \sum_{n=1}^{S^{-1}(T)} \mathbb{Q}_1^T \sup_{h\in\mathcal{H}}\left[\sum_{j=S(n)+1}^{S(n+1)} \ell(\hat{y}_j,y_j)-\ell(h(\mathbf{x}_j),y_j)\right]\\ &\overset{(b)}{=} \sum_{n=1}^{S^{-1}(T)} \mathbb{E}_{\mathbf{x}^{S(n)}}\mathbb{Q}_{S(n)+1}^{S(n+1)} \sup_{h\in\mathcal{H}}\left[\sum_{j=S(n)+1}^{S(n+1)} \ell(\hat{y}_j,y_j)-\ell(h(\mathbf{x}_j),y_j)\right]\\ &\overset{(c)}{=} \sum_{n=1}^{S^{-1}(T)} \tilde{r}^{\mathsf{side}}_{M(n),S(n)}(\mathcal{H},\Phi),\end{aligned}$$

where (a) follows by $\sup(A+B)\le\sup A+\sup B$ and linearity of expectation; (b) follows since $\hat{y}_j$ depends only on $\mathbf{x}^j$ and $y^j_{S(n)}$ for $j\in(S(n),S(n+1)]$; (c) follows by definition. □

Proof of Theorem 9.1. Assume $\mathsf{Rad}_T(\mathcal{H})\le O(T^q)$ for some $q\in[\frac{1}{2},1]$. By Theorem 9.8 and $M(n),S(n)\le T$ we have

$$\tilde{r}^{\mathsf{side}}_{M(n),S(n)}(\mathcal{H},\Phi)\le O\left(LM(n)^q+\sqrt{\frac{M(n)^3L^2\log(LT^2)}{S(n)}}\right).$$

Let $M(n)=n^\alpha$ for some $\alpha>0$ to be determined later. We have $S(n)=\sum_{i=1}^{n-1} i^\alpha=\Theta(n^{\alpha+1})$ by integration approximation, and $S^{-1}(T)\le O(T^{1/(\alpha+1)})$. This implies that

$$\tilde{r}^{\mathsf{side}}_{M(n),S(n)}(\mathcal{H},\Phi)\le O(Ln^{\alpha q}+L\sqrt{\log(LT^2)}n^{\alpha-\frac{1}{2}}).$$

By Lemma 9.9 and integration approximation again, we conclude

$$\tilde{r}_T(\mathcal{H},\Psi)\le O\left(LT^{\frac{\alpha q+1}{\alpha+1}}+L\sqrt{\log(LT^2)}T^{\frac{\alpha+\frac{1}{2}}{\alpha+1}}\right). \tag{9.13}$$

Optimizing $\arg\min_{\alpha>0}\max\{\frac{\alpha q+1}{\alpha+1},\frac{\alpha+\frac{1}{2}}{\alpha+1}\}$, we find (9.13) is minimized when taking $\alpha=\frac{1}{2(1-q)}$. Plugging back to (9.13), we find $\tilde{r}_T(\mathcal{H},\Psi)\le O\left(L\sqrt{\log(LT)}T^{\frac{2-q}{3-2q}}\right)$. This completes the proof of the first part. The second and third parts follow by the facts that $\mathsf{Rad}_T(\mathcal{H}) \leq O(\sqrt{\mathsf{VC}(\mathcal{H})T})$ for finite-VC class [34, Example 5.24], and $\mathsf{Rad}_T(\mathcal{H}) \leq \tilde{O}(T^{\max\{\frac{1}{2},\frac{p-1}{p}\}})$ for classes with α-fat shattering dimension of order α^{-p} [63]. This completes the proof and the big-O notations and $M(n) \leq S(n)/2$ are justified by noting that $\alpha \geq 1$ since $q \geq \frac{1}{2}$. □

9.2.3 Tighter Bounds for Special Classes

As demonstrated in Section 9.2.1, the main technical obstacle for analyzing the regret is to upper bound the discrepancies between $\tilde{R}_j$ and R_j as in Lemma 9.4. It was shown in Lemma 9.6 that such discrepancies can be upper bounded by the Rademacher complexity of the class $\mathcal{G}_{\mathbf{z}_j}$ in (9.9). We demonstrate in this section how to leverage the *structural* information of $\mathcal{G}_{\mathbf{z}_j}$ leading to tighter regret bounds for certain special classes when compared to the general bounds from Theorem 9.1.

Binary valued classes. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary valued class and $\ell(\hat{y},y) = |\hat{y}-y|$. For any given $\mathbf{z}_j$ (assume, w.l.o.g., $\epsilon_{j+1} = 1$) and $y^j \in \{0,1\}^j$, the function $f_{\mathbf{z}_j,y^j}$ can be expressed as $f_{\mathbf{z}_j,y^j}(\mathbf{x}) = \sup_h\{2h(\mathbf{x}) + F(h)\}$ (see definition in (9.8)), where $F(h)$ is a discrete valued function taking values in $[-2M,2M]$. Define

$$\mathcal{H}^0 = \left\{h \in \mathcal{H} : F(h) = \sup_{h' \in \mathcal{H}} F(h')\right\}$$

and

$$\mathcal{H}^1 = \left\{h \in \mathcal{H} : F(h) = \sup_{h' \in \mathcal{H}} F(h') - 1\right\}.$$

Let $h^0(\mathbf{x}) = \sup_{h\in\mathcal{H}^0}\{h(\mathbf{x})\}$, $h^1(\mathbf{x}) = \sup_{h\in\mathcal{H}^1}\{h(\mathbf{x})\}$ and $\hat{h} = \arg\max_{h\in\mathcal{H}} F(h)$. The following *structural* characterization of $f_{\mathbf{z}_j,y^j}$ holds:

Lemma 9.10. For any $\mathbf{x} \in \mathcal{X}$, we have:

$$f_{\mathbf{z}_j,y^j}(\mathbf{x}) = \begin{cases} F(\hat{h}) + 2, & \text{if } h^0(\mathbf{x}) = 1 \\ F(\hat{h}) + 1, & \text{if } h^0(\mathbf{x}) = 0 \text{ and } h^1(\mathbf{x}) = 1 \\ F(\hat{h}), & \text{else} \end{cases}. \tag{9.14}$$

Proof. Note that $f_{\mathbf{z}_j,y^j}(\mathbf{x}) = \sup_h\{2h(\mathbf{x}) + F(h)\}$. If $h^0(\mathbf{x}) = 1$, then $\exists h \in \mathcal{H}^0$ such that $h(\mathbf{x}) = 1$ and $F(h) = F(\hat{h})$, thus $f_{\mathbf{z}_j,y^j}(\mathbf{x}) \geq 2+F(\hat{h})$. Clearly, we also have $f_{\mathbf{z}_j,y^j}(\mathbf{x}) \leq 2\sup_h h(\mathbf{x}) + \sup_h F(h) \leq 2 + F(\hat{h})$, the first case follows. If $h^0(\mathbf{x}) = 0$ and $h^1(\mathbf{x}) = 1$, then there exists $h \in \mathcal{H}^1$ such that $h(\mathbf{x}) = 1$ and $F(h) = F(\hat{h}) - 1$, thus $f_{\mathbf{z}_j,y^j}(\mathbf{x}) \geq F(\hat{h})-1+2 = F(\hat{h})+1$. On the other-hand, since $h^0(\mathbf{x}) = 0$, we have for all $h \in \mathcal{H}^0$, $2h(\mathbf{x}) + F(h) = F(\hat{h})$. For any other $h \notin \mathcal{H}^0 \cup \mathcal{H}^1$, we have $2h(\mathbf{x})+F(h) \leq 2+F(\hat{h})-2 = F(\hat{h})$. Therefore, $f_{\mathbf{z}_j,y^j}(\mathbf{x}) \leq F(\hat{h})+1$, this completes the second case. Finally, if both $h^0(\mathbf{x}) = h^1(\mathbf{x}) = 0$, we have for any $h \in \mathcal{H}^0$, $2h(\mathbf{x}) + F(h) = F(\hat{h})$, i.e., $f_{\mathbf{z}_j,y^j}(\mathbf{x}) \geq F(\hat{h})$. Moreover, for any $h \notin \mathcal{H}^0$, it is easy to verify that $2h(\mathbf{x}) + F(h) \leq F(\hat{h})$. □

Theorem 9.11. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$, $\mathcal{F}^{\mathsf{u}} = \{f_{\mathcal{H}'}(\mathbf{x}) = \sup_{h\in\mathcal{H}'}\{h(\mathbf{x})\} : \mathcal{H}' \subset \mathcal{H}\}$, $\mathcal{F}^{\mathsf{i}} = \{f_{\mathcal{H}'}(\mathbf{x}) = \inf_{h\in\mathcal{H}'}\{h(\mathbf{x})\} : \mathcal{H}' \subset \mathcal{H}\}$ be two classes of functions and ℓ be the absolute loss. Then there exists an oracle-efficient predictor Φ satisfying $\tilde{r}_T(\mathcal{H}, \Phi) \leq O(\sqrt{\max\{\mathsf{VC}(\mathcal{F}^{\mathsf{u}}), \mathsf{VC}(\mathcal{F}^{\mathsf{i}})\}T})$.

Proof. Assume, w.o.l.g., $\epsilon_{j+1} = 1$. The functions h^0, h^1 as in Fact 9.10 are within $\mathcal{F}^{\mathsf{u}}$. For *any* $\mathbf{x}^{2N} \in \mathcal{X}^{2N}$ and $\hat{\mu}$ uniform over $\mathbf{x}^{2N}$, there exists a γ-cover $\mathcal{C}_\gamma$ of $\mathcal{F}^{\mathsf{u}}$ under distance $d_{\hat{\mu}}(f_1, f_2) \overset{\text{def}}{=} \Pr_{\mathbf{x}\sim\hat{\mu}}[f_1(\mathbf{x}) \neq f_2(\mathbf{x})]$ such that $|\mathcal{C}_\gamma| \leq O(\frac{1}{\gamma^{\mathsf{VC}(\mathcal{F}^{\mathsf{u}})}})$ [69]. By Fact 9.10, there exists a function $\mathcal{T} : (\mathcal{F}^{\mathsf{u}})^2 \to \{0,1,2\}^{\mathcal{X}}$ such that for *any* $f_{\mathbf{z}_j,y^j}$, there exist $h^0, h^1 \in \mathcal{F}^{\mathsf{u}}$ such that $f_{\mathbf{z}_j,y^j}(\mathbf{x}) = \mathcal{T}(h^0(\mathbf{x}), h^1(\mathbf{x})) + c_{\mathbf{z}_j,y^j}$, where $c_{\mathbf{z}_j,y^j} = F(\hat{h})$ as in Fact 9.10. Therefore, the function class $\mathcal{C}' \overset{\text{def}}{=} \{\mathcal{T}(h^0, h^1) : h^0, h^1 \in \mathcal{C}_\gamma\}$ forms a 2γ-cover of $\{(f_{\mathbf{z}_j,y^j}(\mathbf{x}) - c_{\mathbf{z}_j,y^j}) : y^j \in [0,1]^j\}$ under distance $d_{\hat{\mu}}(f_1, f_2)$ and $|\mathcal{C}'| \leq O(\frac{1}{\gamma^{2\mathsf{VC}(\mathcal{F}^{\mathsf{u}})}})$. This implies that the function class $\mathcal{C}'' \overset{\text{def}}{=} \{g(\mathbf{x}', \mathbf{x}) = f(\mathbf{x}') - f(\mathbf{x}) : f \in \mathcal{C}',\ (\mathbf{x}', \mathbf{x}) \in \mathcal{X}^2\}$ forms a 4γ-cover of

$$\mathcal{G}_{\mathbf{z}_j} = \{g_{\mathbf{z}_j,y^j}(\mathbf{x}', \mathbf{x}) = f_{\mathbf{z}_j,y^j}(\mathbf{x}') - f_{\mathbf{z}_j,y^j}(\mathbf{x}) : y^j \in [0,1]^j,\ (\mathbf{x}', \mathbf{x}) \in \mathcal{X}^2\}$$

under distance $d_{\hat{\nu}}(g_1, g_2) = \Pr_{(\mathbf{x}',\mathbf{x})\sim\hat{\nu}}[g_1(\mathbf{x}', \mathbf{x}) \neq g_2(\mathbf{x}', \mathbf{x})]$ for any distribution $\hat{\nu}$ uniform over a fixed *pairing* of $\mathbf{x}^{2N}$ and $|\mathcal{C}''| \leq O(\frac{1}{\gamma^{2\mathsf{VC}(\mathcal{F}^{\mathsf{u}})}})$. We have by the chaining argument [34, Example 5.24] that $\mathsf{Rad}_N(\mathcal{G}_{\mathbf{z}_j}) \leq$

$O(\sqrt{\mathsf{VC}(\mathcal{F}^{\mathsf{u}})N})$. This implies by Lemma 9.4 & 9.6 that

$$\tilde{r}^{\mathsf{side}}_{M,N}(\mathcal{H},\Phi) \leq O\left(\sqrt{\mathsf{VC}(\mathcal{H})M} + \frac{M\sqrt{\mathsf{VC}(\mathcal{F}^{\mathsf{u}})}}{\sqrt{N}}\right). \tag{9.15}$$

Taking $M(n) = 1.5^n$ in (9.12), we have $N = S(n) = 2 \cdot 1.5^n - 3$, which ensures $M(n) \leq S(n)/2 + O(1)$ (as required for (9.15) to hold). Invoking Lemma 9.9, we conclude

$$\tilde{r}_T(\mathcal{H},\Psi) \leq O(\sqrt{\mathsf{VC}(\mathcal{H})} + \sqrt{\mathsf{VC}(\mathcal{F}^{\mathsf{u}})}) \sum_{n=1}^{\lceil \log_{1.5}(T) \rceil} 1.5^{n/2} \leq O(\sqrt{\mathsf{VC}(\mathcal{F}^{\mathsf{u}})T}),$$

where the last inequality follows by $\mathcal{H} \subset \mathcal{F}^{\mathsf{u}}$. This completes the proof and the case for $\epsilon_{j+1} = -1$ is symmetric with $\mathcal{F}^{\mathsf{i}}$. □

Note that for the threshold functions $\mathcal{H} = \{1\{x \geq a\} : a, x \in [0,1]\}$ we have $\mathcal{F}^{\mathsf{u}} = \mathcal{F}^{\mathsf{i}} = \mathcal{H}$. Theorem 9.11 implies an oracle efficient $O(\sqrt{T})$ regret, which matches the information-theoretical lower bound and is tighter than the covering-based $O(\sqrt{T \log T})$ bound implied by [60]. Another example is the class of indicators of intervals with bounded length $\{1\{x \in [a,b]\} : b - a \geq \gamma, [a,b] \subset [0,1]\}$, for which we have $\mathsf{VC}(\mathcal{F}^{\mathsf{i}}) = 2$ and $\mathsf{VC}(\mathcal{F}^{\mathsf{u}}) \leq O(\frac{1}{\gamma})$.

Lipschitz functions. Let $\mathcal{X} = [0,1]^d$ and $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be the class of *all* 1-Lipschitz functions under L_∞ norm. Assume $\ell(\hat{y}, y) = |\hat{y} - y|$ is the absolute loss. Let μ and $\hat{\mu}_N$ be the true and empirical distributions, respectively, as in Section 9.2.1. By Lemma 9.5 and assuming that $\tilde{\mathbf{x}}^M_{j+1}$ is sampled *i.i.d.* from $\hat{\mu}_N$, we have $\mathbb{E}_{\mathbf{x}^j_{-N+1}} \sup_{y^j}(\tilde{R}_j - R_j) \leq \mathbb{E}_{\mathbf{x}^0_{-N+1}} \sup_{y^j,\mathbf{z}_j}(\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\hat{\mu}_N}[f_{\mathbf{z}_j,y^j}(\mathbf{x})])$. By the same argument as Proposition 9.1 (second part) and Lipschitz property of $h \in \mathcal{H}$, we have:

Fact 1. For all $\mathbf{z}_j, y^j$ and $\mathbf{x}, \mathbf{x}'$, $|f_{\mathbf{z}_j,y^j}(\mathbf{x}) - f_{\mathbf{z}_j,y^j}(\mathbf{x}')| \leq 2||\mathbf{x} - \mathbf{x}'||_\infty$.

Theorem 9.12. Let $\mathcal{H}$ and ℓ be as above. Then, there exists an oracle-efficient predictor Φ such that $\tilde{r}_T(\mathcal{H},\Phi) \leq \tilde{O}(T^{\max\{\frac{1}{2},\frac{d-1}{d}\}})$, and this bound is tight upto poly-logarithmic factors.

Proof. By Fact 1, we know that for all $\mathbf{z}_j, y^j$ the function $f_{\mathbf{z}_j,y^j}(\mathbf{x})$ is 2-Lipschitz. Therefore, by Kantorovich-Rubinstein duality [73] we have $\sup_{y^j,\mathbf{z}_j}(\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\hat{\mu}_N}[f_{\mathbf{z}_j,y^j}(\mathbf{x})]) \le 2W_1(\mu,\hat{\mu}_N)$, where $W_1(\mu,\hat{\mu}_N) = \inf_{\gamma\in\Gamma(\mu,\hat{\mu}_N)} \mathbb{E}_{(\mathbf{x},\mathbf{x}')\sim\gamma}[||\mathbf{x}-\mathbf{x}'||_\infty]$ is the Wasserstein 1-distance with $\Gamma(\mu,\hat{\mu}_N)$ being the class of all coupling between $\mu,\hat{\mu}_N$. Therefore, we have $\mathbb{E}_{\mathbf{x}^j_{-N+1}} \sup_{y^j}(\tilde{R}_j - R_j) \le 2\mathbb{E}_{\mathbf{x}^0_{-N+1}}[W_1(\mu,\hat{\mu}_N)]$, i.e., the discrepancy is upper bounded by the convergence rate of empirical distribution under Wasserstein 1-distance. Invoking [74, Thm 1] and boundedness of $\mathcal{X}$, we have $\mathbb{E}_{\mathbf{x}^0_{-N+1}}[W_1(\mu,\hat{\mu}_N)] \le \tilde{O}(N^{-1/d})$. Let Φ be the predictor in (9.2). By Lemma 9.4 and $\mathsf{Rad}_M(\mathcal{H}) \le \tilde{O}(M^{\max\{\frac{1}{2},\frac{d-1}{d}\}})$ [34], we have

$$\tilde{r}^{\mathsf{side}}_{M,N}(\mathcal{H}) \le \tilde{O}(M^{\max\{\frac{1}{2},\frac{d-1}{d}\}} + MN^{-1/d}).$$

The result then follows by Lemma 9.9 with $M(n) = 2^n$ (which ensures $N = S(n) = M(n) - 1$). The last part follows by that the ϵ-metric entropy of $\mathcal{H}$ is $\Theta(\frac{1}{\epsilon^d})$ [34]. □

Remark 9.3. Note that, if we assume certain structure on μ that admits a computationally efficient estimator $\hat{\mu}_N$ that satisfies $||\mu - \hat{\mu}_N||_{\mathsf{TV}} \le O(\frac{1}{\sqrt{N}})$ (such as for Gaussian distributions [75]), then the (optimal) $O(\mathsf{Rad}_T(\mathcal{H}) + \sqrt{T})$ bound is achievable for *any* class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$.

9.2.4 Proof of Lemma 9.3

In this section, we establish the *approx-admissibility* of our predictor in (9.2). The reasoning follows closely to the arguments as in [26, Lemma 11&12] but needs careful adaption for handling the dummy samples $\tilde{\mathbf{x}}$s generated from $\hat{\mu}_N$. We have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_j} \sup_{y_j} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\left[\ell(\hat{y}_j, y_j) + R_j\right] \\
&\quad = \mathbb{E}_{\mathbf{x}_j} \sup_{y_j} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\left[\ell(\hat{y}_j, y_j) + \sup_{h\in\mathcal{H}} 2L \sum_{i=j+1}^{M} \epsilon_i h(\tilde{\mathbf{x}}_i) - L^h_j\right] \\
&\quad \overset{(a)}{\le} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\sup_{y_j} \ell(\hat{y}_j, y_j) + \sup_{h\in\mathcal{H}} 2L \sum_{i=j+1}^{M} \epsilon_i h(\tilde{\mathbf{x}}_i) - \ell(h(\mathbf{x}_j), y_j) - L^h_{j-1}\right]
\end{aligned}$$

$$\begin{aligned}
&\overset{(b)}{\leq} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\inf_{\hat{y}}\sup_{y_j}\ell(\hat{y},y_j)+\sup_{h\in\mathcal{H}}2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i)-\ell(h(\mathbf{x}_j),y_j)-L^h_{j-1}\right]\\
&= \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\inf_{\hat{y}}\sup_{y_j}\sup_{h\in\mathcal{H}}2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i)-L^h_{j-1}+\ell(\hat{y},y_j)-\ell(h(\mathbf{x}_j),y_j)\right]\\
&\overset{(c)}{\leq} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\inf_{\hat{y}}\sup_{y_j}\sup_{h\in\mathcal{H}}2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i)-L^h_{j-1}+\partial\ell(\hat{y},y_j)(\hat{y}-h(\mathbf{x}_j))\right]\\
&\overset{(d)}{\leq} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\inf_{\hat{y}}\sup_{y_j}\sup_{g_j\in[-L,L]}\sup_{h\in\mathcal{H}}2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i)-L^h_{j-1}+g_j(\hat{y}-h(\mathbf{x}_j))\right]\\
&\overset{(e)}{\leq} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\inf_{\hat{y}}\sup_{g_j\in\{-L,L\}}\sup_{h\in\mathcal{H}}2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i)-L^h_{j-1}+g_j(\hat{y}-h(\mathbf{x}_j))\right]
\end{aligned}$$

where (a) follows by the definition of R_j and that $\hat{y}_j$ has the same randomness as R_j (i.e, the $\tilde{\mathbf{x}}$s and ϵs); (b) is due to definition of $\hat{y}_j$; (c) is due to convexity of ℓ; (d) is due to L-Lipschitz property of ℓ; (e) follows by that the inner function is convex w.r.t. g_j and thus the $\sup_{g_j\in[-L,L]}$ is attained on the boundary $\{-L,L\}$. We have

$$\begin{aligned}
&\mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\inf_{\hat{y}}\sup_{g_j\in\{-L,L\}}\sup_{h\in\mathcal{H}}2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i)-L^h_{j-1}+g_j(\hat{y}-h(\mathbf{x}_j))\right]\\
&\quad\overset{(a)}{=} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\inf_{\hat{y}}\sup_{d_j\in\Delta(\{-L,L\})}\mathbb{E}_{g_j\sim d_j}\left[\sup_{h\in\mathcal{H}}2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i)-L^h_{j-1}+g_j(\hat{y}-h(\mathbf{x}_j))\right]\right]\\
&\quad\overset{(b)}{=} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\sup_{d_j\in\Delta(\{-L,L\})}\inf_{\hat{y}}\mathbb{E}_{g_j\sim d_j}\left[\sup_{h\in\mathcal{H}}2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i)-L^h_{j-1}+g_j(\hat{y}-h(\mathbf{x}_j))\right]\right]\\
&\quad\overset{(c)}{=} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\sup_{d_j}\inf_{\hat{y}}\mathbb{E}_{g_j\sim d_j}\left[g_j\hat{y}+\sup_{h\in\mathcal{H}}2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i)-L^h_{j-1}-g_jh(\mathbf{x}_j)\right]\right]\\
&\quad= \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\sup_{d_j}\inf_{\hat{y}}\left(\mathbb{E}_{g_j\sim d_j}[g_j\hat{y}]+\mathbb{E}_{g_j\sim d_j}\left[\sup_{h\in\mathcal{H}}2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i)-L^h_{j-1}-g_jh(\mathbf{x}_j)\right]\right)\right]\\
&\quad\overset{(d)}{=} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\sup_{d_j}\left(\inf_{\hat{y}}\mathbb{E}_{g'_j\sim d_j}[g'_j\hat{y}]\right)+\mathbb{E}_{g_j\sim d_j}\left[\sup_{h\in\mathcal{H}}2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i)-L^h_{j-1}-g_jh(\mathbf{x}_j)\right]\right]\\
&\quad= \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\sup_{d_j}\mathbb{E}_{g_j\sim d_j}\left[\inf_{\hat{y}}\mathbb{E}_{g'_j\sim d_j}[g'_j\hat{y}]+\sup_{h\in\mathcal{H}}2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i)-L^h_{j-1}-g_jh(\mathbf{x}_j)\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\sup_{d_j}\mathbb{E}_{g_j\sim d_j}\left[\sup_{h\in\mathcal{H}} 2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i) - L_{j-1}^{h} + \inf_{\hat{y}}\mathbb{E}_{g_j'\sim d_j}[g_j'\hat{y}] - g_j h(\mathbf{x}_j)\right]\right]\\
&\overset{(e)}{\leq} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\sup_{d_j}\mathbb{E}_{g_j\sim d_j}\left[\sup_{h\in\mathcal{H}} 2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i) - L_{j-1}^{h} + \mathbb{E}_{g_j'\sim d_j}[g_j' h(\mathbf{x}_j)] - g_j h(\mathbf{x}_j)\right]\right]\\
&\overset{(f)}{\leq} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\sup_{d_j}\mathbb{E}_{g_j,g_j'\sim d_j}\left[\sup_{h\in\mathcal{H}} 2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i) - L_{j-1}^{h} + (g_j' - g_j)h(\mathbf{x}_j)\right]\right]\\
&\overset{(g)}{=} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\sup_{d_j}\mathbb{E}_{g_j,g_j'\sim d_j}\mathbb{E}_{\epsilon_j}\left[\sup_{h\in\mathcal{H}} 2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i) - L_{j-1}^{h} + \epsilon_j(g_j' - g_j)h(\mathbf{x}_j)\right]\right]\\
&\overset{(h)}{\leq} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\sup_{d_j}\mathbb{E}_{g_j\sim d_j}\mathbb{E}_{\epsilon_j}\left[\sup_{h\in\mathcal{H}} 2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i) - L_{j-1}^{h} + 2\epsilon_j g_j h(\mathbf{x}_j)\right]\right]\\
&\overset{(i)}{=} \mathbb{E}_{\epsilon,\tilde{\mathbf{x}}}\mathbb{E}_{\mathbf{x}_j}\left[\mathbb{E}_{\epsilon_j}\left[\sup_{h\in\mathcal{H}} 2L\sum_{i=j+1}^{M}\epsilon_i h(\tilde{\mathbf{x}}_i) - L_{j-1}^{h} + 2\epsilon_j L h(\mathbf{x}_j)\right]\right]\\
&= \tilde{R}_{j-1},
\end{aligned}
$$

where (a) follows by $\sup_{g_j\in\{-L,L\}} \equiv \sup_{d_j\in\Delta(\{-L,L\})}\mathbb{E}_{g_j\sim d_j}$ where $\Delta(\{-L,L\})$ is the set of all probability distributions over $\{-L,L\}$; (b) follows by the minimax theorem and noticing that the inner expectation is bilinear w.r.t. $\hat{y}$ and d_j; (c) follows by the fact that $g_j\hat{y}$ is independent of $\sup_h$; (d) follows by that the $\sup_h$ term is independent of $\hat{y}$ and introducing an *i.i.d.* copy g_j' of g_j; (e) follows by the fact that replacing $\hat{y}$ with $h(\mathbf{x}_j)$ does not decrease the inf term; (f) is due to $\sup\mathbb{E} \leq \mathbb{E}\sup$; (g) is due to symmetries of g_j, g_j' and ϵ_j is uniform over $\{-1,1\}$; (h) follows by $\sup(A+B+C) \leq \sup(A/2+B)+\sup(A/2+C) = (\sup(A+2B)+\sup(A+2C))/2$, the linearity of expectation and symmetries of B, C; (i) follows by that the inner expectation takes the same value for all $g_j \in \{-L,L\}$ and therefore the $\sup_{d_j}\mathbb{E}_{g_j\sim d_j}$ can be eliminated. This completes the proof.

9.2.5 Proof of Lemma 9.6

We have

$$
\mathbb{E}_{\mathbf{x}^0_{-N+1}}\mathbb{E}_{\mathbf{x}^j}\sup_{y^j}(\tilde{R}_j - R_j)
$$

$$\begin{aligned}
&\overset{(a)}{=} \mathbb{E}_{\mathbf{x}^0_{-N+1}} \mathbb{E}_{\mathbf{x}^j} \sup_{y^j} \mathbb{E}_{\tilde{\mathbf{x}}^M_{j+1}, \epsilon^M_{j+1}} \mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x}) - f_{\mathbf{z}_j,y^j}(\tilde{\mathbf{x}}_{j+1})] \\
&\le \mathbb{E}_{\mathbf{x}^0_{-N+1}} \mathbb{E}_{\mathbf{x}^j} \mathbb{E}_{\tilde{\mathbf{x}}^M_{j+2}, \epsilon^M_{j+1}} \sup_{y^j} \mathbb{E}_{\tilde{\mathbf{x}}_{j+1},\mathbf{x}}[f_{\mathbf{z}_j,y^j}(\mathbf{x}) - f_{\mathbf{z}_j,y^j}(\tilde{\mathbf{x}}_{j+1})] \\
&\overset{(b)}{=} \mathbb{E}_{\mathbf{x}^0_{-N+1}} \mathbb{E}_{\mathbf{z}_j} \sup_{y^j} \mathbb{E}_{\tilde{\mathbf{x}}_{j+1},\mathbf{x}}[f_{\mathbf{z}_j,y^j}(\mathbf{x}) - f_{\mathbf{z}_j,y^j}(\tilde{\mathbf{x}}_{j+1})] \\
&= \mathbb{E}_{\mathbf{x}^0_{-N+1}} \mathbb{E}_{\mathbf{z}_j} \sup_{y^j} [\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})] - \mathbb{E}_{\tilde{\mathbf{x}}_{j+1}}[f_{\mathbf{z}_j,y^j}(\tilde{\mathbf{x}}_{j+1})]] \\
&\overset{(c)}{=} \mathbb{E}_{\mathbf{z}_j} \mathbb{E}_{\mathbf{x}^{-N+B}_{-N+1}} \sup_{y^j} \left[\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})] - \frac{1}{B}\sum_{i=1}^{B} f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i})\right] \\
&\overset{(d)}{=} \mathbb{E}_{\mathbf{z}_j} \mathbb{E}_{\mathbf{x}^{-N+B}_{-N+1}} \sup_{y^j} \left[\frac{1}{B}\sum_{i=1}^{B} \mathbb{E}_{\mathbf{x}'_i\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x}'_i)] - f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i})\right] \\
&= \mathbb{E}_{\mathbf{z}_j} \mathbb{E}_{\mathbf{x}^{-N+B}_{-N+1}} \sup_{y^j} \mathbb{E}_{\mathbf{x}'^B\sim\mu^{\otimes B}} \left[\frac{1}{B}\sum_{i=1}^{B} f_{\mathbf{z}_j,y^j}(\mathbf{x}'_i) - f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i})\right] \\
&\le \mathbb{E}_{\mathbf{z}_j} \mathbb{E}_{\mathbf{x}^{-N+B}_{-N+1}} \mathbb{E}_{\mathbf{x}'^B} \sup_{y^j} \left[\frac{1}{B}\sum_{i=1}^{B} f_{\mathbf{z}_j,y^j}(\mathbf{x}'_i) - f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i})\right] \\
&\overset{(e)}{=} \mathbb{E}_{\mathbf{z}_j} \mathbb{E}_{\mathbf{x}^{-N+B}_{-N+1}} \mathbb{E}_{\mathbf{x}'^B} \mathbb{E}_{\epsilon'^B} \sup_{y^j} \left[\frac{1}{B}\sum_{i=1}^{B} \epsilon'_j(f_{\mathbf{z}_j,y^j}(\mathbf{x}'_i) - f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i}))\right] \\
&\le \sup_{\mathbf{x}^{-N+B}_{-N+1}, \mathbf{x}'^B, \mathbf{z}_j} \mathbb{E}_{\epsilon'^B} \sup_{y^j} \left[\frac{1}{B}\sum_{i=1}^{B} \epsilon'_j(f_{\mathbf{z}_j,y^j}(\mathbf{x}'_i) - f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i}))\right]
\end{aligned}$$

where (a) follows by Lemma 9.5 (in Section 9.2.1); (b) follows by definition of $\mathbf{z}_j$; (c) follows by Lemma 9.13 below and taking $B = N-M+j+1$; (d) follows by introducing B fresh $i.i.d.$ samples $\mathbf{x}'^B \sim \mu^{\otimes B}$; (e) follows by symmetries of $\mathbf{x}'^B$ and $\mathbf{x}^{-N+B}_{-N+1}$ (which are independent of $\mathbf{z}_j$) and introducing the $i.i.d.$ random variables ϵ'^B uniform over $\{-1,1\}^B$;

Lemma 9.13. Let $B = N - M + j + 1$, then

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}^0_{-N+1}} \mathbb{E}_{\mathbf{z}_j} \sup_{y^j} [\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})] - \mathbb{E}_{\tilde{\mathbf{x}}_{j+1}}[f_{\mathbf{z}_j,y^j}(\tilde{\mathbf{x}}_{j+1})]] \\
&\qquad = \mathbb{E}_{\mathbf{z}_j} \mathbb{E}_{\mathbf{x}^{-N+B}_{-N+1}} \sup_{y^j} \left[\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})] - \frac{1}{B}\sum_{i=1}^{B} f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i})\right].
\end{aligned}$$

Proof. Note that $\mathbf{z}_j = (\mathbf{x}^j, \tilde{\mathbf{x}}_{j+2}^M, \epsilon_{j+1}^M)$, where $\tilde{\mathbf{x}}_{j+1}^M$ are sampled uniformly from $\mathbf{x}_{-N+1}^0$ *without replacement*, and $\mathbf{x}^j$, ϵ_{j+1}^M are independent of $\mathbf{x}_{-N+1}^0$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_{-N+1}^0}\mathbb{E}_{\mathbf{z}_j}\sup_{y^j}[\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})] - \mathbb{E}_{\tilde{\mathbf{x}}_{j+1}}[f_{\mathbf{z}_j,y^j}(\tilde{\mathbf{x}}_{j+1})]] \\
&= \mathbb{E}_{\mathbf{x}^j,\epsilon_{j+1}^M}\mathbb{E}_{\mathbf{x}_{-N+1}^0}\mathbb{E}_{\tilde{\mathbf{x}}_{j+2}^M}\sup_{y^j}[\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})] - \mathbb{E}_{\tilde{\mathbf{x}}_{j+1}}[f_{\mathbf{z}_j,y^j}(\tilde{\mathbf{x}}_{j+1})]] \\
&\overset{(\star)}{=} \mathbb{E}_{\mathbf{x}^j,\epsilon_{j+1}^M}\mathbb{E}_{\mathbf{x}_{-N+1}^0}\mathbb{E}_I\sup_{y^j}\left[\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})] - \frac{1}{B}\sum_{i\in[N]\backslash I} f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i})\right]
\end{aligned}$$

where the key step $(\star)$ follows by noticing that the randomness of $\tilde{\mathbf{x}}_{j+2}^M$ is equivalent to selecting a *random* index set $I \subset [N]$ uniformly with size $|I| = M - j - 1$ and the index of $\tilde{\mathbf{x}}_{j+1}$ (in $\mathbf{x}_{-N+1}^0$) is then uniform over $[N]\backslash I$ [6], where the size of $[N]\backslash I$ is $B = N - M + j + 1$; Therefore,

$$\mathbb{E}_{\tilde{\mathbf{x}}_{j+1}}[f_{\mathbf{z}_j,y^j}(\tilde{\mathbf{x}}_{j+1})] = \frac{1}{B}\sum_{i\in[N]\backslash I} f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i}).$$

Note that $\mathbf{x}_{-N+1}^0$ is an *i.i.d.* sample, by *symmetries*, we can *fix* $I = \{B+1, \cdots, N\}$ (i.e., we take $\tilde{\mathbf{x}}_{j+2}^M$ being $\mathbf{x}_{-N+B+1}^0$) and therefore $\tilde{\mathbf{x}}_{j+2}^M$ can be decoupled from $\mathbf{x}_{-N+1}^{-N+B}$, leading to

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}^j,\epsilon_{j+1}^M}\mathbb{E}_{\mathbf{x}_{-N+1}^0}\mathbb{E}_I\sup_{y^j}\left[\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})] - \frac{1}{B}\sum_{i\in[N]\backslash I} f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i})\right] \\
&= \mathbb{E}_{\mathbf{x}^j,\epsilon_{j+1}^M}\mathbb{E}_{\tilde{\mathbf{x}}_{j+2}^M}\mathbb{E}_{\mathbf{x}_{-N+1}^{-N+B}}\sup_{y^j}\left[\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})] - \frac{1}{B}\sum_{i=1}^{B} f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i})\right] \\
&= \mathbb{E}_{\mathbf{z}_j}\mathbb{E}_{\mathbf{x}_{-N+1}^{-N+B}}\sup_{y^j}\left[\mathbb{E}_{\mathbf{x}\sim\mu}[f_{\mathbf{z}_j,y^j}(\mathbf{x})] - \frac{1}{B}\sum_{i=1}^{B} f_{\mathbf{z}_j,y^j}(\mathbf{x}_{-N+i})\right].
\end{aligned}$$

This completes the proof of this lemma. □

9.3 Oblivious Adversaries

Finally, we provide the regret analysis for online learning against an *oblivious* adversary. Note that, up to this point, we have assumed that

[6] By the definition of sampling *without replacement.*

the generation of y_t's is *adaptive*, meaning that the selection of y_t at each time step depends on all prior information: $\mathbf{x}^t$, y^{t-1}, and $\hat{y}^{t-1}$. For comparison, we now introduce a weaker notion of an adversary—namely, the *oblivious* adversary—which selects the y_t's based only on the current instance $\mathbf{x}_t$. Interestingly, in the case of obliviously chosen labels, the hybrid minimax regret can be substantially improved.

We follow the same online learning game as in (9.1) with the exception that the adversary fixes functions $f_1, \cdots, f_T : \mathcal{X} \to [0,1]$ before the game and sets the adversary labels $y_t = f_t(\mathbf{x}_t)$ for each time step $t \in [T]$. Formally, for any expert class $\mathcal{H}$ and prediction rule Φ, we are interested in the following *oblivious* minimax regret:

$$\begin{aligned}&\tilde{r}^{\mathsf{ob}}_T(\mathcal{H},\Phi)\\&\quad= \sup_{f_1,\cdots,f_T\in[0,1]^{\mathcal{X}}} \sup_{\mu} \mathbb{E}_{\mathbf{x}^T}\mathbb{E}_{\hat{y}^T}\left[\sum_{t=1}^{T} \ell(\hat{y}_t, f_t(\mathbf{x}_t)) - \inf_{h\in\mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), f_t(\mathbf{x}_t))\right],\end{aligned}$$

where $\mathbf{x}^T$ are sampled *i.i.d.* from μ and $\hat{y}_t \sim \Phi(\mathbf{x}^t, y^{t-1})$ for $t \in [T]$. For the clarity of presentation, we assume that $\ell(\hat{y}, y) = |\hat{y} - y|$ is the absolute loss.

We now ready to state the main result of this section:

Theorem 9.14. Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be a class of Rademacher complexity $\mathsf{Rad}_T(\mathcal{H}) = O(T^q)$ for some $q \in [\frac{1}{2}, 1]$ and ℓ be the absolute loss. Then there exists an oracle-efficient prediction rule Φ with at most $O(\sqrt{T}\log T)$ calls to the ERM oracle per round, such that

$$\tilde{r}^{\mathsf{ob}}_T(\mathcal{H},\Phi) \leq O(T^q).$$

In particular, for finite-VC class $\mathcal{H}$, we have $\tilde{r}^{\mathsf{ob}}_T(\mathcal{H},\Phi) \leq O(\sqrt{\mathsf{VC}(\mathcal{H})T})$. For a class $\mathcal{H}$ with α-fat shattering dimension $O(\alpha^{-p})$ for some $p > 0$, we have $\tilde{r}^{\mathsf{ob}}_T(\mathcal{H},\Phi) \leq \tilde{O}(T^{\max\{\frac{1}{2},\frac{p-1}{p}\}})$.

Proof. We will follow the same path as the regret analysis for the *non-oblivious* adversaries as established in Section 9.2. We first consider the scenario with side-information $\mathbf{x}^0_{-N+1}$, and define for any predictor Φ the following oblivious minimax regret with side-information:

$$\tilde{r}^{\mathsf{ob,side}}_{M,N}(\mathcal{H},\Phi) = \sup_{f_1,\cdots,f_M\in[0,1]^{\mathcal{X}}} \sup_{\mu} \mathbb{E}_{\mathbf{x}^M_{-N+1}} \mathbb{E}_{\hat{y}^M}\left[\sum_{j=1}^{M} \ell(\hat{y}_j, f_j(\mathbf{x}_j))\right.$$

$$- \inf_{h\in\mathcal{H}} \sum_{j=1}^{M} \ell(h(\mathbf{x}_j), f_j(\mathbf{x}_j)) \Bigg] ,$$

where $\mathbf{x}_{-N+1}^{M}$ are sampled *i.i.d.* (9.2) and R_j and $\tilde{R}_j$ be the same *surrogate* relaxations as in (9.4) and (9.5). We claim that:

$$\tilde{r}^{\mathsf{ob,side}}_{M,N}(\mathcal{H}, \Phi) \leq \sup_{f^M} \sup_{\mu} \mathbb{E}_{\mathbf{x}^{0}_{-N+1}} \left[\tilde{R}_0 + \sum_{j=1}^{M-1} \mathbb{E}_{\mathbf{x}^j} [\tilde{R}_j - R_j] \right] . \tag{9.16}$$

To see this, we find

$$\begin{aligned}
\tilde{r}^{\mathsf{ob,side}}_{M,N}(\mathcal{H}, \Phi) &= \sup_{f^M} \sup_{\mu} \mathbb{E}_{\mathbf{x}^{M}_{-N+1}} \mathbb{E}_{\hat{y}^M} \left[\sum_{j=1}^{M} \ell(\hat{y}_j, f_j(\mathbf{x}_j)) - \inf_{h\in\mathcal{H}} \sum_{j=1}^{M} \ell(h(\mathbf{x}_j), f_j(\mathbf{x}_j)) \right] \\
&= \sup_{f^M} \sup_{\mu} \mathbb{E}_{\mathbf{x}^{M}_{-N+1}} \mathbb{E}_{\hat{y}^M} \left[\sum_{j=1}^{M} \ell(\hat{y}_j, f_j(\mathbf{x}_j)) + R_M \right] \\
&= \sup_{f^M} \sup_{\mu} \mathbb{E}_{\mathbf{x}^{M-1}_{-N+1}} \mathbb{E}_{\hat{y}^{M-1}} \left[\sum_{j=1}^{M-1} \ell(\hat{y}_j, f_j(\mathbf{x}_j)) + \mathbb{E}_{\mathbf{x}_M} \mathbb{E}_{\hat{y}_M} [\ell(\hat{y}_M, f_M(\mathbf{x}_M)) + R_M] \right] \\
&\overset{(a)}{\leq} \sup_{f^M} \sup_{\mu} \mathbb{E}_{\mathbf{x}^{M-1}_{-N+1}} \mathbb{E}_{\hat{y}^{M-1}} \left[\sum_{j=1}^{M-1} \ell(\hat{y}_j, f_j(\mathbf{x}_j)) + \mathbb{E}_{\mathbf{x}_M} \sup_{y_M} \mathbb{E}_{\hat{y}_M} [\ell(\hat{y}_M, y_M) + R_M] \right] \\
&\overset{(b)}{\leq} \sup_{f^M} \sup_{\mu} \mathbb{E}_{\mathbf{x}^{M-1}_{-N+1}} \mathbb{E}_{\hat{y}^{M-1}} \left[\sum_{j=1}^{M-1} \ell(\hat{y}_j, f_j(\mathbf{x}_j)) + \tilde{R}_{M-1} \right] \\
&= \sup_{f^M} \sup_{\mu} \mathbb{E}_{\mathbf{x}^{M-1}_{-N+1}} \mathbb{E}_{\hat{y}^{M-1}} \left[\sum_{j=1}^{M-1} \ell(\hat{y}_j, f_j(\mathbf{x}_j)) + R_{M-1} + \tilde{R}_{M-1} - R_{M-1} \right] \\
&= \sup_{f^M} \sup_{\mu} \left(\mathbb{E}_{\mathbf{x}^{M-1}_{-N+1}} \mathbb{E}_{\hat{y}^{M-1}} \left[\sum_{j=1}^{M-1} \ell(\hat{y}_j, f_j(\mathbf{x}_j)) + R_{M-1} \right] + \mathbb{E}_{\mathbf{x}^{M-1}_{-N+1}} (\tilde{R}_{M-1} - R_{M-1}) \right) \\
&\overset{(c)}{\leq} \sup_{f^M} \sup_{\mu} \mathbb{E}_{\mathbf{x}^{0}_{-N+1}} \left[\tilde{R}_0 + \sum_{j=1}^{M-1} \mathbb{E}_{\mathbf{x}^j} [\tilde{R}_j - R_j] \right]
\end{aligned}$$

where (a) follows by that replacing $f_M(\mathbf{x}_M)$ with $\sup_{y_M}$ do not decrease the value; (b) follows by Lemma 9.3; (c) follows by repeating the same argument for another $M - 1$ steps.

Now, the key observation is that $\mathbb{E}_{\mathbf{x}^j_{-N+1}}[\tilde{R}_j - R_j] = 0$ for all $j \in [M-1]$ whenever $N \geq M-1$. This follows by the same argument as in the proof of Lemma 9.6 by noticing that the $\sup_{y^j}$ is outside the expectation $\mathbb{E}_{\epsilon'^B}$ for oblivious adversaries. Moreover, this argument holds for all $B = N - M + j + 1 \geq 1$, i.e., $N \geq M - j$ (since by our assumption $N \geq M-1$ and $j \geq 1$). Therefore, we have

$$\tilde{r}^{\mathsf{ob,side}}_{M,N}(\mathcal{H}, \Phi) \leq \mathbb{E}_{\mathbf{x}^0_{-N+1}}[\tilde{R}_0] \leq \mathsf{Rad}_M(\mathcal{H}) \leq O(M^q),$$

whenever $N \geq M-1$. By the epoch approach as in Section 9.2.2 and taking the epoch length $M(n) = 2^n$ (which ensures $S(n) \geq M(n) - 1$) we conclude

$$\tilde{r}^{\mathsf{ob}}_T(\mathcal{H}, \Psi) \leq \sum_{n=1}^{\lceil \log T \rceil} 2^{nq} \leq O(T^q),$$

where Ψ is the epoch predictor derived from Φ as (9.12). The theorem now follows by Lemma 9.2 and noticing that the computational error only contributes $O(\sqrt{T})$ to the regret. □

Remark 9.4. Theorem 9.14 demonstrates that the *oblivious* minimax regret with *unknown i.i.d.* feature generation process is equivalent to the regret achievable with *known* feature generation distribution and non-oblivious adversaries [63, Thm 7], which also matches the information-theoretical lower bound (upto poly-logarithmic factors).

9.4 Bibliographical Notes

The relaxation-based approach was first introduced by [26], providing a generic method for constructing sequential prediction algorithms (albeit potentially inefficient) for a wide range of online learning scenarios. Rakhlin *et al.* [26] demonstrated that an *oracle-efficient* online learning algorithm is feasible via the so-called *random play-out* approach, provided one can access a sampling oracle for *future* features. This includes applications such as transductive online learning [76], [77] and settings with *known* i.i.d. feature generation distributions [78]–[80]. A more sophisticated scenario—the *smooth adversarial* setting—was investigated by [25], [63], [68], [81]. In this setting, the future sampling distribution is not directly accessible but can be stochastically controlled via a *coupling* argument introduced by [68]. However, this approach still requires access to a sampling oracle for the underlying reference measure.

The results on oracle-efficient algorithm with unknown feature distributions presented in this chapter are based on the work of [82].

10

Online Learning with Noisy Labels

This chapter studies online learning from noisy labels, where examples arrive sequentially with adversarially chosen features and stochastically corrupted labels. Unlike classical agnostic online learning—where regret is evaluated on *observable* labels and both features and labels may be adversarial—our focus is on a more structured setting: (i) we assume that the noisy labels are derived from a semi-*stochastic* mechanism rather than from purely adversarial selections; (ii) our predictions are evaluated on the (unobservable) *true* labels, not *noisy* observations.

It is instructive to start with the following example from [24]:

Example 10.1. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a finite hypothesis class. Consider the following online learning game between Nature/Adversary and Learner over a time horizon T. Nature fixes a ground truth $h \in \mathcal{H}$ to start the game. At each time step t, Nature adversarially selects feature $\mathbf{x}_t \in \mathcal{X}$ and reveals it to the learner. Learner makes a prediction $\hat{y}_t$ based on prior features $\mathbf{x}^t = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$ and *noisy* labels $\tilde{y}^{t-1} = \{\tilde{y}_1, \cdots, \tilde{y}_{t-1}\}$. Nature then selects an (unknown) noise parameter $\eta_t \in [0, \eta]$ for some given η (known to learner), and generates [1]

$$\tilde{y}_t = \mathsf{Bernoulli}(\eta_t) \oplus y_t,$$

[1] This is typically referred to as Massart's noise.

where $\oplus$ denotes binary addition and $y_t = h(\mathbf{x}_t)$ is the *true* label. It was demonstrated by Ben-David *et al.* [24, Thm 15] that there exist predictors $\hat{y}^T$ such that

$$\sup_{h\in\mathcal{H},\mathbf{x}^T\in\mathcal{X}^T} \mathbb{E}\left[\sum_{t=1}^{T} 1\{\hat{y}_t \neq h(\mathbf{x}_t)\}\right] \leq \frac{\log|\mathcal{H}|}{1-2\sqrt{\eta(1-\eta)}}. \tag{10.1}$$

Note that the risk bound in (10.1) is surprising: although the cumulative noise grows linearly with ηT, the risk remains *independent* of the time horizon T, even when evaluated on the *unseen* true labels. Despite its foundational nature, understanding this phenomenon beyond simple Massart's noise remains largely unexplored in the literature.

We present in this chapter a theoretical framework that systematically addresses this gap, offering a more *principled approach* to understanding the intrinsic complexity of the problem that determines the risk under various noise mechanisms.

10.1 Problem Formulation and Preliminaries

Let $\mathcal{X}$ be a set of features (or instances), $\mathcal{Y}$ be a set of labels, and $\tilde{\mathcal{Y}}$ be a set of *noisy observations*. We assume throughout this chapter that $|\mathcal{Y}| = N$ and $|\tilde{\mathcal{Y}}| = M$ for some integers $N, M \geq 2$. We denote

$$\mathcal{D}(\tilde{\mathcal{Y}}) = \left\{p = (p[1], \ldots, p[M]) \in [0,1]^M : \sum_{m=1}^{M} p[m] = 1\right\}$$

as the set of all *probability distributions* over $\tilde{\mathcal{Y}}$.

A *noise kernel* is defined as a map

$$\mathcal{K} : \mathcal{X} \times \mathcal{Y} \to 2^{\mathcal{D}(\tilde{\mathcal{Y}})},$$

where $2^{\mathcal{D}(\tilde{\mathcal{Y}})}$ is the set of all *subsets* of $\mathcal{D}(\tilde{\mathcal{Y}})$, i.e., the kernel $\mathcal{K}$ maps each $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ to a *subset of distributions* $\mathcal{K}(\mathbf{x}, y) \subset \mathcal{D}(\tilde{\mathcal{Y}})$.

Note that the noise kernel provides a compact way of modeling the *noisy label distribution* directly without explicitly referring to the *noise*. This is more convenient for our discussion, as ultimately the statistical information is solely determined by the noisy label distributions.

Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a class of *hypotheses* and $\mathcal{K}$ be a noise kernel as defined above. Denote $\mathcal{Q}_y^{\mathbf{x}} := \mathcal{K}(\mathbf{x}, y)$ for notational convenience. We consider the following *robust online classification* protocol:

Protocol 10.1 Adversarial Online Learning with Noisy Labels

```
Initialization: Nature fixes a ground truth $h \in \mathcal{H}$
for $t = 1$ to $T$ do
    Nature selects instance $\mathbf{x}_t \in \mathcal{X}$ adversarially
    Learner predicts $\hat{y}_t \in \mathcal{Y}$ based on history $(\mathbf{x}^t, \tilde{y}^{t-1})$
    Adversary selects $\tilde{p}_t \in \mathcal{Q}_{h(\mathbf{x}_t)}^{\mathbf{x}_t}$ and samples noisy label $\tilde{y}_t \sim \tilde{p}_t$
end for
```

The goal of *learner* is to minimize the *cumulative risk*

$$\sum_{t=1}^{T} 1\{h(\mathbf{x}_t) \neq \hat{y}_t\}.$$

Note that the cumulative error is a *random variable* that depends on all the randomness associated with the game. To remove the dependency on such randomness and to assess the fundamental limits of the prediction quality, we consider the following two measures [2]:

Definition 10.1. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a set of hypotheses and $\mathcal{K} : \mathcal{X} \times \mathcal{Y} \to 2^{\mathcal{D}(\tilde{\mathcal{Y}})}$ be a noise kernel. We denote by Φ the (possibly randomized) strategies of the *learner*. The *expected minimax risk* is defined as:

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}) = \inf_{\Phi} \sup_{h \in \mathcal{H}} \mathbb{Q}_{\mathcal{K}}^T \mathbb{E}_{\hat{y}^T} \left[\sum_{t=1}^{T} 1\{h(\mathbf{x}_t) \neq \hat{y}_t\} \right], \tag{10.2}$$

where $\hat{y}_t \sim \Phi(\mathbf{x}^t, \tilde{y}^{t-1})$ and $\mathbb{Q}_{\mathcal{K}}^T$ denotes for operator

$$\mathbb{Q}_{\mathcal{K}}^T \equiv \sup_{\mathbf{x}_1 \in \mathcal{X}} \sup_{\tilde{p}_1 \in \mathcal{Q}_{h(\mathbf{x}_1)}^{\mathbf{x}_1}} \mathbb{E}_{\tilde{y}_1 \sim \tilde{p}_1} \cdots \sup_{\mathbf{x}_T \in \mathcal{X}} \sup_{\tilde{p}_T \in \mathcal{Q}_{h(\mathbf{x}_T)}^{\mathbf{x}_T}} \mathbb{E}_{\tilde{y}_T \sim \tilde{p}_T}.$$

[2] We assume here the selection of $\tilde{p}^T$ and $\mathbf{x}^T$ are oblivious to the learner's action for simplicity. This is equivalent to the adaptive case if the learner's internal randomness are independent among different time steps by a standard argument from Cesa-Bianchi *et al.* [9, Lemma 4.1].

By *skolemization* (Lemma 2.2), we have operator identity:

$$\mathbb{Q}_{\mathcal{K}}^T \equiv \sup_{\psi^T} \sup_{\tilde{p}^T} \mathbb{E}_{\tilde{y}^T \sim \tilde{p}^T},$$

where $\psi^T = \{\psi_1, \cdots, \psi_T\}$ runs over all functions $\psi_t : \tilde{\mathcal{Y}}^{t-1} \to \mathcal{X}$ for $t \in [T]$ and $\tilde{p}^T$ runs over all (joint) distributions over $\tilde{\mathcal{Y}}^T$ subject to the constraints that for any $t \in [T]$ and $\tilde{y}^{t-1}$ the *conditional* marginal $\tilde{p}_t$ of $\tilde{p}^T$ at $\tilde{y}_t$ conditioning on $\tilde{y}^{t-1}$ satisfies $\tilde{p}_t \in \mathcal{Q}_{h(\mathbf{x}_t)}^{\mathbf{x}_t}$ for $\mathbf{x}_t = \psi_t(\tilde{y}^{t-1})$. This leads to our next definition of the *high probability* minimax risk:

Definition 10.2. Let $\mathcal{H}$, $\mathcal{K}$ and Φ be as in Definition 10.1. For any confidence parameter $\delta > 0$, the *high probability minimax risk* at confidence δ is defined as the minimum number $B^\delta(\mathcal{H}, \mathcal{K}) \geq 0$ such that there exists a predictor Φ satisfying:

$$\sup_{h \in \mathcal{H}, \psi^T, \tilde{p}^T} \Pr\left[\sum_{t=1}^T 1\{h(\mathbf{x}_t) \neq \hat{y}_t\} \geq B^\delta(\mathcal{H}, \mathcal{K})\right] \leq \delta, \tag{10.3}$$

where the selection of ψ^T and $\tilde{p}^T$ are as in the discussion above with $\mathbf{x}_t = \psi_t(\tilde{y}^{t-1})$ and the probability is over both $\tilde{y}^T \sim \tilde{p}^T$ and $\hat{y}^T$ for $\hat{y}_t \sim \Phi(\mathbf{x}^t, \tilde{y}^{t-1})$.

Note that the kernel map $\mathcal{K}$ is generally *known* to the learner when constructing the predictor Φ. However, the induced kernel sets $\mathcal{Q}_{h(\mathbf{x}_t)}^{\mathbf{x}_t}$ are not, since they depend on the *unknown* ground truth classifier h and *adversarially* generated features $\mathbf{x}^T$. In certain cases, such as Theorem 10.4, the kernel map $\mathcal{K}$ is also *not* required to be known.

We assume, w.l.o.g., that $\mathcal{Q}_y^{\mathbf{x}}$s are *convex* and *closed* sets for all $(\mathbf{x}, y)$, since the adversary can select an arbitrary distribution from $\mathcal{Q}_y^{\mathbf{x}}$s at each time t, including randomized strategies that effectively sample from a mixture (i.e., convex combination) of distributions in $\mathcal{Q}_y^{\mathbf{x}}$s.

Clearly, one must introduce some constraints on the kernel $\mathcal{K}$ in order to obtain meaningful results. To do so, we introduce the following *well-separateness* condition:

Definition 10.3 (Well-Separated Kernel)**.** Let $L : \mathcal{D}(\tilde{\mathcal{Y}}) \times \mathcal{D}(\tilde{\mathcal{Y}}) \to \mathbb{R}^{\geq 0}$ be a divergence, we say a kernel $\mathcal{K}$ is *well-separated* w.r.t. L at scale $\gamma > 0$, if $\forall \mathbf{x} \in \mathcal{X}$, $\forall y, y' \in \mathcal{Y}$ with $y \neq y'$,

$$L(\mathcal{Q}_y^{\mathbf{x}}, \mathcal{Q}_{y'}^{\mathbf{x}}) \overset{\mathsf{def}}{=} \inf_{p \in \mathcal{Q}_y^{\mathbf{x}}, q \in \mathcal{Q}_{y'}^{\mathbf{x}}} L(p, q) \geq \gamma.$$

Example 10.2. Let $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$ be the label and noisy observation sets. We can specify for any $y \in \mathcal{Y}$ a canonical distribution $p_y \in \mathcal{D}(\tilde{\mathcal{Y}})$. A natural kernel would be to define:

$$\mathcal{Q}_y^{\mathbf{x}} = \{p \in \mathcal{D}(\tilde{\mathcal{Y}}) : ||p - p_y||_{\mathsf{TV}} \leq \epsilon\}.$$

In this case, the kernel is well-separated with the gap γ under total variation if:

$$\inf_{y \neq y' \in \mathcal{Y}} ||p_y - p_{y'}||_{\mathsf{TV}} \geq \gamma + 2\epsilon.$$

10.2 The Binary Label Case

We initiate our discussion with a simple case, where we assume the label space $\mathcal{Y} = \{0, 1\}$ is binary-valued. This will provide us with an intuitive understanding of how the stochastic nature of noisy labels impacts the risk bounds. We state our first main result:

Theorem 10.1. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any *finite* binary valued class, $\mathcal{K}$ be any noise kernel that is well-separated at scale γ_{L} w.r.t. L^2 divergence. Then, the *expected* minimax risk, defined in Definition 10.1, is upper bounded by:

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}) \leq \frac{16 \log |\mathcal{H}|}{\gamma_{\mathrm{L}}}.$$

10.2.1 Proof of Theorem 10.1

We begin with the following simple geometry fact that is crucial for our proof.

Lemma 10.2. Let $\mathcal{Q} \subset \mathcal{D}(\tilde{\mathcal{Y}})$ be a convex and closed set, p be a point outside of $\mathcal{Q}$ with $\gamma \overset{\mathsf{def}}{=} \inf_{q \in \mathcal{Q}} L^2(p, q)$. Denote by $q^* \in \mathcal{Q}$ the (unique) point that attains $L^2(p, q^*) = \gamma$. Then for any $q \in \mathcal{Q}$, we have $L^2(q, p) - L^2(q, q^*) \geq L^2(p, q^*) = \gamma$.

Proof. By the *hyperplane separation theorem*, the hyperplane perpendicular to line segment $p - q^*$ at q^* separates $\mathcal{Q}$ and p. Therefore, the degree θ of angle formed by $p - q^* - q$ is greater than $\pi/2$. By the law of cosines, $L^2(q, p) \geq L^2(q, q^*) + L^2(q^*, p) = L^2(q, q^*) + \gamma$. □

Our key idea of proving Theorem 10.1 is to reduce the robust (noisy) online classification problem to a suitable conditional distribution estimation problem, as discussed next.

Online conditional distribution estimation. Let $\mathcal{F} \subset \mathcal{D}(\tilde{\mathcal{Y}})^{\mathcal{X}}$ be a class of functions mapping $\mathcal{X}$ to *distributions* in $\mathcal{D}(\tilde{\mathcal{Y}})$. *Online Conditional Distribution Estimation* (OCDE) is a game between *Nature* and an *estimator* that follows the following protocol: (1) at each time step t, Nature selects some $\mathbf{x}_t \in \mathcal{X}$ and reveals it to the estimator; (2) the estimator then makes an estimation $\hat{p}_t \in \mathcal{D}(\tilde{\mathcal{Y}})$, based on $\mathbf{x}^t, \tilde{y}^{t-1}$; (3) Nature then selects some $\tilde{p}_t \in \mathcal{D}(\tilde{\mathcal{Y}})$, samples $\tilde{y}_t \sim \tilde{p}_t$ and reveals $\tilde{y}_t$ to the estimator. The goal is to find a (deterministic) estimator Φ that minimizes the *regret*:

$$\mathsf{Reg}_T(\mathcal{F}, \Phi) = \sup_{f \in \mathcal{F}} \mathbb{Q}^T \left[\sum_{t=1}^{T} L(\tilde{p}_t, \hat{p}_t) - L(\tilde{p}_t, f(\mathbf{x}_t)) \right], \tag{10.4}$$

where $\hat{p}_t = \Phi(\mathbf{x}^t, \tilde{y}^{t-1})$, $\mathbb{Q}^T$ is the operator specified in Definition 10.1 by setting $\mathcal{Q}_y^{\mathbf{x}} = \mathcal{D}(\tilde{\mathcal{Y}})$ for all $\mathbf{x}, y$, and L is any divergence. We emphasize that distributions $\tilde{p}^T$ are *not* necessarily realizable by f and are selected completely arbitrarily. This is the key that allows us to deal with *unknown* noisy label distributions.

We now establish the following key technical lemma:

Lemma 10.3. Let $\mathcal{F}$ be any distribution-valued finite class and L be a Bregman divergence such that the induced loss $\ell(p, \tilde{y}) \overset{\mathsf{def}}{=} L(e_{\tilde{y}}, p)$ is α-Exp-concave. Then, there exists an estimator Φ, such that

$$\mathsf{Reg}_T(\mathcal{F}, \Phi) \le \frac{\log |\mathcal{F}|}{\alpha}.$$

Moreover, estimation $\hat{p}_t$ is a convex combination of $\{f(\mathbf{x}_t) : f \in \mathcal{F}\}$.

Proof. Let Φ be the EWA algorithm as in Algorithm 3.3 with input $\mathcal{F}$, $\eta := \alpha$ and loss $\ell(p, \tilde{y}) \overset{\mathsf{def}}{=} L(e_{\tilde{y}}, p)$. Let $\tilde{y}^T$ be any realization of the labels and e_t be the standard base of $\mathbb{R}^M$ with value 1 at position $\tilde{y}_t$ and zeros otherwise. By α-Exp-concavity of loss ℓ and the regret bound from Theorem 3.5 (view $\mathbf{x}_t = \psi_t(\tilde{y}^{t-1})$ and $\hat{p}_t := \hat{y}_t$), we have:

$$\sup_{f \in \mathcal{F}, \psi^T, \tilde{y}^T \in \tilde{\mathcal{Y}}^T} \sum_{t=1}^{T} L(e_t, \hat{p}_t) - L(e_t, f(\psi_t(\tilde{y}^{t-1}))) \le \frac{\log |\mathcal{F}|}{\alpha}, \tag{10.5}$$

where $\psi^T = \{\psi_1, \cdots, \psi_T\}$ runs over all functions $\psi_t : \tilde{\mathcal{Y}}^{t-1} \to \mathcal{X}$ for $t \in [T]$. Note that this bound holds *point-wise* w.r.t. any individual $\psi^T, \tilde{y}^T$.

Fix any ψ^T and distribution $\tilde{p}^T$ over $\tilde{\mathcal{Y}}^T$. We denote $\mathbb{E}_t$ as the conditional expectation on $\tilde{y}_t$ over the randomness of $\tilde{y}^T \sim \tilde{p}^T$ conditioning on $\tilde{y}^{t-1}$ and denote $\tilde{p}_t$ as the *conditional* marginal. By Proposition 2.1, we have for all $t \in [T]$ that:

$$\mathbb{E}_t \left[L(e_t, \hat{p}_t) - L(e_t, f(\psi_t(\tilde{y}^{t-1})))\right] = L(\tilde{p}_t, \hat{p}_t) - L(\tilde{p}_t, f(\psi_t(\tilde{y}^{t-1}))),$$

since $\mathbb{E}_t[e_t] = \tilde{p}_t$ for $\tilde{y}_t \sim \tilde{p}_t$, $\hat{p}_t$ depending only on $\tilde{y}^{t-1}$ and L is a Bregman divergence. We now take $\mathbb{E}_{\tilde{y}^T}$ on both sides of (10.5). By $\sup \mathbb{E} \leq \mathbb{E} \sup$ and the law of total probability (i.e., $\mathbb{E}_{\tilde{y}^T}[X_1 + \cdots + X_T] = \mathbb{E}_{\tilde{y}^T}[\mathbb{E}_1[X_1] + \cdots + \mathbb{E}_T[X_T]]$ for any random variables X^T), we have:

$$\sup_{f \in \mathcal{F}} \sup_{\psi^T, \tilde{p}^T} \mathbb{E}_{\tilde{y}^T \sim \tilde{p}^T} \left[\sum_{t=1}^{T} L(\tilde{p}_t, \hat{p}_t) - L(\tilde{p}_t, f(\psi_t(\tilde{y}^{t-1})))\right] \leq \frac{\log |\mathcal{F}|}{\alpha},$$

where $\tilde{p}^T$ runs over all distributions over $\tilde{\mathcal{Y}}^T$ and ψ^T runs over all functions $\psi_t : \tilde{\mathcal{Y}}^{t-1} \to \mathcal{X}$. The lemma then follows by the operator equivalence $\mathbb{Q}^T \equiv \sup_{\psi^T, \tilde{p}^T} \mathbb{E}_{\tilde{y}^T}$ and taking the kernel $\mathcal{Q}_y^{\mathbf{x}} := \mathcal{D}(\tilde{\mathcal{Y}})$ (see the discussion following Definition 10.1). The last part follows by the fact that the EWA algorithm automatically ensures $\hat{p}_t$ is a convex combination of $\{f(\mathbf{x}_t) : f \in \mathcal{F}\}$ for all $t \in [T]$. □

Proof of Theorem 10.1. We define the following distribution valued function class $\mathcal{F}$ using hypothesis class $\mathcal{H}$ and noise kernel $\mathcal{K}$. For any $\mathbf{x} \in \mathcal{X}$, we denote by $\mathcal{Q}_0^{\mathbf{x}}$ and $\mathcal{Q}_1^{\mathbf{x}}$ the sets of noisy label distributions corresponding to labels 0 and 1, respectively. Since the kernel $\mathcal{K}$ is well-separated at scale γ_{L} under L^2 divergence, we have, by the *hyperplane separation theorem*, that there must exist $q_0^{\mathbf{x}} \in \mathcal{Q}_0^{\mathbf{x}}$ and $q_1^{\mathbf{x}} \in \mathcal{Q}_1^{\mathbf{x}}$ such that $L^2(q_0^{\mathbf{x}}, q_1^{\mathbf{x}}) = L^2(\mathcal{Q}_0^{\mathbf{x}}, \mathcal{Q}_1^{\mathbf{x}}) \geq \gamma_{\text{L}}$. We now define, for any $h \in \mathcal{H}$ the function f_h such that $\forall \mathbf{x} \in \mathcal{X},\ f_h(\mathbf{x}) = q_{h(\mathbf{x})}^{\mathbf{x}}$. Let $\mathcal{F} = \{f_h : h \in \mathcal{H}\}$ and Φ be the estimator in Online Conditional Density Estimation (OCDE) game from Lemma 10.3 with class $\mathcal{F}$ and L^2 divergence (using $\mathbf{x}^T, \tilde{y}^T$

from the *original* noisy classification game). Our *classification* predictor is as follows:

$$\hat{y}_t = \arg\min_y \{L^2(q_y^{\mathbf{x}_t}, \hat{p}_t) : y \in \{0,1\}\}. \tag{10.6}$$

That is, we predict the label y so that $q_y^{\mathbf{x}_t}$ is closer to $\hat{p}_t$ under L^2 divergence, where $\hat{p}_t = \Phi(\mathbf{x}^t, \tilde{y}^{t-1})$.

Let $h^* \in \mathcal{H}$ be the underlying true classification function. We have by Lemma 10.3 and 1/4-Exp-concavity of L^2 divergence that [3]

$$\mathbb{Q}_{\mathcal{K}}^T \left[\sum_{t=1}^T L^2(\tilde{p}_t, \hat{p}_t) - L^2(\tilde{p}_t, f_{h^*}(\mathbf{x}_t)) \right] \leq 4 \log |\mathcal{F}|, \tag{10.7}$$

where $\mathbb{Q}_{\mathcal{K}}^T$ is the operator in Definition 10.1.

For any time step t, we denote by $y_t = h^*(\mathbf{x}_t)$ the true label. Since $q_y^{\mathbf{x}_t} \in \mathcal{Q}_y^{\mathbf{x}_t}$ are the elements satisfying $L^2(q_0^{\mathbf{x}_t}, q_1^{\mathbf{x}_t}) = L^2(\mathcal{Q}_0^{\mathbf{x}_1}, \mathcal{Q}_1^{\mathbf{x}_t}) \geq \gamma_{\text{L}}$ and $\hat{p}_t$ is a *convex* combination of $q_0^{\mathbf{x}_t}$ and $q_1^{\mathbf{x}_t}$ (Lemma 10.3), we have $q_{y_t}^{\mathbf{x}_t}$ is the closest element in $\mathcal{Q}_{y_t}^{\mathbf{x}_t}$ to $\hat{p}_t$ under L^2 divergence. Note that, we also have $\tilde{p}_t \in \mathcal{Q}_{y_t}^{\mathbf{x}_t}$. Invoking Lemma 10.2, we find

$$L^2(\tilde{p}_t, \hat{p}_t) - L^2(\tilde{p}_t, q_{y_t}^{\mathbf{x}_t}) \geq L^2(\hat{p}_t, q_{y_t}^{\mathbf{x}_t}). \tag{10.8}$$

Denote $a_t = L^2(\tilde{p}_t, \hat{p}_t) - L^2(\tilde{p}_t, f_{h^*}(\mathbf{x}_t))$. We have, by (10.8) and $f_{h^*}(\mathbf{x}_t) = q_{y_t}^{\mathbf{x}_t}$ that $a_t \geq L^2(\hat{p}_t, f_{h^*}(\mathbf{x}_t))$. Therefore:

1. For all $t \in [T]$, $a_t \geq 0$, since $\forall p, q,\ L^2(p,q) \geq 0$;

2. If $\hat{y}_t \neq y_t$, then $a_t \geq \gamma_{\text{L}}/4$. This is because the event $\{\hat{y}_t \neq y_t\}$ implies that $L^2(\hat{p}_t, q_{y_t}^{\mathbf{x}_t}) \geq L^2(\hat{p}_t, q_{1-y_t}^{\mathbf{x}_t})$. Hence, $L^2(\hat{p}_t, f_{h^*}(\mathbf{x}_t)) = L^2(\hat{p}_t, q_{y_t}^{\mathbf{x}_t}) \geq \gamma_{\text{L}}/4$. Here, we used the following geometric fact:

$$\begin{aligned} 2\sqrt{L^2(\hat{p}_t, q_{y_t}^{\mathbf{x}_t})} &\geq \sqrt{L^2(\hat{p}_t, q_{y_t}^{\mathbf{x}_t})} + \sqrt{L^2(\hat{p}_t, q_{1-y_t}^{\mathbf{x}_t})} \\ &= \sqrt{L^2(q_{y_t}^{\mathbf{x}_t}, q_{1-y_t}^{\mathbf{x}_t})} \geq \sqrt{\gamma_{\text{L}}}. \end{aligned}$$

This implies that $\forall t \in [T],\ a_t \geq \frac{\gamma_{\text{L}}}{4} 1\{\hat{y}_t \neq y_t\}$, therefore:

$$\sum_{t=1}^T 1\{\hat{y}_t \neq y_t\} \leq \frac{4}{\gamma_{\text{L}}} \sum_{t=1}^T L^2(\tilde{p}_t, \hat{p}_t) - L^2(\tilde{p}_t, f_{h^*}(\mathbf{x}_t)).$$

[3] Since $\mathbb{Q}_{\mathcal{K}}^T[F(\psi^T, \tilde{y}^T)] \leq \mathbb{Q}^T[F(\psi^T, \tilde{y}^T)]$ for any kernel $\mathcal{K}$ and function F, where $\mathbb{Q}^T$ is the *unconstrained* operator in (10.4).

The expected minimax risk now follows from (10.7) since $|\mathcal{F}| \leq |\mathcal{H}|$. $\square$

Note that, for the specific setting in Example 10.1, our result yields risk of the same order up to a constant factor, since $1 - 2\sqrt{\eta(1-\eta)} = \Theta((1-2\eta)^2)$ for $\eta \in [0, \frac{1}{2})$.

Remark 10.1. Note that the selection of L^2 divergence plays a central rule in the proof of Theorem 10.1 thanks to Lemma 10.2. A naive extension to KL-divergence does not work, mainly due to the fact that if q is a projection of point p onto a convex set under KL-divergence, it does not necessarily imply that q is the projection of any point along the line segment of p and q. Therefore, our central argument in the proof of Theorem 10.1 that relates $1\{\hat{y}_t \neq y_t\}$ and $L(\tilde{p}_t, \hat{p}_t) - L(\tilde{p}_t, f_{h^*}(\mathbf{x}_t))$ will not go through. This can be remedied for certain special noise kernels, as discussed in Section 10.5.

10.3 Reduction to Pairwise Comparison: a Generic Approach

As we showed in Section 10.2, minimax risk can be upper bounded by $\frac{16 \log |\mathcal{H}|}{\gamma_{\mathrm{L}}}$ if the kernel is uniformly separated by an L^2 gap γ_{L}. However, two issues remain: (i) the proof technique is not directly generalizable to the multi-class label case. For instance, in the binary case we define a class $\mathcal{F}$ with values $q_0^{\mathbf{x}}, q_1^{\mathbf{x}}$ that satisfy $L^2(q_0^{\mathbf{x}}, q_1^{\mathbf{x}}) = L^2(\mathcal{Q}_0^{\mathbf{x}}, \mathcal{Q}_1^{\mathbf{x}})$. However, in the multi-class case, this selection is less obvious since for any $y \in \mathcal{Y}$, the closest points in $\mathcal{Q}_y^{\mathbf{x}}$ to different sets $\mathcal{Q}_{y'}^{\mathbf{x}}$ are *different*. There is no canonical way of assigning the value $f_h(\mathbf{x})$; (ii) it is unclear whether L^2 gap is the right information-theoretical measure for characterizing minimax risk, compared to, for instance, the more natural f-divergences. This section presents a general approach for addressing these issues via a novel reduction to *pairwise comparison* of two-hypotheses.

We first introduce a few technical concepts before presenting our main results. Recall that our robust online classification problem is completely determined by the pair $(\mathcal{H}, \mathcal{K})$ of hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ and noise kernel $\mathcal{K}$.

Definition 10.4. A robust online classification problem $(\mathcal{H}, \mathcal{K})$ is said to be *pairwise testable* with confidence $\delta > 0$ and error bound $C(\delta) \geq 0$, if for any pair $h_i, h_j \in \mathcal{H}$, the sub-problem $(\{h_i, h_j\}, \mathcal{K})$ admits a *high probability minimax risk* $B^\delta(\{h_i, h_j\}, \mathcal{K}) \leq C(\delta)$ at confidence δ (see Definition 10.2).

Clearly, if $(\mathcal{H}, \mathcal{K})$ admits a high probability minimax risk $B^\delta(\mathcal{H}, \mathcal{K})$, then it is also pairwise testable with the same risk by taking $C(\delta) = B^\delta(\mathcal{H}, \mathcal{K})$. Perhaps surprisingly, we will show in this section that the *converse* holds as well up to a logarithmic factor.

Assume for now that the pair $(\mathcal{H}, \mathcal{K})$ is *pairwise testable* and class $\mathcal{H} = \{h_1, \cdots, h_K\}$ is finite of size K. Let $\Phi_{i,j}$ be the predictor for the sub-problem $(\{h_i, h_j\}, \mathcal{K})$ with error bound $C(\delta/(2K))$ and confidence $\delta/(2K) > 0$. Let $\mathbf{x}^T, \tilde{y}^T$ be any realization of problem $(\mathcal{H}, \mathcal{K})$. We define, for any $h_i \in \mathcal{H}$ and $t \in [T]$, a *surrogate loss* vector:

$$\forall j \in [K],\ \mathbf{v}_t^i[j] = 1\{\Phi_{i,j}(\mathbf{x}^t, \tilde{y}^{t-1}) \neq h_i(\mathbf{x}_t) \text{ and } h_i(\mathbf{x}_t) \neq h_j(\mathbf{x}_t)\}, \tag{10.9}$$

That is, the loss $\mathbf{v}_t^i[j] = 1$ if and only if $h_i(\mathbf{x}_t) \neq h_j(\mathbf{x}_t)$ *and* the predictor $\Phi_{i,j}(\mathbf{x}^t, \tilde{y}^{t-1})$ differs from $h_i(\mathbf{x}_t)$. Given access to predictors $\Phi_{i,j}$s, our prediction rule for $(\mathcal{H}, \mathcal{K})$ is then presented in Algorithm 10.2.

At a high level, Algorithm 10.2 tries to identify the ground truth classifier h_{k^*} using the testing results of $\Phi_{i,j}$s. Note that pairwise testability implies, w.h.p., the errors made by tester Φ_{k,k^*} on h_{k^*} is upper bounded by C for all $k \in [K]$ simultaneously. However, for any other pair $i, j \neq k^*$, the tester $\Phi_{i,j}$ does not provide any guarantees, since the samples used to test h_i, h_j originate from h_{k^*} and is not *realizable* for $\Phi_{i,j}$. The key technical challenge is to extract the testing results for Φ_{k,k^*} from the other irrelevant tests (i.e., $\Phi_{i,j}$ with $k^* \not\in \{i, j\}$), even when the k^* is *unknown*. This is resolved by our definition of l_t^i in Algorithm 10.2, which computes for each i the *maximum* testing loss over all of its competitors. This ensures that, for ground truth k^*, loss $l_t^{k^*} \leq C$. While for any other $i \neq k^*$, we have $l_t^i \geq \sum_{r=1}^t \mathbf{v}_r^i[k^*] \geq \sum_{r=1}^t 1\{h_i(\mathbf{x}_r) \neq h_{k^*}(\mathbf{x}_r)\} - C$. Therefore, any hypothesis h_i for which $l_t^i > C$ cannot be the ground truth. Algorithm 10.2 then maintains an index set S^t that eliminates all h_i for which $l_t^i > C$, and makes prediction $\hat{y}_t = h_{\hat{k}_t}(\mathbf{x}_t)$ with $\hat{k}_t$ sampling *uniformly* from S^t. In particular, Algorithm 10.2 enjoys the following risk bound:

Algorithm 10.2 Predictor via Pairwise Hypothesis Testing

Input: Class $\mathcal{H} = \{h_1, \ldots, h_K\}$, testers $\Phi_{i,j}$ for $i, j \in [K]$, and error bound C

```
Initialize $S^1 \leftarrow \{1, \ldots, K\}$.
for $t \leftarrow 1$ to $T$ do
    Receive $\mathbf{x}_t$.
    Sample index $\hat{k}_t$ from $S^t$ uniformly and make prediction:
        $$\hat{y}_t \leftarrow h_{\hat{k}_t}(\mathbf{x}_t).$$
    Receive noisy label $\tilde{y}_t$.
    Set $S^{t+1} \leftarrow \emptyset$.
    for $i \in S^t$ do
        Compute
            $$l_t^i = \max_{j \in [K]} \sum_{r=1}^{t} \mathbf{v}_r^i[j],$$
        where $\mathbf{v}_t^i[j]$ is computed via $\Phi_{i,j}$ as in Equation (10.9).
        if $l_t^i \leq C$ then
            Update $S^{t+1} \leftarrow S^{t+1} \cup \{i\}$.
        end if
    end for
end for
```

Theorem 10.4. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be any finite hypothesis class of size K and $\mathcal{K}$ be any noisy kernel. If the pair $(\mathcal{H}, \mathcal{K})$ is pairwise testable with error bound $C(\delta)$ as in Definition 10.4, then for any $\delta > 0$, the predictor in Algorithm 10.2 with $C = C(\delta/(2K))$ achieves the *high probability* minimax risk (Definition 10.2) upper bounded by:

$$B^{\delta}(\mathcal{H}, \mathcal{K}) \leq 2(1 + 2C(\delta/(2K)) \log K) + \log(2/\delta). \tag{10.10}$$

Proof. Let $h_{k^*} \in \mathcal{H}$ be the underlying true classification function and ψ^T be any fixed functions realizing the features $\mathbf{x}_t = \psi_t(\tilde{y}^{t-1})$ (see Definition 10.2). We take $C = C(\delta/2K)$ in Algorithm 10.2. By definition of *pairwise testability* and union bound, we have w.p. $\geq 1 - \delta/2$ over the randomness of $\tilde{y}^T$ and the internal randomness of Φ_{k,k^*}s that for all $k \in [K]$,

$$\sum_{t=1}^{T} 1\{h_{k^*}(\mathbf{x}_t) \neq \Phi_{k,k^*}(\mathbf{x}^t, \tilde{y}^{t-1})\} \leq C(\delta/(2K)). \tag{10.11}$$

Note that for any other $\{i, j\} \not\ni k^*$, equation (10.11) may not hold for predictor $\Phi_{i,j}$. However, our following argument relies only on the guarantees for predictors Φ_{k,k^*}, which effectively makes our pairwise testing *realizable.*

We now condition on the event defined in (10.11). Let $\mathbf{v}_t^k$ with $k \in [K]$ and $t \in [T]$ be the *surrogate loss* vector, as defined in (10.9). We observe the following key properties

1. We have for all $t \in [T]$ that:

$$\max_{j \in [K]} \sum_{r=1}^{t} \mathbf{v}_r^{k^*}[j] \leq C\left(\frac{\delta}{2K}\right); \tag{10.12}$$

2. For any $k \neq k^*$, we have for all $t \in [T]$:

$$\max_{j \in [K]} \sum_{r=1}^{t} \mathbf{v}_r^{k}[j] \geq \left(\sum_{r=1}^{t} 1\{h_k(\mathbf{x}_r) \neq h_{k^*}(\mathbf{x}_t)\}\right) - C\left(\frac{\delta}{2K}\right). \tag{10.13}$$

The first property follows from the definition of $\mathbf{v}_t^k$ and (10.11). The second property holds since the lower bound is attained when $j = k^*$.

We now analyze the performance of Algorithm 10.2. By property (10.12), we know that $k^* \in S^t$ for all $t \in [T]$, i.e., $|S^t| \geq 1$. Let $N_t = |S^t|$. We define for all $t \in [T]$ the *potential*:

$$E_t = \sum_{k \in S^t} \max\left\{0, 2C(\delta/(2K)) - \sum_{r=1}^{t} 1\{h_k(\mathbf{x}_r) \neq h_{k^*}(\mathbf{x}_r)\}\right\}.$$

Clearly, we have $E_t \leq 2C(\delta/(2K))N_t$. Let $D_t = |\{k \in S^t : h_k(\mathbf{x}_t) \neq h_{k^*}(\mathbf{x}_t)\}|$. We have:

$$D_t \leq N_t - N_{t+1} + E_t - E_{t+1}, \tag{10.14}$$

Since for any $k \in S_t$ such that $h_k(\mathbf{x}_t) \neq h_{k^*}(\mathbf{x}_t)$, either k is removed from S^{t+1} (which contributes at most $N_t - N_{t+1}$), or its contribution to

E_{t+1} decreases by 1 compared to E_t. This follows from the construction of Algorithm 10.2 and property (10.13): once the contribution of k to E_t reaches 0, it must be excluded from S^{t+1}. By the definition of $\hat{y}_t$, we have:

$$\mathbb{E}\left[1\{h_{k^*}(\mathbf{x}_t) \neq \hat{y}_t\}\right] = \frac{D_t}{|S^t|} \leq \frac{N_t - N_{t+1} + E_t - E_{t+1}}{N_t}. \tag{10.15}$$

From [76, Thm 2], we have:

$$\begin{aligned} \sum_{t=1}^{T} \frac{N_t - N_{t+1}}{N_t} &\leq \sum_{t=1}^{T} \left(\frac{1}{N_t} + \frac{1}{N_t - 1} + \cdots + \frac{1}{N_{t+1} + 1} \right) \\ &\leq \sum_{k=1}^{K} \frac{1}{k} \leq \log K. \end{aligned}$$

Moreover, we observe that:

$$\begin{aligned} \sum_{t=1}^{T} \frac{E_t - E_{t+1}}{N_t} &\overset{(a)}{\leq} \frac{2C(\delta/(2K))N_1 - E_2}{N_1} + \sum_{t=2}^{T} \frac{E_t - E_{t+1}}{N_t} \\ &\overset{(b)}{\leq} \frac{2C(\delta/(2K))(N_1 - N_2)}{N_1} \\ &\quad + \frac{2C(\delta/(2K))N_2 - E_3}{N_2} + \sum_{t=3}^{T} \frac{E_t - E_{t+1}}{N_t} \\ &\overset{(c)}{\leq} 2C(\delta/(2K)) \sum_{t=1}^{T} \frac{N_t - N_{t+1}}{N_t} \\ &\leq 2C(\delta/(2K)) \log K, \end{aligned}$$

where (a) and (b) follow by $E_t \leq 2C(\delta/(2K))N_t$ and $N_t \geq N_{t+1}$; (c) follows by repeating the same argument for another $T - 1$ steps.

Therefore, we conclude

$$\mathbb{E}\left[\sum_{t=1}^{T} 1\{h_{k^*}(\mathbf{x}_t) \neq \hat{y}_t\}\right] \leq (1 + 2C(\delta/(2K))) \log K,$$

where the randomness is on the selection of $\hat{k}_t \sim S^t$. Since our selection of $\hat{k}_t$s are independent (conditioning on S^t) for different t, and the indicator is bounded by 1 and non-negative, we can invoke Lemma 2.8

(second part) to obtain a high probability guarantee of confidence $\delta/2$ by introducing an extra $\log(2/\delta)$ additive term. The theorem now follows by a union bound with the event (10.11). □

Remark 10.2. Note that, it is *not* immediately obvious that pairwise testing of two hypotheses can be converted into a general prediction rule a-priori. This is because the underlying true hypothesis is *unknown*, and therefore many pairs tested do not provide any guarantees. We are able to resolve this issue due to the definition of the loss l_t^i (in Algorithm 10.2) for each hypothesis i, which considers the *maximum* loss among all its competitors.

Theorem 10.4 provides a *black box* reduction for converting any testing rule for two hypotheses into a prediction rule for a general hypothesis class $\mathcal{H}$, introducing only an additional $\log|\mathcal{H}|$ factor. This effectively decouples the adversarial properties of the features $\mathbf{x}^T$ from the statistical properties of the noisy labels $\tilde{y}^T$. The rest of this section is devoted to instantiating Theorem 10.4 into various scenarios by providing explicit pairwise testing rules.

10.3.1 Pairwise-Testing via Hellinger Gap.

As discussed above, the risk of noisy online *classification* can be reduced to the *pairwise testing* Φ_{ij} of two hypotheses. However, we still need to construct the explicit pairwise testing rules. This section is devoted to providing a generic testing rule for *general* kernels.

Let h_1, h_2 be any two hypotheses. We may assume that $h_1(\mathbf{x}) \neq h_2(\mathbf{x})$ for all features $\mathbf{x}$, since the agreed features do not impact our pairwise testing risk. We now provide a more compact characterization of the kernel $\mathcal{K}$ without explicitly referring to the feature $\mathbf{x}$. Following the discussion after Definition 10.1, we can fix the feature selection rule ψ^T, and define the kernel by specifying the constrained sets $\mathcal{Q}_y^{\mathbf{x}_t}$ using only prior noisy labels $\tilde{y}^{t-1}$. Thus, we denote $\mathcal{Q}_i^{\tilde{y}^{t-1}} := \mathcal{Q}_{h_i(\mathbf{x}_t)}^{\mathbf{x}_t}$, where $\mathbf{x}_t = \psi_t(\tilde{y}^{t-1})$ and $i \in \{1, 2\}$.

For any $J \leq T$, we denote $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ as the sets of all (joint) distributions over $\tilde{\mathcal{Y}}^J$ induced by the kernel for h_1, h_2, respectively.

Equivalently, $p \in \mathcal{Q}_i^J$ if and only if for all $t \in [J]$ and $\tilde{y}^{t-1} \in \tilde{\mathcal{Y}}^{t-1}$, we have the *conditional* marginal $p_{\tilde{y}_t|\tilde{y}^{t-1}} \in \mathcal{Q}_i^{\tilde{y}^{t-1}}$.

The pairwise testing of h_1, h_2 at time step $J+1$ is then equivalent to the (composite) *hypothesis testing* w.r.t. sets $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$. This is typically resolved using Le Cam-Birgé testing [20, Chapter 32.2] if the distributions are of *product* form. However, this does not serve our purpose, since the distributions in $\mathcal{Q}_i^J$ can have highly correlated marginals. Our main result for addressing this issue is a *conditional* version of Le Cam-Birgé testing, as stated in Theorem 10.5 below. To the best of our knowledge, this conditional version is novel.

Recall that the squared Hellinger divergence is defined as $H^2(\mathcal{P}, \mathcal{Q}) = \inf_{p\in\mathcal{P}, q\in\mathcal{Q}} H^2(p, q)$.

Theorem 10.5 (*Conditional* Le Cam-Birgé Testing)**.** Let $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ be the class of distributions induced by a kernel upto time J, as defined above. If for all $t \in [J]$ and $\tilde{y}^{t-1} \in \tilde{\mathcal{Y}}^{t-1}$, sets $\mathcal{Q}_1^{\tilde{y}^{t-1}}, \mathcal{Q}_2^{\tilde{y}^{t-1}}$ are convex and $H^2(\mathcal{Q}_1^{\tilde{y}^{t-1}}, \mathcal{Q}_2^{\tilde{y}^{t-1}}) \geq \gamma_t$ for some $\gamma_t \geq 0$. Then, there exists a testing rule $\phi : \tilde{\mathcal{Y}}^J \to \{1, 2\}$ such that: [4]

$$\sup_{p\in\mathcal{Q}_1^J, q\in\mathcal{Q}_2^J} \left\{ \Pr_{\tilde{y}^J\sim p}[\phi(\tilde{y}^J) \neq 1] + \Pr_{\tilde{y}^J\sim q}[\phi(\tilde{y}^J) \neq 2] \right\}$$

$$\leq 2\prod_{t=1}^{J}(1 - \gamma_t/2) \leq 2e^{-\frac{1}{2}\sum_{t=1}^{J}\gamma_t}.$$

Sketch. The proof requires a suitable application of the minimax theorem by expressing the testing error as a *linear function* and arguing that $\mathcal{Q}_i^J$s are convex. The error bound is then controlled by a careful application of the *chain-rule* of Rényi divergence. We defer the detailed proof to Section 10.4. □

Theorem 10.5 immediately implies the following *cumulative* risk bound:

[4] Note that the tester ϕ implicitly depends on the feature selector ψ^J. This is not essential for our purposes, since such a dependency can be reduced to that of $\mathbf{x}^J$ (via a more tedious minimax analysis that considers the joint distribution over $\mathbf{x}^J, \tilde{y}^J$), which are observable to the tester.

Proposition 10.1. Let $\{h_1, h_2\} \subset \mathcal{Y}^{\mathcal{X}}$ and $\mathcal{K}$ be a noise kernel. For any $t \in [T]$, we denote $\gamma_t = \inf_{\tilde{y}^{t-1}} H^2(\mathcal{Q}_1^{\tilde{y}^{t-1}}, \mathcal{Q}_2^{\tilde{y}^{t-1}})$, where $\mathcal{Q}_i^{\tilde{y}^{t-1}}$ is the distribution class induced by $\mathcal{K}$ as discussed above. Then, for any $\delta > 0$, the *high probability* cumulative risk:

$$B^\delta(\{h_1, h_2\}, \mathcal{K}) \leq \arg\min_n \left\{ n \in \mathbb{N} : \sum_{t=1}^{n} \gamma_t \geq 2\log(2/\delta) \right\}.$$

Proof. Let n^* be the smallest number satisfying the RHS. If $t \leq n^*$ (this can be checked at each time step t using only $\mathbf{x}^t$ and $\mathcal{K}$), we predict arbitrarily. If $t \geq n^* + 1$, we use the tester ϕ in Theorem 10.5 with $J = n^*$ to produce an index $\hat{\imath} \in \{1, 2\}$ and make the prediction $h_{\hat{\imath}}(\mathbf{x}_t)$ for *all* following time steps. That is, we only use the tester at step $n^* + 1$ and reuse the *same* testing result for all following time steps. By Theorem 10.5, the probability of making errors after step $n^* + 1$ is upper bounded by δ. Therefore, the cumulative risk is upper bounded by n^* with probability $\geq 1 - \delta$. □

Instantiating to the *well-separated* kernels, we arrive at:

Corollary 10.6. Let $\{h_1, h_2\} \subset \mathcal{Y}^{\mathcal{X}}$ and $\mathcal{K}$ be a well-separated kernel with gap γ_{H} under Hellinger distance (Definition 10.3). Then, for any $\delta \geq 0$ we have the *high probability* cumulative risk:

$$B^\delta(\{h_1, h_2\}, \mathcal{K}) \leq \frac{2\log(1/\delta)}{\gamma_{\mathrm{H}}}.$$

Proof. Note that, for any time step t such that $h_1(\mathbf{x}_t) \neq h_2(\mathbf{x}_t)$, we have the gap γ_t in Proposition 10.1 equals γ_{H}. We now have the following prediction rule: for any time step t such that $h_1(\mathbf{x}_t) = h_2(\mathbf{x}_t)$, we predict the agreed label; else, we predict the same way as in Proposition 10.1. Clearly, we only make errors for the second case. By Proposition 10.1, we have that the number of errors is upper bounded by $\frac{2\log(1/\delta)}{\gamma_{\mathrm{H}}}$. □

10.3.2 Characterization for Well-Separated Kernels

In this section, we establish *matching* lower and upper bounds (up to a $\log|\mathcal{H}|$ factor) for the minimax risk of a general multi-class hypothesis class w.r.t. the *Hellinger gap*, in contrast to Theorem 10.1, which applies only to binary label classes w.r.t. L^2 gap.

Theorem 10.7. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a finite class of size K, and $\mathcal{K}$ be a kernel that is well-separated at scale $\gamma_{\mathbf{H}}$ w.r.t. Hellinger divergence. Then, the high probability minimax risk with confidence $\delta > 0$ is upper bounded by:

$$B^{\delta}(\mathcal{H}, \mathcal{K}) \leq \frac{8 \log(4K/\delta) \log K}{\gamma_{\mathbf{H}}} + \log(2/\delta). \tag{10.16}$$

Moreover, for any kernel $\mathcal{K}$ such that there exist at least $\log K$ features $\mathbf{x}$ for which there exists $y \neq y' \in \mathcal{Y}$ such that we have $H^2(\mathcal{Q}_y^{\mathbf{x}}, \mathcal{Q}_{y'}^{\mathbf{x}}) \leq \gamma_{\mathbf{H}}$, then there exists a class $\mathcal{H}$ of size K for which:

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}) \geq \Omega\left(\frac{\log K}{\gamma_{\mathbf{H}}}\right).$$

Proof. By Corollary 10.6, we know that $(\mathcal{H}, \mathcal{K})$ is pairwise testable with error bound $C(\delta) = \frac{2\log(2/\delta)}{\gamma_{\mathbf{H}}}$. The upper bound on *classification* risk then follows from Theorem 10.4 by noticing that $C(\delta/(2K)) = \frac{2\log(4K/\delta)}{\gamma_{\mathbf{H}}}$.

To prove the lower bound, we denote $\tau = \log K$ with $K = |\mathcal{H}|$, and $\mathbf{x}_1, \cdots, \mathbf{x}_\tau$ be τ distinct elements in $\mathcal{X}$ satisfying the condition of the theorem. We define for any $\mathbf{b} \in \{0,1\}^\tau$ a function $h_{\mathbf{b}}$ such that for all $i \in [\tau]$, $h_{\mathbf{b}}(\mathbf{x}_i) = y_i$ if $\mathbf{b}[i] = 0$ and $h_{\mathbf{b}}(\mathbf{x}_i) = y_i'$ otherwise, where $y_i \neq y_i' \in \mathcal{Y}$ are the elements that satisfy $\inf_{p \in \mathcal{Q}_{y_i}^{\mathbf{x}_i}, q \in \mathcal{Q}_{y_i'}^{\mathbf{x}_i}} \{H^2(p,q)\} \leq \gamma_{\mathbf{H}}$. Let $\mathcal{H}$ be the class consisting of all such $h_{\mathbf{b}}$. Let $q_i \in \mathcal{Q}_{y_i}^{\mathbf{x}_i}$ and $q_i' \in \mathcal{Q}_{y_i'}^{\mathbf{x}_i}$ be the elements satisfying $H^2(q_i, q_i') \leq \gamma_{\mathbf{H}}$. We now partition the features $\mathbf{x}^T$ into τ epochs, each of length T/τ, such that each epoch i has constant feature $\mathbf{x}_i$. Let $\mathbf{h}$ be a random function selected uniformly from $\mathcal{H}$. We claim that for any prediction rule $\hat{y}_t$ and any epoch i we have:

$$\mathbb{E}_{\mathbf{h}, \tilde{y}^T}\left[\sum_{t=iT/\tau-1}^{(i+1)T/\tau} 1\{\mathbf{h}(\mathbf{x}_t) \neq \hat{y}_t\}\right] \geq \Omega\left(\frac{1}{\gamma_{\mathbf{H}}}\right), \tag{10.17}$$

where $\tilde{y}_t \sim q_i$ if $\mathbf{h}(\mathbf{x}_i) = y_i$ and $\tilde{y}_t \sim q_i'$ otherwise. The theorem now follows by counting the errors for all τ epochs.

We now establish (10.17) using the Le Cam's two point method. Clearly, for each epoch i, the prediction performance depends only on the label $\mathbf{y}_i = \mathbf{h}(\mathbf{x}_i)$, which is uniform over $\{y_i, y_i'\}$ and independent for different epochs by construction. For any time step j during the ith epoch, we denote by $\tilde{y}^{j-1}$ and $\tilde{y}'^{j-1}$ the samples generated from q_i and q_i', respectively. By the Le Cam's two point method [20, Theorem 7.7] the expected error at step j is lower bounded by:

$$\frac{1 - \mathsf{TV}(\tilde{y}^{j-1}, \tilde{y}'^{j-1})}{2} \geq \frac{1 - \sqrt{H^2(\tilde{y}^{j-1}, \tilde{y}'^{j-1})(1 - H^2(\tilde{y}^{j-1}, \tilde{y}'^{j-1})/4)}}{2} \tag{10.18}$$

where the inequality follows from [20, Equation 7.20]. Note that the RHS of (10.18) is *monotone decreasing* w.r.t. $H^2(\tilde{y}^{j-1}, \tilde{y}'^{j-1})$, since $H^2(p, q) \leq 2$ for all p, q.

By the *tensorization* of Hellinger divergence [20, Equation 7.23], we have:

$$H^2(\tilde{y}^{j-1}, \tilde{y}'^{j-1}) = 2 - 2(1 - H^2(q_i, q_i')/2)^{j-1} \leq 2 - 2(1 - \gamma_{\mathrm{H}}/2)^{j-1},$$

where the last inequality is implied by $H^2(q_i, q_i') \leq \gamma_{\mathrm{H}}$. Using the fact that $\log(1 - x) \geq \frac{-x}{1-x}$, we have, if $\gamma_{\mathrm{H}} \leq 1$ and $j - 1 \leq \frac{1}{\gamma_{\mathrm{H}}}$ then $2 - 2(1 - \gamma_{\mathrm{H}}/2)^{j-1} \leq 2(1 - e^{-1}) < 2$. Therefore, the RHS of (10.18) is lower bounded by an *absolute* positive constant for all $j - 1 \leq \frac{1}{\gamma_{\mathrm{H}}}$, and hence the expected cumulative error will be lower bounded by $\Omega(1/\gamma_{\mathrm{H}})$ during epoch i. This completes the proof. □

It is interesting to note that the bound in Theorem 10.7 is *independent* of *both* the size of label set $\mathcal{Y}$ and the noisy observation set $\tilde{\mathcal{Y}}$, as well as the time horizon T. Moreover, the dependency on the Hellinger gap γ_{H} is *tight* upto only a logarithmic factor $\log |\mathcal{H}|$. This factor is inherent from our reduction to pairwise testing in Algorithm 10.2 and we believe that removing it would require new techniques.

Remark 10.3. Note that $H^2(p, q) \geq 4L^2(p, q)$ holds for any p, q. Thus, the Hellinger dependency of Theorem 10.7 on γ_{H} is tighter than the L^2 dependency of Theorem 10.1. Specifically, if we take p to be the uniform distribution over $\tilde{\mathcal{Y}}$ and q to be the distribution that takes half of the elements with probability mass $\frac{1+\epsilon}{M}$ and half with $\frac{1-\epsilon}{M}$, then, $L^2(p, q) = \frac{\epsilon^2}{M}$, while $H^2(p, q) \geq \Omega(\epsilon^2)$. Therefore, the differences can grow linearly w.r.t. the size of set $\tilde{\mathcal{Y}}$.

10.3.3 Soft-Constrained Gaps

The well-separatedness condition in Theorem 10.1 and Theorem 10.7 requires a *uniform* gap for all $\mathbf{x}_t$s. This may sometimes be too restrictive. We demonstrate in this section that such a “hard" gap can be relaxed to a “soft" gap, while still achieving sub-linear risk.

To this end, we consider a slightly relaxed adversary, where we require that for some constant $A > 0$ and $0 \leq \alpha < 1$, the following soft-constraint holds:

$$\forall r \in (0, 1/2], \frac{1}{T}\sum_{t=1}^{T} 1\left\{\inf_{\tilde{y}^{t-1} \in \tilde{\mathcal{Y}}^{t-1}} \inf_{y \neq y' \in \mathcal{Y}} H^2(\mathcal{Q}_y^{\tilde{y}^{t-1}}, \mathcal{Q}_{y'}^{\tilde{y}^{t-1}}) \leq r\right\} \leq Ar^{\frac{\alpha}{1-\alpha}}, \tag{10.19}$$

where $\mathcal{Q}_y^{\tilde{y}^{t-1}} := \mathcal{Q}_y^{\psi_t(\tilde{y}^{t-1})}$ for some fixed (unknown) feature selector ψ^T as in Section 10.3.1.

The following result follows similarly as Theorem 10.7:

Proposition 10.2. We have:

$$\sup_{\mathcal{K}} \sup_{\mathcal{H}:|\mathcal{H}| \leq K} \tilde{r}_T(\mathcal{H}, \mathcal{K}) = \tilde{\Theta}(T^{1-\alpha}),$$

where the $\tilde{\Theta}$ hides poly-logarithmic factors w.r.t. T and K, and $\mathcal{K}$ runs over all kernels that satisfy (10.19).

Proof. By Theorem 10.4, we only need to consider the testing of two hypotheses $\{h_1, h_2\}$ to derive an upper bound. Let γ be a parameter to be determined later. We have by (10.19) that the number of steps t for which $\inf_{\tilde{y}^{t-1} \in \tilde{\mathcal{Y}}^{t-1}} \inf_{y \neq y' \in \mathcal{Y}} H^2(\mathcal{Q}_y^{\tilde{y}^{t-1}}, \mathcal{Q}_{y'}^{\tilde{y}^{t-1}}) \leq \gamma$ is upper bounded by $A\gamma^{\frac{\alpha}{1-\alpha}}T$. We may assume, w.l.o.g., that all such steps are within the *first* $A\gamma^{\frac{\alpha}{1-\alpha}}T$ time steps, since we can simply filter out such steps (using kernel map $\mathcal{K}$ and the observed features $\mathbf{x}_t$s) when constructing the testing rule. Note that the rest of the steps satisfy for all $\tilde{y}^{t-1}$ and $y \neq y' \in \mathcal{Y}$ that $H^2(\mathcal{Q}_y^{\tilde{y}^{t-1}}, \mathcal{Q}_{y'}^{\tilde{y}^{t-1}}) \geq \gamma$. By Corollary 10.6, the number of errors after step $A\gamma^{\frac{\alpha}{1-\alpha}}T$ is upper bounded by $\tilde{O}(\frac{1}{\gamma})$. Therefore, the total number of errors is upper bounded by

$$\inf_{0 \leq \gamma < 1/2} A\gamma^{\frac{\alpha}{1-\alpha}}T + \frac{2\log(1/\delta)}{\gamma} \leq \tilde{O}(T^{1-\alpha}),$$

where the upper bound follows by taking $\gamma = T^{-(1-\alpha)}$.

To see the lower bound, we define a kernel with the first $A\gamma^{\frac{\alpha}{1-\alpha}}T$ steps of gap γ (to be determined) and define the remaining steps arbitrarily as long as it satisfies (10.19). By Theorem 10.7, we have if $A\gamma^{\frac{\alpha}{1-\alpha}}T \geq \frac{\log|\mathcal{H}|}{\gamma}$, then an $\Omega(\frac{\log|\mathcal{H}|}{\gamma})$ lower bound holds. This is satisfied when taking $\gamma = \left(\frac{\log|\mathcal{H}|}{T}\right)^{1-\alpha}$, which completes the proof. □

10.3.4 Unknown Gap Parameters.

While our previous results provide sub-linear risk that is tight up to poly-logarithmic factors, we have assumed that full knowledge of the kernel sets $\mathcal{Q}_y^{\mathbf{x}_t}$s is available to the learner. In some cases, such information cannot be known completely (or only partially known). For instance, in the classical setting of *Tsybakov noise* as discussed in Diakonikolas et al. (2021), the gap parameters are not assumed to be known.

To account for this, we introduce the following noise kernel, analogous to the *Tsybakov noise* in batch learning. For simplicity, we take $\mathcal{Y} = \tilde{\mathcal{Y}} = \{0, 1\}$. Let $\tilde{y} \in \tilde{\mathcal{Y}}$, we denote $e_{\tilde{y}}$ as the distribution over $\tilde{\mathcal{Y}}$ that assigns probability 1 on $\tilde{y}$ and denote u as uniform distribution over $\tilde{\mathcal{Y}}$. For any $\mathbf{x}^T$, the kernel $\mathcal{K}$ satisfies $\mathcal{Q}_y^{\mathbf{x}_t} = \{\lambda' e_y + (1-\lambda')u : \lambda' \geq \lambda_t\}$, subject to the condition that for some $A > 0$ and $0 \leq \alpha < 1$:

$$\forall r \in (0, 1/2],\ \frac{1}{T}\sum_{t=1}^{T} 1\left\{\frac{\lambda_t}{2} \leq r\right\} \leq A r^{\frac{\alpha}{1-\alpha}}. \tag{10.20}$$

We assume that the parameters λ_ts are (obliviously) selected *independent* of the noisy observation $\tilde{y}^T$. Crucially, we assume that the parameters λ_ts are *unknown* to the learner. Observe that, the set $\mathcal{Q}_y^{\mathbf{x}_t}$ is completely determined by the parameters λ_t and y, irrespective of $\mathbf{x}_t$.

Theorem 10.8. Let $\mathcal{H} \subset \{0, 1\}^{\mathcal{X}}$ be any finite class and $\mathcal{K}$ be a kernel that satisfies condition (10.20). Then, the expected minimax risk is upper bounded by:

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}) \leq \tilde{O}(T^{\frac{2(1-\alpha)}{2-\alpha}}),$$

where $\tilde{O}$ hides poly-logarithmic factors on T and $|\mathcal{H}|$. Moreover, there exist class $\mathcal{H}$ and kernel $\mathcal{K}$ satisfying (10.20), such that:

$$\tilde{r}_T(\mathcal{H},\mathcal{K}) \geq \tilde{\Omega}(T^{\frac{2(1-\alpha)}{2-\alpha}}).$$

Proof. The lower bound follows by the same argument as in Proposition 10.2 by noticing that $H^2(\mathcal{Q}_0^{\mathbf{x}_t},\mathcal{Q}_1^{\mathbf{x}_t}) = \Theta(\lambda_t^2)$ for sufficiently small λ_t. Therefore, it is sufficient to find the λ for which $A\lambda^{\frac{\alpha}{1-\alpha}}T \geq \frac{\log|\mathcal{H}|}{\lambda^2}$. This is satisfied when $\lambda = \left(\frac{\log|\mathcal{H}|}{AT}\right)^{\frac{1-\alpha}{2-\alpha}}$.

For the upper bound, we leverage Theorem 10.4 by constructing an explicit *universal* pairwise testing rule. Let h_1, h_2 be any two hypothesises. We assume, w.l.o.g. (by relabeling), that $h_1(\mathbf{x}) = 0$ and $h_2(\mathbf{x}) = 1$ for all $\mathbf{x}$. At each time step t, we compute the empirical mean $\hat{\mu}_t = \frac{\tilde{y}_1+\cdots+\tilde{y}_{t-1}}{t-1}$, and predict 0 if $\hat{\mu}_t \leq \frac{1}{2}$ and predict 1 otherwise. Let $\lambda_1,\cdots,\lambda_T$ be any configuration of the parameters. Assume, w.l.o.g., that h_1 is the ground truth classifier. We have for any given $\tilde{y}^{t-1}$ the conditional expectation $\mathbb{E}[\tilde{y}_t \mid \tilde{y}^{t-1}] \leq \frac{1}{2} - \frac{\lambda_t}{2}$. By the Hoeffding-Azuma inequality (Lemma 2.7), we have for all $t \in [T]$, the error probability:

$$\Pr\left[\hat{\mu}_t > \frac{1}{2}\right] \leq e^{-(\sum_{i=1}^{t-1}\lambda_i)^2/2(t-1)}.$$

Therefore, for any given $\delta > 0$, we have by the union bound that w.p. $\geq 1-\delta$ the total number of errors made by the predictor is upper bounded by

$$\mathrm{err}_T = \sum_{t=1}^{T} 1\left\{\sum_{j=1}^{t-1}\lambda_j \leq \sqrt{2t\log(T/\delta)}\right\}. \tag{10.21}$$

We now upper bound err_T using property (10.20). Note that, for any given gap parameters $\lambda_1,\cdots,\lambda_T$, the worst configuration for err_T is when $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_T$. To see this, we use the following "switching" argument. Suppose otherwise, there exists some j for which $\lambda_{j+1} < \lambda_j$. We show that switching λ_j and λ_{j+1} will not decrease err_T. This follows from the fact that the switch will not effect any time steps except step $j+1$ in which case the sum of gap parameters *decreases*. We can therefore assume, w.l.o.g., that the gap parameters are monotone increasing. Now, we have by (10.20) that for all $j \in [T]$:

$$\sum_{t=1}^{T} 1\left\{\lambda_t \leq (j/AT)^{\frac{1-\alpha}{\alpha}}\right\} \leq j.$$

This implies that for any time step j, we have $\lambda_j > \left(\frac{j}{AT}\right)^{\frac{1-\alpha}{\alpha}}$ since the gap parameters are monotone *increasing*. Therefore, by integration approximation, we have:

$$\sum_{j=1}^{n} \lambda_j \geq \Omega(n^{\frac{1}{\alpha}} T^{-\frac{1-\alpha}{\alpha}}).$$

Setting $n^{\frac{1}{\alpha}} T^{-\frac{1-\alpha}{\alpha}} \leq n^{\frac{1}{2}} \cdot \sqrt{2\log(T/\delta)}$, we find that $n = \tilde{O}(T^{\frac{2(1-\alpha)}{2-\alpha}})$. This implies that for any time step $t \geq n$, the t'th indicator in (10.21) equals 0. Therefore, the risk of pairwise testing is upper bound by $\mathsf{err}_T \leq \tilde{O}(T^{\frac{2(1-\alpha)}{2-\alpha}})$ w.p. $\geq 1-\delta$, where $\tilde{O}$ hides the factor $\log(T/\delta)$. The upper bound of the theorem now follows by Theorem 10.4. □

Remark 10.4. Observe that the lower and upper bounds of Theorem 10.8 *match* up to poly-logarithmic factors w.r.t. T and $|\mathcal{H}|$. Moreover, the proof technique for the upper bound can be generalized to the case when $\mathcal{Q}_0^{\mathbf{x}}$ encompasses *any* distributions over $[0,1]$ with means in $[0, \frac{1-\lambda_t}{2}]$ (and in $[\frac{1+\lambda_t}{2}, 1]$ for $\mathcal{Q}_1^{\mathbf{x}}$), not only for Bernoulli distributions as in (10.20).

Note that, the pairwise testing rule derived in the proof of Theorem 10.8 requires no information about the underlying distributions. This differs from the general testing rule derived from Theorem 10.5, which requires the likelihood ratio of distributions $p_1^* \in \mathcal{Q}_1^J$ and $p_2^* \in \mathcal{Q}_2^J$ that achieve $||p_1^* - p_2^*||\mathsf{TV} = \mathsf{TV}(\mathcal{Q}_1^J, \mathcal{Q}_2^J)$ (see Section 10.4).

10.4 Proof of Theorem 10.5

We start with an application of the minimax theorem to hypothesis testing [5].

Lemma 10.9. Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be two sets of distributions over a finite domain Ω. If $\mathcal{P}_0$ and $\mathcal{P}_1$ are convex under L_1 distance (i.e., total variation), then

$$\min_{\phi\,:\,\Omega\to[0,1]} \sup_{p_0\in\mathcal{P}_0, p_1\in\mathcal{P}_1} \{\mathbb{E}_{\omega\sim p_0}[1-\phi(\omega)] + \mathbb{E}_{\omega\sim p_1}[\phi(\omega)]\} =$$

[5] This result was mentioned in [20, Chapter 32.2], without providing a proof.

$$= 1 - \inf_{p_0 \in \mathcal{P}_0, p_1 \in \mathcal{P}_1} ||p_0 - p_1||_{\mathsf{TV}}.$$

Moreover, if ϕ^* is the function that attains minimal, then the tester $\psi^*(\omega) = 1\{\phi^*(\omega) < 0.5\}$ achieves:

$$\sup_{p_0 \in \mathcal{P}_0, p_1 \in \mathcal{P}_1} \{\Pr_{\omega \sim p_0}[\psi^*(\omega) \neq 0] + \Pr_{\omega \sim p_1}[\psi^*(\omega) \neq 1]\} \leq$$

$$\leq 2(1 - \inf_{p_0 \in \mathcal{P}_0, p_1 \in \mathcal{P}_1} ||p_0 - p_1||_{\mathsf{TV}}).$$

Proof. Observe that the function ϕ can be viewed as a vector in $[0,1]^\Omega$. Moreover, the distributions over Ω can be viewed as vectors in $[0,1]^\Omega$ as well. Therefore, we have

$$\mathbb{E}_{\omega \sim p_0}[1 - \phi(\omega)] + \mathbb{E}_{\omega \sim p_1}[\phi(\omega)] = \langle p_0, 1 - \phi \rangle + \langle p_1, \phi \rangle,$$

which is a linear function w.r.t. both (p_0, p_1) and ϕ. Since the both $\mathcal{P}_0 \times \mathcal{P}_1$ and $[0,1]^\Omega$ are convex and $[0,1]^\Omega$ is compact, we can invoke the minimax theorem [9, Thm 7.1] to obtain:

$$\begin{aligned}\min_{\phi\,:\,\Omega \to [0,1]} \sup_{p_0 \in \mathcal{P}_0, p_1 \in \mathcal{P}_1} &\{\mathbb{E}_{\omega \sim p_0}[1 - \phi(\omega)] + \mathbb{E}_{\omega \sim p_1}[\phi(\omega)]\} \\ &= \sup_{p_0 \in \mathcal{P}_0, p_1 \in \mathcal{P}_1} \min_{\phi\,:\,\Omega \to [0,1]} \{\mathbb{E}_{\omega \sim p_0}[1 - \phi(\omega)] + \mathbb{E}_{\omega \sim p_1}[\phi(\omega)]\} \\ &= \sup_{p_0 \in \mathcal{P}_0, p_1 \in \mathcal{P}_1} \{1 - ||p_0 - p_1||_{\mathsf{TV}}\},\end{aligned}$$

where the last equality follows by Le Cam's two point lemma [20, Theorem 7.7]. Let ϕ^* be the function that attains minimal and $\psi^*(\omega) = 1\{\phi^*(\omega) < 0.5\}$. We have $1\{\psi^*(\omega) \neq i\} \leq 2(1 - i - \phi^*(\omega))$ for all $i \in \{0,1\}$. To see this, for $i = 0$, we have $\psi^*(\omega) \neq 0$ only if $\phi^*(\omega) < 0.5$, thus $1 - \phi^*(\omega) \geq 0.5$ (the case for $i = 1$ follows similarly). Therefore, we have for all $p_0 \in \mathcal{P}_0, p_1 \in \mathcal{P}_1$:

$$\Pr_{\omega \sim p_0}[\psi^*(\omega) \neq 0] + \Pr_{\omega \sim p_1}[\psi^*(\omega) \neq 1] \leq 2(\mathbb{E}_{\omega \sim p_0}[1 - \phi^*(\omega)] + \mathbb{E}_{\omega \sim p_1}[\phi^*(\omega)]).$$

This completes the proof. □

We have the following key property:

Lemma 10.10. Let $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ be the sets in Theorem 10.5. Then $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ are convex.

Proof. Let $p_1, p_2 \in \mathcal{Q}_i^J$ for $i \in \{1,2\}$ and $\lambda \in [0,1]$. We need to show that $p = \lambda p_1 + (1-\lambda)p_2 \in \mathcal{Q}_i^J$ as well. For any given $t \in [J]$ and $\tilde{y}^{t-1} \in \tilde{\mathcal{Y}}^{t-1}$, we have:

$$\begin{aligned} p(\tilde{y}_t \mid \tilde{y}^{t-1}) &= \frac{\lambda p_1(\tilde{y}^t) + (1-\lambda)p_2(\tilde{y}^t)}{\lambda p_1(\tilde{y}^{t-1}) + (1-\lambda)p_2(\tilde{y}^{t-1})} \\ &= \lambda \frac{p_1(\tilde{y}^{t-1})}{p(\tilde{y}^{t-1})} p_1(\tilde{y}_t \mid \tilde{y}^{t-1}) + (1-\lambda)\frac{p_2(\tilde{y}^{t-1})}{p(\tilde{y}^{t-1})} p_2(\tilde{y}_t \mid \tilde{y}^{t-1}) \in \mathcal{Q}_i^{\tilde{y}^{t-1}} \end{aligned}$$

where the last inclusion follows by convexity of $\mathcal{Q}_i^{\tilde{y}^{t-1}}$ as assumed in Theorem 10.5. Therefore, we have $p \in \mathcal{Q}_i^J$ by definition of $\mathcal{Q}_i$. □

Now, our main technical problem is to bound the total variation $\mathsf{TV}(\mathcal{Q}_1^J, \mathcal{Q}_2^J)$. The primary challenge comes from controlling the dependencies of conditional marginals of the distributions. To proceed, we now introduce the concept of *Renyi divergence*. Let p_1, p_2 be two distributions over the same finite domain Ω, the α-Renyi divergence is defined as:

$$D_\alpha(p_1, p_2) = \frac{1}{\alpha - 1} \log \mathbb{E}_{\omega \sim p_2}\left[\left(\frac{p_1(\omega)}{p_2(\omega)}\right)^\alpha\right].$$

If p, q are distributions over domain $\Omega_1 \times \Omega_2$ and r is a distribution over Ω_1, then the *conditional* α-Renyi divergence is defined as:

$$D_\alpha(p, q \mid r) = \frac{1}{\alpha - 1} \log \mathbb{E}_{\omega_1 \sim r}\left[\sum_{\omega_2 \in \Omega_2} p(\omega_2 \mid \omega_1)^\alpha q(\omega_2 \mid \omega_1)^{1-\alpha}\right].$$

The following property about Renyi divergence is well known [20, Chapter 7.12]:

Lemma 10.11. Let p, q be two distributions over $\Omega_1 \times \Omega_2$ and $p^{(1)}$ and $q^{(1)}$ be the restrictions of p, q on Ω_1, respectively. Then the following chain rule holds:

$$D_\alpha(p, q) = D_\alpha(p^{(1)}, q^{(1)}) + D_\alpha(p, q \mid r),$$

where $r(\omega_1) = p^{(1)}(\omega_1)^\alpha q^{(1)}(\omega_1)^{1-\alpha} e^{-(\alpha-1)D_\alpha(p^{(1)}, q^{(1)})}$ is a distribution over Ω_1.

We now arrive at our main technical result for bounding the Renyi divergence between $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ in Theorem 10.5:

Proposition 10.3. Let $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ be the sets in Theorem 10.5. If for all $t \in [J]$ and $\tilde{y}^{t-1} \in \tilde{\mathcal{Y}}^{t-1}$, we have $\inf_{p_t \in \mathcal{Q}_1^{\tilde{y}^{t-1}}, q_t \in \mathcal{Q}_2^{\tilde{y}^{t-1}}} D_\alpha(p_t, q_t) \geq \eta_t$ for some $\eta_t \geq 0$. Then

$$\inf_{p \in \mathcal{Q}_1^J, q \in \mathcal{Q}_2^J} D_\alpha(p, q) \geq \sum_{t=1}^{J} \eta_t.$$

Proof. We prove by induction on J. The base case for $J = 1$ is trivial. We now prove the induction step with $J \geq 2$. For any pair $p \in \mathcal{Q}_1^J$ and $q \in \mathcal{Q}_2^J$, we have by Lemma 10.11 that $D_\alpha(p, q) = D_\alpha(p^{(1)}, q^{(1)}) + D_\alpha(p, q \mid r)$, where $p^{(1)}$, $q^{(1)}$ are restrictions of p, q on $\tilde{y}^{J-1}$ and r is a distribution over $\tilde{\mathcal{Y}}^{J-1}$. By definition of α-Renyi divergence, we have:

$$\begin{aligned} D_\alpha(p, q \mid r) &\geq \inf_{\tilde{y}^{J-1}} \frac{1}{\alpha - 1} \log \sum_{\tilde{y}_J \in \tilde{\mathcal{Y}}} p(\tilde{y}_J \mid \tilde{y}^{J-1})^\alpha q(\tilde{y}_J \mid \tilde{y}^{J-1})^{1-\alpha} \\ &= \inf_{\tilde{y}^{J-1}} D_\alpha(p_{\tilde{y}_J \mid \tilde{y}^{J-1}}, q_{\tilde{y}_J \mid \tilde{y}^{J-1}}) \\ &\overset{(a)}{\geq} \inf_{p \in \mathcal{Q}_1^{\tilde{y}^J}, q \in \mathcal{Q}_2^{\tilde{y}^J}} D_\alpha(p, q) \overset{(b)}{\geq} \eta_J, \end{aligned}$$

where (a) follows since $p_{\tilde{y}_J \mid \tilde{y}^{J-1}} \in \mathcal{Q}_1^{\tilde{y}^J}$ and $q_{\tilde{y}_J \mid \tilde{y}^{J-1}} \in \mathcal{Q}_2^{\tilde{y}^J}$ by the definition of $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$; (b) follows by assumption. The result then follows by induction hypothesis $D_\alpha(p^{(1)}, q^{(1)}) \geq \sum_{t=1}^{J-1} \eta_t$, since $p^{(1)} \in \mathcal{Q}_1^{J-1}$ and $q^{(1)} \in \mathcal{Q}_2^{J-1}$. □

The following result converts the Renyi divergence based bounds to that with Hellinger divergence.

Proposition 10.4. Let $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ be the sets in Theorem 10.5. If for all $t \in [J]$ and $\tilde{y}^{t-1} \in \tilde{\mathcal{Y}}^{t-1}$, we have $H^2(\mathcal{Q}_1^{\tilde{y}^{t-1}}, \mathcal{Q}_2^{\tilde{y}^{t-1}}) \geq \gamma_t$ for some $\gamma_t \geq 0$. Then:

$$\inf_{p \in \mathcal{Q}_1^J, q \in \mathcal{Q}_2^J} H^2(p, q) \geq 2 \left(1 - \prod_{t=1}^{J} (1 - \gamma_t/2)\right).$$

Proof. Observe that, for any distributions p, q we have:

$$H^2(p,q) = 2(1 - e^{-\frac{1}{2}D_{1/2}(p,q)}). \tag{10.22}$$

Specifically, for given $p \in \mathcal{Q}_1^J$ and $q \in \mathcal{Q}_2^J$, we have:

$$1 - H^2(p,q)/2 = e^{-\frac{1}{2}D_{1/2}(p,q)} \leq e^{-\frac{1}{2}\sum_{t=1}^J \eta_t} = \prod_{t=1}^J e^{-\frac{1}{2}\eta_t} \leq \prod_{t=1}^J (1 - \gamma_t/2),$$

where η_ts are the constants in Proposition 10.3 and the last inequality follows by $e^{-\frac{1}{2}\eta_t} \leq 1 - \gamma_t/2$ due to (10.22) again. This completes the proof. □

Proof of Theorem 10.5. We have by Lemma 10.9 that the testing error is upper bounded by $1 - \inf_{p \in \mathcal{Q}_1, q \in \mathcal{Q}_2} ||p - q||_{\mathsf{TV}}$. Fix any pair p, q, we have by relation between Hellinger and total variation that $1 - ||p - q||_{\mathsf{TV}} \leq 1 - \frac{1}{2}H^2(p,q)$. The result follows by Proposition 10.4. □

10.5 Tight Bounds via Log-loss

In this section, we introduce a refined technique based on the reduction to *online conditional distribution estimation* as discussed in Section 10.2. We shall use again Lemma 10.3 but with *log-loss*. This yields tight risk dependency on *both* $\log |\mathcal{H}|$ and the gap parameter for certain special, yet important, noise kernels.

10.5.1 The Randomized Response Mechanism

Let $\mathcal{Y} = \tilde{\mathcal{Y}} = \{1, \cdots, M\}$. We denote by u the *uniform* distribution over $\tilde{\mathcal{Y}}$ and $e_{\tilde{y}}$ the distribution that assigns probability 1 on $\tilde{y} \in \tilde{\mathcal{Y}}$. For any $\eta > 0$, we define a *homogeneous* (i.e., independent of $\mathbf{x}$) kernel:

$$\forall \mathbf{x} \in \mathcal{X},\ y \in \mathcal{Y},\ \mathcal{K}^\eta(\mathbf{x}, y) = \{(1 - \eta')e_y + \eta' u : \eta' \in [0, \eta)\}.$$

Note that, this kernel can be interpreted as the *randomized response mechanism* with multiple outcomes in differential privacy [83], where η is interpreted as the noise level of *perturbing* the true labels. For instance, it achieves $(\epsilon, 0)$-local differential privacy if we set $\eta = \frac{M}{e^\epsilon - 1 + M}$.

Theorem 10.12. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be any finite class and $\mathcal{K}^\eta$ be as defined above with $0 \leq \eta < 1$. Then, the expected minimax risk is upper bounded by:

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}^\eta) \leq \frac{\log |\mathcal{H}|}{(1-\eta)^2/2}.$$

Moreover, the *high probability* minimax risk at confidence $\delta > 0$ is upper bounded by:

$$B^\delta(\mathcal{H}, \mathcal{K}^\eta) \leq \frac{\log |\mathcal{H}| + 2\log(1/\delta)}{(1-\eta)^2/4}.$$

Furthermore, for $1 - \eta \ll \frac{1}{M}$ we have $B^\delta(\mathcal{H}, \mathcal{K}^\eta) \leq O\left(\frac{\log|\mathcal{H}|+\log(1/\delta)}{M(1-\eta)^2}\right)$.

Proof. Our proof follows a similar path as the proof of Theorem 10.1. For any $h \in \mathcal{H}$, we define a distribution-valued function f_h such that $f_h(\mathbf{x}) = (1-\eta)e_{h(\mathbf{x})} + \eta u$. Let $\mathcal{F} = \{f_h : h \in \mathcal{H}\}$. Invoking Lemma 10.3 with log-loss and using the fact the KL-divergence is Bregman and 1-Exp-concave, there exist estimators $\hat{p}^T$ such that:

$$\sup_{f \in \mathcal{F}} \mathbb{Q}_{\mathcal{K}}^T \left[\sum_{t=1}^T \mathsf{KL}(\tilde{p}_t, \hat{p}_t) - \mathsf{KL}(\tilde{p}_t, f(\mathbf{x}_t)) \right] \leq \log |\mathcal{H}|,$$

where $\mathbb{Q}_{\mathcal{K}}^T$ is the operator in Definition 10.1. We now define the following classifier:

$$\hat{y}_t = \arg\max_y \{\hat{p}_t[y] : y \in \mathcal{Y}\}.$$

Note that, this is a *multi-class* classifier. Let $h^* \in \mathcal{H}$ be the underlying true classification function and $\tilde{p}^T$ be the noisy label distributions selected by the adversary. We have:

Lemma 10.13. The following holds for all $t \leq T$:

$$\mathsf{KL}(\tilde{p}_t, \hat{p}_t) - \mathsf{KL}(\tilde{p}_t, f_{h^*}(\mathbf{x}_t)) \geq 0.$$

Moreover, if $\hat{y}_t \neq h^*(\mathbf{x}_t)$ then:

$$\mathsf{KL}(\tilde{p}_t, \hat{p}_t) - \mathsf{KL}(\tilde{p}_t, f_{h^*}(\mathbf{x}_t)) \geq (1-\eta)^2/2.$$

Proof of the Lemma. Let $y_t = h^*(\mathbf{x}_t)$ and $e_t \in \mathcal{D}(\tilde{\mathcal{Y}})$ be the distribution that assigns probability 1 on y_t. By the definition $f_{h^*}(\mathbf{x}_t) = \lambda e_t + (1-\lambda)u$ and $\tilde{p}_t = \lambda_t e_t + (1-\lambda_t)u$, where $\lambda = 1-\eta$ and $\lambda_t = 1-\eta_t$ for some $\eta_t \leq \eta$. Since $0 \leq \eta_t \leq \eta$, we have $1 \geq \lambda_t \geq \lambda$. Note that, $\mathsf{KL}(\tilde{p}_t, \hat{p}_t) - \mathsf{KL}(\tilde{p}_t, f_{h^*}(\mathbf{x}_t))$ is a linear function w.r.t. λ_t (Proposition 2.1), and it takes the minimal value at $\lambda_t \in \{1, \lambda\}$; therefore:

$$\mathsf{KL}(\tilde{p}_t, \hat{p}_t) - \mathsf{KL}(\tilde{p}_t, f_{h^*}(\mathbf{x}_t)) \geq \min\{\log(f_{h^*}(\mathbf{x}_t)[y_t]/\hat{p}_t[y_t]), \mathsf{KL}(f_{h^*}(\mathbf{x}_t), \hat{p}_t)\}.$$

Clearly, the second KL-divergence term is positive. We now show that $\log(f_{h^*}(\mathbf{x}_t)[y_t]/\hat{p}_t[y_t]) \geq 0$. To see this, we have by Lemma 10.3 that $\hat{p}_t$ is a *convex* combination of $\{f(\mathbf{x}_t) : f \in \mathcal{F}\}$ and therefore $\hat{p}_t = \lambda a_t + (1-\lambda)u$ for some $a_t \in \mathcal{D}(\tilde{\mathcal{Y}})$. This implies that $\hat{p}_t[y_t] = \lambda a_t[y_t] + (1-\lambda)\frac{1}{M}$ and $f_{h^*}(\mathbf{x}_t)[y_t] = \lambda + (1-\lambda)\frac{1}{M}$. Since $a_t[y_t] \leq 1$, we have $f_{h^*}(\mathbf{x}_t)[y_t] \geq \hat{p}_t[y_t]$. The first part of the claim now follows.

We now prove the second part of the claim. Note that in order for $\hat{y}_t \neq y_t$ we must have $a_t[y_t] \leq \frac{1}{2}$, since $\hat{y}_t$ is defined to be the label with maximum probability mass under $\hat{p}_t$. Therefore,

$$\begin{aligned} \log(f_{h^*}(\mathbf{x}_t)[y_t]/\hat{p}_t[y_t]) &\geq \log\left(\frac{\lambda + (1-\lambda)/M}{\lambda/2 + (1-\lambda)/M}\right) \\ &= \log\left(1 + \frac{\lambda/2}{\lambda/2 + (1-\lambda)/M}\right) \geq \log(1+\lambda) \end{aligned}$$

where the second inequality follows from $\lambda/2 + (1-\lambda)/M \leq 1/2$. Furthermore, we have:

$$\mathsf{KL}(f_{h^*}(\mathbf{x}_t), \hat{p}_t) \geq \frac{1}{2}||f_{h^*}(\mathbf{x}_t) - \hat{p}_t||_1^2 \geq \lambda^2/2,$$

where the first inequality is a consequence of Pinsker's inequality [20] and the second inequality follows by $||f_{h^*}(\mathbf{x}_t) - \hat{p}_t||_1 = \lambda||e_{y_t} - a_t||_1 = \lambda(2|1 - a_t[y_t]|) \geq \lambda$, since $a_t[y_t] \leq \frac{1}{2}$. The claim now follows by the fact that $\log(1+\lambda) \geq \lambda^2/2$ for all $0 \leq \lambda \leq 1$. □

The first part of the theorem now follows by the same argument as the proof of Theorem 10.1. The proof of the second and third parts requires a careful analysis relating log-loss with the Hellinger distance and employing a martingale concentration inequality similar to [84, Lemma A.14]. We defer the technical proof to Section 10.5.3 for readability. □

To complement the upper bounds of Theorem 10.12, we have the following matching lower bound follows directly from Theorem 10.7:

Corollary 10.14. There exists a class $\mathcal{H}$ such that for $1-\eta \ll \frac{1}{M}$ we have:

$$\tilde{r}(\mathcal{H}, \mathcal{K}^\eta) \geq \Omega\left(\frac{\log|\mathcal{H}|}{M(1-\eta)^2}\right).$$

Proof. Specializing to the setting in Theorem 10.7, we know that the squared Helliger gap is of order:

$$\left(\sqrt{\frac{\eta}{M}} - \sqrt{1-\frac{(M-1)\eta}{M}}\right)^2 \sim \frac{M(1-\eta)^2}{4},$$

when $1-\eta \ll \frac{1}{M}$ (by Taylor expansion). This implies an $\Omega\left(\frac{\log|\mathcal{H}|}{M(1-\eta)^2}\right)$ lower bound. □

Remark 10.5. Taking $\eta = \frac{M}{e^\epsilon - 1 + M}$ for sufficiently small ϵ, we have

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}^\eta) = \Theta\left(\frac{M \log|\mathcal{H}|}{\epsilon^2}\right),$$

and the randomized response mechanism with kernel $\mathcal{K}^\eta$ achieves $(\epsilon, 0)$-local differential privacy. This holds even when the noise parameters used by different local parties vary, as long as they are upper bounded by η.

10.5.2 Kernel Set of Size One

In this section, we establish an upper bound for the special case when the kernel set size $|\mathcal{Q}_y^{\mathbf{x}}| = 1$ for all $\mathbf{x}, y$. This matches the lower bound in Theorem 10.7 up to a *constant* factor.

Theorem 10.15. Let $\mathcal{H} \subset \mathcal{Y}^\mathcal{X}$ be any finite class and $\mathcal{K}$ be any noise kernel that is well-separated at scale γ_{H} w.r.t. squared Hellinger distance such that $|\mathcal{Q}_y^{\mathbf{x}}| = 1$ for all $\mathbf{x}, y$. Then the high probability minimax risk at confidence $\delta > 0$ is upper bounded by:

$$B^\delta(\mathcal{H}, \mathcal{K}) \leq O\left(\frac{\log(|\mathcal{H}|/\delta)}{\gamma_{\text{H}}}\right).$$

Proof. Our proof follows a similar path as in the proof of Theorem 3.3, but replaces L^2 loss with log-loss. Specifically, for any $h \in \mathcal{H}$, we define

$f_h(\mathbf{x}) = q^{\mathbf{x}}_{h(\mathbf{x})}$, where $q^{\mathbf{x}}_{h(\mathbf{x})}$ is the unique element in $\mathcal{Q}^{\mathbf{x}}_{h(\mathbf{x})}$. Denote $\mathcal{F} = \{f_h : h \in \mathcal{H}\}$. We run the EWA algorithm (Algorithm 3.1) over $\mathcal{F}$ with $\eta = 1$ and ℓ being the log-loss, and produce an estimator $\hat{p}^T$. The classifier is then given by:

$$\hat{y}_t = \arg\min_{y \in \mathcal{Y}} \{H^2(q^{\mathbf{x}_t}_y, \hat{p}_t)\}.$$

Now, our key observation is that the noisy label distribution $\tilde{p}_t = f_{h^*}(\mathbf{x}_t)$ is *well-specified* (since $|\mathcal{Q}^{\mathbf{x}}_y| = 1$, the only choice for $\tilde{p}_t$ is $f_{h^*(\mathbf{x}_t)}$), where h^* is the ground truth classifier. Therefore, invoking [84, Lemma A.14], we find:

$$\Pr\left[\sum_{t=1}^{T} H^2(\tilde{p}_t, \hat{p}_t) \leq \log|\mathcal{F}| + 2\log(1/\delta)\right] \geq 1 - \delta.$$

We claim that $1\{\hat{y}_t \neq h^*(\mathbf{x}_t)\} \leq \frac{4}{\gamma_{\mathrm{H}}} H^2(\tilde{p}_t, \hat{p}_t)$. Clearly, this automatically satisfies if $\hat{y}_t = h^*(\mathbf{x}_t)$. For $\hat{y}_t \neq h^*(\mathbf{x}_t)$, we have $H^2(q^{\mathbf{x}_t}_{\hat{y}_t}, \hat{p}_t) \leq H^2(q^{\mathbf{x}_t}_{h^*(\mathbf{x}_t)}, \hat{p}_t) = H^2(\tilde{p}_t, \hat{p}_t)$ by definition of $\hat{y}_t$. This implies that:

$$H^2(\tilde{p}, \hat{p}_t) \geq \frac{1}{4} H^2(q^{\mathbf{x}_t}_{\hat{y}_t}, q^{\mathbf{x}_t}_{h^*(\mathbf{x}_t)}) \geq \frac{\gamma_{\mathrm{H}}}{4},$$

where the first inequality follows by triangle inequality of Hellinger distance (the factor $\frac{1}{4}$ comes from the conversion from squared Hellinger distance to Hellinger distance), and the second inequality follows by definition of γ_{H}. Therefore, we have w.p. $\geq 1 - \delta$ that:

$$\sum_{t=1}^{T} 1\{\hat{y}_t \neq h^*(\mathbf{x}_t)\} \leq \frac{4}{\gamma_{\mathrm{H}}}(\log|\mathcal{F}| + 2\log(1/\delta)).$$

This completes the proof since $|\mathcal{H}| \geq |\mathcal{F}|$. □

Observe that the key ingredient in the proof of Theorem 10.15 is the realizability of $\tilde{p}_t$ by f_{h^*} due to the property $|\mathcal{Q}^{\mathbf{x}}_y| = 1$, which does not hold for general kernels.

10.5.3 Proof of High Probability Minimax Risk of Theorem 10.12

We begin with the following key inequality:

Lemma 10.16. Let $\tilde{p} = (1-\eta')e_{\tilde{y}} + \eta' u$, $p = (1-\eta)e_{\tilde{y}} + \eta u$ and $\hat{p} = (1-\eta)a + \eta u$, where $e_{\tilde{y}}, a, u \in \mathcal{D}(\tilde{\mathcal{Y}})$ and $0 \le \eta' \le \eta < 1$, such that $e_{\tilde{y}}$ is the distribution assigning probability 1 on $\tilde{y}$, u is uniform over $\tilde{\mathcal{Y}}$ and $a \in \mathcal{D}(\tilde{\mathcal{Y}})$ is arbitrary. Then:

$$\sum_{\tilde{y}' \in \tilde{\mathcal{Y}}} \tilde{p}[\tilde{y}'] \sqrt{\frac{\hat{p}[\tilde{y}']}{p[\tilde{y}']}} \le \sum_{\tilde{y}' \in \tilde{\mathcal{Y}}} p[\tilde{y}'] \sqrt{\frac{\hat{p}[\tilde{y}']}{p[\tilde{y}']}} = \sum_{\tilde{y}' \in \tilde{\mathcal{Y}}} \sqrt{p[\tilde{y}']\hat{p}[\tilde{y}']}. \tag{10.23}$$

Proof. Denote $|\tilde{\mathcal{Y}}| = M$, and let $r \in \mathbb{R}^{\tilde{\mathcal{Y}}}$ be the vector such that $r[\tilde{y}'] = \sqrt{\hat{p}[\tilde{y}']/p[\tilde{y}']}$. We have the LHS of (10.23) equals $e_{\tilde{y}}^{\mathsf{T}} r + \eta'(u - e_{\tilde{y}})^{\mathsf{T}} r$. We claim that $f(\eta') \overset{\mathsf{def}}{=} e_{\tilde{y}}^{\mathsf{T}} r + \eta'(u - e_{\tilde{y}})^{\mathsf{T}} r$ attains maximum when $\eta' = \eta$, which will finish the proof. It is sufficient to prove that $(u - e_{\tilde{y}})^{\mathsf{T}} r \ge 0$ since $f(\eta')$ is a linear function w.r.t. η'. We have:

$$u^{\mathsf{T}} r = \frac{1}{M} \sum_{\tilde{y}' \in \tilde{\mathcal{Y}}} \sqrt{\frac{\hat{p}[\tilde{y}']}{p[\tilde{y}']}},\ e_{\tilde{y}}^{\mathsf{T}} r = \sqrt{\frac{\hat{p}[\tilde{y}]}{p[\tilde{y}]}}.$$

We only need to show that $\forall \tilde{y}' \in \tilde{\mathcal{Y}}$ with $\tilde{y}' \ne \tilde{y}$, we have $\sqrt{\hat{p}[\tilde{y}']/p[\tilde{y}']} \ge \sqrt{\hat{p}[\tilde{y}]/p[\tilde{y}]}$, i.e.,

$$\frac{p[\tilde{y}]}{p[\tilde{y}']} \ge \frac{\hat{p}[\tilde{y}]}{\hat{p}[\tilde{y}']}.$$

Note that, $p[\tilde{y}] = 1 - \eta + \frac{\eta}{M}$, $p[\tilde{y}'] = \frac{\eta}{M}$, $\hat{p}[\tilde{y}] = (1-\eta)a[\tilde{y}] + \frac{\eta}{M}$ and $\hat{p}[\tilde{y}'] = (1-\eta)a[\tilde{y}'] + \frac{\eta}{M}$, i.e., we have $p[\tilde{y}] \ge \hat{p}[\tilde{y}], \hat{p}[\tilde{y}'] \ge p[\tilde{y}]$. The result now follows by the simple fact that for *any* $a \ge b, c \ge d \ge 0$ we have $\frac{a}{d} \ge \frac{b}{c}$. □

We are now ready to state our main result, which establishes the high probability bounds in Theorem 10.12.

Theorem 10.17. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be any finite class and $\mathcal{K}^{\eta}$ be the kernel in Section 10.5 with $0 \le \eta < 1$. Then, the *high probability* minimax risk at confidence δ is upper bounded by:

$$B^{\delta}(\mathcal{H}, \mathsf{P}, \mathcal{K}^{\eta}) \le \frac{\log|\mathcal{H}| + 2\log(1/\delta)}{(1-\eta)^2/4}.$$

Furthermore, for $1 - \eta \ll \frac{1}{M}$ we have $B^{\delta}(\mathcal{H}, \mathsf{P}, \mathcal{K}^{\eta}) \le O\left(\frac{\log|\mathcal{H}| + \log(1/\delta)}{M(1-\eta)^2}\right)$.

Proof. Let $\mathcal{F}$ be the class as in the proof of Theorem 10.12 and $\hat{p}_t$ be produced by the EWA algorithm under Log-loss. We have by Proposition 2.2 and Theorem 3.5 that for *any* $\tilde{y}^T \in \tilde{\mathcal{Y}}^T$:

$$\sup_{\mathbf{x}^T \in \mathcal{X}^T} \sum_{t=1}^{T} \log \frac{f^*(\mathbf{x}_t)[\tilde{y}_t]}{\hat{p}_t[\tilde{y}_t]} \leq \log |\mathcal{F}|$$

where f^* is the corresponding function of the underlying truth $h^* \in \mathcal{H}$ (see the proof of Theorem 10.12). We now assume $\tilde{y}^T$ are sampled from $\tilde{p}^T$, where $\tilde{p}^T$ are the noisy label distributions selected by the adversary. Denote by $\mathbb{E}_t$ the conditional expectation on $\tilde{y}^{t-1}$. We have:

$$\mathbb{E}_t\left[e^{-\frac{1}{2}\log\frac{f^*(\mathbf{x}_t)[\tilde{y}_t]}{\hat{p}_t[\tilde{y}_t]}}\right] = \mathbb{E}_{\tilde{y}_t \sim \tilde{p}_t}\sqrt{\frac{\hat{p}[\tilde{y}_t]}{f^*(\mathbf{x}_t)[\tilde{y}_t]}} \leq \sum_{\tilde{y}_t \in \tilde{\mathcal{Y}}} \sqrt{\hat{p}[\tilde{y}_t] f^*(\mathbf{x}_t)[\tilde{y}_t]},$$

where the inequality follows from Lemma 10.16. By a similar argument as in the proof of [84, Lemma A.14], we have:

$$\log \sum_{\tilde{y}_t \in \tilde{\mathcal{Y}}} \sqrt{\hat{p}[\tilde{y}_t] f^*(\mathbf{x}_t)[\tilde{y}_t]} = \log\left(1 - \frac{1}{2}H^2(\hat{p}_t, f^*(\mathbf{x}_t))\right) \leq -\frac{1}{2}H^2(\hat{p}_t, f^*(\mathbf{x}_t)),$$

where the first equality follows by definition of squared Hellinger divergence. Taking $X_t = \log\frac{f^*(\mathbf{x}_t)[\tilde{y}_t]}{\hat{p}_t[\tilde{y}_t]}$, $\alpha = \frac{1}{2}$ and invoking Lemma 2.9 we have w.p. $\geq 1 - \delta$

$$\Pr\left[\sum_{t=1}^{T} H^2(\hat{p}_t, f^*(\mathbf{x}_t)) \leq \log|\mathcal{F}| + 2\log(1/\delta)\right] \geq 1 - \delta.$$

Let now $\hat{y}_t = \arg\max_{\tilde{y}}\{\hat{p}_t[\tilde{y}] : \tilde{y} \in \tilde{\mathcal{Y}}\}$. We have, if $\hat{y}_t \neq h^*(\mathbf{x}_t)$

$$H^2(\hat{p}_t, f^*(\mathbf{x}_t)) \geq ||\hat{p}_t - f(\mathbf{x}_t)||_1^2/4 \geq (1-\eta)^2/4,$$

where the first inequality follows from $\sqrt{H^2(p,q)} \geq ||p - q||_1/2$ [20, Equation 7.20] and the second inequality follows from the proof of Lemma 10.13. Since $H^2(p,q) \geq 0$ for all p, q, we have w.p. $\geq 1 - \delta$ that:

$$\sum_{t=1}^{T} 1\{\hat{y}_t \neq h^*(\mathbf{x}_t)\} \leq \frac{\log|\mathcal{H}| + 2\log(1/\delta)}{(1-\eta)^2/4}.$$

To prove the second part, we observe that if $\hat{y}_t \neq h^*(\mathbf{x}_t)$, then $\hat{p}_t = (1-\eta)a_t + \eta u$ such that $a_t[h^*(\mathbf{x}_t)] \leq \frac{1}{2}$. Since $f^*(\mathbf{x}_t) = (1-\eta)e_{h^*(\mathbf{x}_t)} + \eta u$, we have by direct computation that:

$$H^2(\hat{p}_t, f^*(\mathbf{x}_t)) \geq \left(\sqrt{(1-\eta)/2 + \frac{\eta}{M}} - \sqrt{1-\eta+\frac{\eta}{M}}\right)^2 \sim \frac{M(1-\eta)^2}{16},$$

where the last asymptote follows by Taylor expansion

$$\frac{M(\eta-1)^2}{16} + O(\sum_{n=3}^{\infty} M^{n-1}(1-\eta)^n)$$

and the remainder term converges when $1-\eta \ll \frac{1}{M}$. □

Remark 10.6. Note that, Lemma 10.16 is the key that allows us to reduce our *mis-specified* setting to the well-specified case, such as [84, Lemma A.14], for which a reduction to the Hellinger divergence is possible.

10.6 Extensions for Stochastically Generated Features

We have demonstrated in previous sections that the minimax risk of our robust online classification problem can be effectively bounded for a finite hypothesis class $\mathcal{H}$ and adversarially generated features $\mathbf{x}^T$. We now demonstrate how this result can be generalized to infinite classes and general *stochastic* feature generating processes via the notion of stochastic sequential covering introduced in Chapter 7.2.

Infinite Classes. The following result that reduces the minimax risk of an infinite class to the size of the stochastic sequential cover.

Theorem 10.18. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be any hypothesis class, P be any class of random processes over $\mathcal{X}^T$ and $\mathcal{K}$ be a noise kernel that is well-separated w.r.t. Hellinger divergence at scale γ_{H}. If there exists a finite stochastic sequential cover $\mathcal{G} \subset \mathcal{Y}^{\mathcal{X}^*}$ of $\mathcal{H}$ w.r.t. P at scale 0 and confidence $\delta/2 > 0$, then there exists a predictor such that for all $\boldsymbol{\nu}^T \in \mathsf{P}$, if $\mathbf{x}^T \sim \boldsymbol{\nu}^T$ then w.p. $\geq 1-\delta$ over all randomness involved, the risk is upper bounded by:

$$O\left(\frac{\log(|\mathcal{G}|)\log(4|\mathcal{G}|/\delta)}{\gamma_{\mathrm{H}}}\right).$$

Proof. Let A be the event over $\mathbf{x}^T$ so that $\forall h \in \mathcal{H}$, $\exists g \in \mathcal{G}$ such that $\forall t \in [T],\ h(\mathbf{x}_t) = g(\mathbf{x}^t)$. Let now $\boldsymbol{\nu}^T \in \mathsf{P}$ be the underlying true feature generating process. We have by the definition of stochastic sequential covering that $\Pr_{\mathbf{x}^T}[A] \geq 1 - \delta/2$. We now observe that Theorem 10.7 holds for sequential functions as well. Therefore, taking confidence parameter $\delta/2$, the prediction rule derived from Theorem 10.7 w.r.t. class $\mathcal{G}$ yields high probability minimax risk upper bounded by:

$$O\left(\frac{\log(|\mathcal{G}|)\log(4|\mathcal{G}|/\delta)}{\gamma_{\mathbf{H}}}\right). \tag{10.24}$$

Let $h^* \in \mathcal{H}$ be the underlying true function, $\mathbf{x}^T \in A$ be any realization of the feature, and g^* be the sequential covering function of h^* at scale 0. Note that, g^* has the same labeling as h^* on $\mathbf{x}^T$. Therefore, any predictor has the same behaviours when running on h^* and g^*, and thus the high probability minimax risk for $\mathcal{H}$ is upper bounded by that of $\mathcal{G}$. The theorem now follows by a union bound. □

Note that, any bounds that we have established in the previous sections for finite class can be extended to infinite classes; these bounds depend only on the stochastic sequential cover size using a similar argument as Theorem 10.18. We will not discuss all such cases in the interest of clarity of presentation. As a demonstration, we establish the following concrete minimax risk bounds:

Corollary 10.19. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a class with finite Littlestone dimension $\mathsf{Ldim}(\mathcal{H})$ [85] and $|\mathcal{Y}| = N$. If the features are generated adversarially, and $\mathcal{K}$ is any noise kernel that is well-separated w.r.t. Hellinger divergence at scale $\gamma_{\mathbf{H}}$. Then, the high probability minimax risk at confidence δ is upper bounded by:

$$B^\delta(\mathcal{H}, \mathcal{K}) \leq O\left(\frac{\mathsf{Ldim}(\mathcal{H})^2 \log^2(TN) + \mathsf{Ldim}(\mathcal{H})\log(4TN/\delta)}{\gamma_{\mathbf{H}}}\right).$$

Moreover, for the noise kernel $\mathcal{K}^\eta$ as in Theorem 10.12, the high probability minimax risk with confidence $\delta > 0$ is upper bounded by:

$$B^\delta(\mathcal{H}, \mathcal{K}^\eta) \leq \frac{(\mathsf{Ldim}(\mathcal{H}) + 1)\log(TN) + 2\log(1/\delta)}{(1-\eta)^2/4}.$$

Proof. The first part follows directly from Theorem 10.18 and the fact that the sequential covering of $\mathcal{H}$ w.r.t. adversarial selection of $\mathcal{X}^T$ is of order $(TN)^{\mathsf{Ldim}(\mathcal{H})+1}$ by [85, Theorem 25]. The second part follows by Theorem 10.12. □

We complement this corollary with the following lower bound:

Proposition 10.5. For any $d, N \in \mathbb{N}$ and $\gamma_{\mathbf{H}} > 0$, there exists a class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ with $\mathsf{Ldim}(\mathcal{H}) \leq d$ and $|\mathcal{Y}| = N$, and a kernel $\mathcal{K}$ with Hellinger gap $\Omega(\gamma_{\mathbf{H}})$, such that:

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}) \geq \Omega\left(\frac{d \log N}{\gamma_{\mathbf{H}}}\right).$$

Proof. We define $\mathcal{Y} := [N]$, $\mathcal{X} := \{\mathbf{x}_1, \cdots, \mathbf{x}_d\}$, and $\mathcal{H} := \mathcal{Y}^{\mathcal{X}}$. It is easy to verify that $\mathsf{Ldim}(\mathcal{H}) = d$. Let $M = c \log N$ and $\mathcal{C} \subset \{-1, +1\}^{2M}$ be a maximum packing such that $\forall \mathbf{v}_1 \neq \mathbf{v}_2 \in \mathcal{C}, \mathsf{Ham}(\mathbf{v}_1, \mathbf{v}_2) \geq \frac{M}{2}$ and $\mathsf{Ham}(\mathbf{v}_i, \mathbf{1}) = M$, where Ham denotes the Hamming distance and $\mathbf{1}$ is the all-1 vector. By [86, Thm D.1], we have $|\mathcal{C}| \geq N$ for an appropriately selected constant c. Therefore, for any $y \in \mathcal{Y}$, we can identify a unique $\mathbf{v}_y \in \mathcal{C}$. We now define, for any $y \in \mathcal{Y}$, the distribution p_y over $\tilde{\mathcal{Y}} := [2M]$ such that

$$\forall \tilde{y} \in \tilde{\mathcal{Y}},\ p_y[\tilde{y}] = \frac{1 + \mathbf{v}_y[\tilde{y}]\epsilon}{2M},$$

where $\epsilon > 0$ is a small parameter to be selected. It is easy to verify that p_y is indeed a probability distribution. Moreover, for all $y_1 \neq y_2 \in \mathcal{Y}$, we have $\mathsf{KL}(p_{y_1}, p_{y_2}) \leq O(\epsilon^2)$ and $H^2(p_{y_1}, p_{y_2}) \geq \Omega(\epsilon^2)$. The first inequality follows from simple approximation, and the second inequality follows from the packing property of $\mathcal{C}$. We now take $\epsilon^2 = \gamma_{\mathbf{H}}$ and define the kernel $\mathcal{K}(\mathbf{x}, y) := \{p_y\}$. To prove the risk lower bound, we partition the time horizon into d blocks, each of size T/d, with the ith block taking feature $\mathbf{x}_i$. By Fano's inequality (cf. Chapter 2.4) and a similar argument as in Theorem 10.7, we have that the expected risk is lower bounded by $\Omega\left(\frac{d \log N}{\gamma_{\mathbf{H}}}\right)$. □

Remark 10.7. Note that in Proposition 10.5, we have a $\log N$ dependency on the label set size. This contrasts with the (agnostic) noiseless case [87], where the regret is independent of the label set size N.

$\boldsymbol{\sigma}$-Smoothed Processes. Finally, we apply our results for a large class of distributions over $\mathcal{X}^T$ known as σ-smoothed processes discussed in Chapter 8. To briefly recall for any given distribution μ over $\mathcal{X}$, we say a distribution ν over $\mathcal{X}$ is σ-smooth w.r.t. μ if for all measurable sets $A \subset \mathcal{X}$, we have $\nu(A) \leq \mu(A)/\sigma$. A random process $\boldsymbol{\nu}^T$ over $\mathcal{X}^T$ is said to be σ-smooth if the *conditional marginal* $\boldsymbol{\nu}^T(\cdot \mid X^{t-1})$ is σ-smooth w.r.t. μ for all $t \leq T$, almost surely. For instance, if $\sigma = 1$, we reduce to the *i.i.d.* process case.

Corollary 10.20. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a class with finite VC-dimension $\mathsf{VC}(\mathcal{H})$ and $|\mathcal{Y}| = 2$, $\mathsf{S}^{\sigma}(\mu)$ be the class of all σ-smoothed processes w.r.t. μ, and $\mathcal{K}^{\eta}$ be the noise kernel as in Theorem 10.12. Then for any $\boldsymbol{\nu}^T \in \mathsf{S}^{\sigma}(\mu)$, if $\mathbf{x}^T \sim \boldsymbol{\nu}^T$, then the *high probability*, minimax risk at confidence $\delta > 0$ is upper bounded by:

$$O\left(\frac{\mathsf{VC}(\mathcal{H})\log(T/\sigma) + \log(1/\delta)}{(1-\eta)^2}\right).$$

Proof. By Proposition 8.4, $\mathcal{H}$ admits a stochastic sequential cover $\mathcal{G}$ at confidence $\delta/2 > 0$ such that:

$$\log|\mathcal{G}| \leq O(\mathsf{VC}(\mathcal{H})\log(T/\sigma) + \log(1/\delta)).$$

We now condition on the event of the exact covering. By Theorem 10.12 (second part), the high probability minimax risk at confidence $\delta/2$ is upper bounded by:

$$O\left(\frac{\log|G| + \log(2/\delta)}{(1-\eta)^2}\right) \leq O\left(\frac{\mathsf{VC}(\mathcal{H})\log(T/\sigma) + \log(1/\delta)}{(1-\eta)^2}\right).$$

The result now follows by a union bound. □

10.7 Bibliographical Notes

Online learning with noisy labels was first studied in [88], which considers sequential prediction with binary outcomes corrupted by a Binary Symmetric Channel (BSC). The BSC can be viewed as a special case of the noise kernel in Example 10.1 with *fixed* noise parameters η_t. The more challenging setting with *varying* noise parameters—corresponding to Massart's noise—was investigated by [24]. Online conditional probability estimation has been extensively studied; see, for example, [36], [44], [50], [51], [56], [66]. The approach of leveraging online conditional density estimation (cf. Theorem 10.1) is conceptually similar to that of [84], within the context of *online decision making*. Analogous ideas of pairwise comparisons have also appeared in the differential privacy literature—see, e.g., [89]—though only in the *batch* setting.

The general theoretical framework and results discussed in this chapter are based on the work of [90].

References

[1] L. D. Davisson, “Universal noiseless coding”, *IEEE Trans. Inf. Theory*, vol. IT-19, no. 6, Nov. 1973, pp. 783–795.

[2] M. Drmota and W. Szpankowski, “Precise minimax redundancy and regrets.”, *IEEE Trans. Inf. Theory*, vol. IT-50, 2004, pp. 2686–2707.

[3] A. Orlitsky and N. P. Santhanam, “Speaking of infinity”, *IEEE Trans. Inf. Theory*, vol. 50, no. 10, Oct. 2004, pp. 2215–2230.

[4] J. Rissanen, “Fisher information and stochastic complexity”, *IEEE Trans. Information Theory*, no. 42, 1996, pp. 40–47.

[5] G. I. Shamir, “On the MDL principle for i.i.d. sources with large alphabets”, *IEEE Trans. Inform. Theory*, vol. 52, no. 5, May 2006, pp. 1939–1955.

[6] W. Szpankowski, “On asymptotics of certain recurrences arising in universal coding”, *Problems of Information Transmission*, vol. 34, 1998, pp. 55–61.

[7] Q. Xie and A. Barron, “Minimax redundancy for the class of memoryless sources”, *IEEE Trans. Information Theory*, 1997, pp. 647–657.

[8] Q. Xie and A. Barron, “Asymptotic minimax regret for data compression, gambling, and prediction”, *IEEE Trans. Information Theory*, vol. 46, 2000, pp. 431–445.

[9] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning and Games.* Cambridge University Press, 2006.

[10] W. Szpankowski and M. Weinberger, “Minimax pointwise redundancy for memoryless models over large alphabets”, *IEEE Trans. Information Theory*, vol. 58, 2012, pp. 4094–4104.

[11] R. E. Krichevsky and V. K. Trofimov, “The performance of universal encoding”, *IEEE Trans. Inform. Theory*, vol. IT-27, no. 2, Mar. 1981, pp. 199–207.

[12] Y. M. Shtarkov, “Universal sequential coding of single messages”, *Problems of Information Transmission*, vol. 23, no. 3, Jul. 1987, pp. 3–17.

[13] A. Barron, J. Rissanen, and B. Yu, “The minimum description length principle in coding and modeling”, *IEEE Trans. Information Theory*, vol. 44, no. 6, Oct. 1998, pp. 2743–2760.

[14] A. Orlitsky, P. Santhanam, and J. Zhang, “Universal compression of memoryless sources over unknown alphabets”, *IEEE Trans. Information Theory*, no. 50, 2004, pp. 1469–1481.

[15] A. Orlitsky and P. Santhanam, “Speaking of infinity”, *IEEE Trans. Information Theory*, no. 50, 2004, pp. 2215–2230.

[16] G. I. Shamir, “Universal source coding for monotonic and fast decaying monotonic distributions”, *IEEE Trans. Inform. Theory*, vol. 59, no. 11, Nov. 2013, pp. 7194–7211.

[17] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[18] M. Mohri, A. Rostamizadeh, and A. Talwalker, *Foudation of Machine Learning.* press, 2018.

[19] E. Hazan *et al.*, “Introduction to online convex optimization”, *Foundations and Trends® in Optimization*, vol. 2, no. 3-4, 2016, pp. 157–325.

[20] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning.* Cambridge University Press, 2022.

[21] T. Zhang, *Mathematical analysis of machine learning algorithms.* Cambridge University Press, 2023.

[22] D. A. Freedman, "On tail probabilities for martingales", *The Annals of Probability*, vol. 3, no. 1, 1975, pp. 100–118. DOI: 10.1214/aop/1176996495. URL: https://doi.org/10.1214/aop/1176996495.

[23] G. Peshkir and A. Shiryaev, "The khintchine inequality and martingale expanding sphere of their action", *Russian Math. Surveys*, vol. 50, no. 5, 1995, pp. 849–904.

[24] S. Ben-David, D. Pál, and S. Shalev-Shwartz, "Agnostic online learning.", in *Conference on Learning Theory*, vol. 3, 2009.

[25] A. Rakhlin, K. Sridharan, and A. Tewari, "Online learning: Stochastic and constrained adversaries", *arXiv preprint arXiv:1104.5070*, 2011.

[26] S. Rakhlin, O. Shamir, and K. Sridharan, "Relax and randomize: From value to algorithms", *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[27] A. Rakhlin, K. Sridharan, and A. Tewari, "Online learning: Random averages, combinatorial parameters, and learnability", in *Advances in Neural Information Processing Systems*, 2010.

[28] Z. Liu, I. Attias, and D. Roy, "Sequential probability assignment with contexts: Minimax regret, contextual shtarkov sums, and contextual normalized maximum likelihood", *arXiv preprint arXiv:2410.03849v1*, 2024.

[29] J. Rissanen, "Modeling by the shortest data description", *Automatica*, vol. 14, 1978, pp. 465–471.

[30] J. Rissanen, "Universal coding, information, prediction, and estimation", *IEEE Trans. Inform. Theory*, vol. IT-30, no. 4, Jul. 1984, pp. 629–636.

[31] N. Cesa-Bianchi and O. Shamir, "Efficient transductive online learning via randomized rounding", *arXiv preprint arXiv:1106.2429*, 2011.

[32] P. Jacquet, G. I. Shamir, and W. Szpankowski, "Precise minimax regret for logistic regression with categorical feature values", in *PRML: ALT'21*, vol. 132, pp. 755–771, 2021.

[33] N. Alon, O. Ben-Eliezer, Y. Dagan, S. Moran, M. Naor, and E. Yogev, "Adversarial laws of large numbers and optimal regret in online classification", in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 447–455, 2021.

[34] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press, 2019.

[35] D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan, "Logistic regression: The importance of being improper", in *Conference on Learning Theory (COLT)*, 2018.

[36] B. Bilodeau, D. Foster, and D. Roy, "Tight bounds on minimax regret under logarithmic loss via self-concordance", in *International Conference on Machine Learning*, PMLR, pp. 919–929, 2020.

[37] G. I. Shamir, "Logistic regression regret: What's the catch?", in *Conference on Learning Theory*, PMLR, pp. 3296–3319, 2020.

[38] P. Jacquet, G. I. Shamir, and W. Szpankowski, "Precise minimax regret for logistic regression", in *IEEE International Symposium on Information Theory (ISIT)*, IEEE, pp. 444–449, 2022.

[39] E. J. McShane, "Extension of range of functions", *Bulletin of the American Mathematical Society*, vol. 40, no. 12, 1934, pp. 837–842.

[40] S. M. Kakade and A. Y. Ng, "Online bounds for bayesian algorithms", in *Advances in Neural Information Processing Systems (NeurIPS)*, L. K. Saul, Y. Weiss, and L. Bottou, Eds., MIT Press, 2005, pp. 641–648.

[41] K. Yamanishi, "Minimax relative loss analysis for sequential prediction algorithms using parametric hypotheses", in *Conference on Learning Theory (COLT)*, pp. 32–43, 1998.

[42] B. Clarke and A. Barron, "Jeffreys' prior is asymptotically least favorable under entropy of risk", *J. Statistical Planning and Inference*, 1994, pp. 453–471.

[43] A. Rakhlin, K. Sridharan, and A. Tewari, "Sequential complexities and uniform martingale laws of large numbers", *Probability Theory and Related Fields*, vol. 161, no. 1, 2015, pp. 111–153.

[44] A. Rakhlin and K. Sridharan, "Sequential probability assignment with binary alphabets and large classes of experts", *arXiv preprint arXiv:1501.07340*, 2015.

[45] L.-A. Gottlieb, A. Kontorovich, and R. Krauthgamer, "Efficient regression in metric spaces via approximate lipschitz extension", *IEEE Trans. on Information Theory*, vol. 63, no. 8, 2017, pp. 4838–4849.

[46] E. Hazan, T. Koren, and K. Y. Levy, "Logistic regression: Tight bounds for stochastic and online optimization", in *Conference on Learning Theory (COLT)*, pp. 197–209, MIT press, 2014.

[47] R. Jézéquel, P. Gaillard, and A. Rudi, "Mixability made efficient: Fast online multiclass logistic regression", *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[48] V. Vovk, "Competitive on-line statistics", *International Statistical Review*, vol. 69, no. 2, 2001, pp. 213–248.

[49] A. Rakhlin and K. Sridharan, "Online nonparametric regression with general los function", in *Conference on Learning Theory*, 2014.

[50] B. Bilodeau, D. J. Foster, and D. M. Roy, "Minimax rates for conditional density estimation via empirical entropy", *The Annals of Statistics*, vol. 51, no. 2, 2023, pp. 762–790.

[51] C. Wu, M. Heidari, A. Grama, and W. Szpankowski, "Precise regret bounds for log-loss via a truncated bayesian algorithm", in *Advances in Neural Information Processing Systems*, vol. 35, pp. 26 903–26 914, 2022.

[52] C. Wu, M. Heidari, A. Grama, and W. Szpankowski, "Regret bounds for log-loss via bayesian algorithms", *IEEE Trans. Inf. Theory*, vol. 6, 2023, pp. 5971–5989.

[53] A. Rakhlin, K. Sridharan, and A. Tewari, "Online learning via sequential complexities.", *J. Mach. Learn. Res.*, vol. 16, no. 1, 2015, pp. 155–186.

[54] N. Littlestone, "Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm", *Machine learning*, vol. 2, 1988, pp. 285–318.

[55] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm", *Information and computation*, vol. 108, no. 2, 1994, pp. 212–261.

[56] A. Bhatt and Y.-H. Kim, "Sequential prediction under log-loss with side information", in *Algorithmic Learning Theory*, PMLR, pp. 340–344, 2021.

[57] D. J. Aldous, "Exchangeability and related topics", in *École d'Été de Probabilités de Saint-Flour XIII—1983*, Springer, 1985, pp. 1–198.

[58] D. Haussler, N. Littlestone, and M. K. Warmuth, “Predicting $\{0, 1\}$-functions on randomly drawn points”, *Information and Computation*, vol. 115, no. 2, 1994, pp. 248–292.

[59] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world.* Springer Science & Business Media, 2005.

[60] A. Lazaric and R. Munos, “Hybrid stochastic-adversarial on-line learning”, in *Conference on Learning Theory*, 2009.

[61] S. Hanneke and L. Yang, “Minimax analysis of active learning.”, *J. Mach. Learn. Res.*, vol. 16, no. 12, 2015, pp. 3487–3602.

[62] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, “Scale-sensitive dimensions, uniform convergence, and learnability”, *Journal of the ACM (JACM)*, vol. 44, no. 4, 1997, pp. 615–631.

[63] A. Block, Y. Dagan, N. Golowich, and A. Rakhlin, “Smoothed online learning is as easy as statistical learning”, *arXiv preprint arXiv:2202.04690*, 2022.

[64] B. Doerr, “Probabilistic tools for the analysis of randomized optimization heuristics”, in *Theory of evolutionary computation*, Springer, 2020, pp. 1–87.

[65] D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan, “Logistic regression: The importance of being improper”, in *Conference On Learning Theory*, PMLR, pp. 167–208, 2018.

[66] C. Wu, M. Heidari, A. Grama, and W. Szpankowski, “Expected worst case regret via stochastic sequential covering”, *Transactions on Machine Learning Research*, 2023.

[67] N. Haghtalab, T. Roughgarden, and A. Shetty, “Smoothed analysis of online and differentially private learning”, *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9203–9215.

[68] N. Haghtalab, T. Roughgarden, and A. Shetty, “Smoothed analysis with adaptive adversaries”, in *IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, IEEE, pp. 942–953, 2022.

[69] D. Haussler, “Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension”, *Journal of Combinatorial Theory, Series A*, vol. 69, no. 2, 1995, pp. 217–232.

[70] M. Blanchard, “Agnostic smoothed online learning”, *arXiv preprint arXiv:2410.05124*, 2024.

[71] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

[72] C. Wu, A. Grama, and W. Szpankowski, “Online learning in dynamically changing environments”, in *Conference on Learning Theory*, PMLR 195, pp. 325–358, 2023.

[73] C. Villani, *Topics in optimal transportation*, vol. 58. American Mathematical Soc., 2021.

[74] N. Fournier and A. Guillin, “On the rate of convergence in wasserstein distance of the empirical measure”, *Probability theory and related fields*, vol. 162, no. 3-4, 2015, pp. 707–738.

[75] H. Ashtiani, S. Ben-David, N. Harvey, C. Liaw, A. Mehrabian, and Y. Plan, “Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes”, *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[76] S. Kakade and A. T. Kalai, “From batch to transductive online learning”, *Advances in Neural Information Processing Systems*, vol. 18, 2005.

[77] N. Cesa-Bianchi and O. Shamir, “Efficient transductive online learning via randomized rounding”, *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, 2013, pp. 177–194.

[78] A. Rakhlin and K. Sridharan, “Bistro: An efficient relaxation-based method for contextual bandits”, in *International Conference on Machine Learning*, ser. PMLR, vol. 48, pp. 1977–1985, PMLR, 20–22 Jun 2016.

[79] V. Syrgkanis, H. Luo, A. Krishnamurthy, and R. E. Schapire, “Improved regret bounds for oracle-based adversarial contextual bandits”, *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[80] K. Banihashem, M. Hajiaghayi, S. Shin, and M. Springer, “An improved relaxation for oracle-efficient adversarial contextual bandits”, in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[81] N. Haghtalab, Y. Han, A. Shetty, and K. Yang, “Oracle-efficient online learning for smoothed adversaries”, in *Advances in Neural Information Processing Systems*, 2022.

[82] C. Wu, J. Sima, and W. Szpankowski, “Oracle-efficient hybrid online learning with unknown distribution”, in *The Thirty Seventh Annual Conference on Learning Theory*, PMLR, pp. 4992–5018, 2024.

[83] C. Dwork, A. Roth, *et al.*, “The algorithmic foundations of differential privacy”, *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, 2014, pp. 211–407.

[84] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin, “The statistical complexity of interactive decision making”, *arXiv preprint arXiv:2112.13487*, 2021.

[85] A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz, “Multiclass learnability and the erm principle.”, *J. Mach. Learn. Res.*, vol. 16, no. 1, 2015, pp. 2377–2404.

[86] C. Wu, Y. Wang, A. Grama, and W. Szpankowski, “Learning functional distributions with private labels”, in *International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 202, pp. 37 728–37 744, PMLR, 23–29 Jul 2023.

[87] S. Hanneke, S. Moran, V. Raman, U. Subedi, and A. Tewari, “Multiclass online learning and uniform convergence”, in *The Thirty Sixth Annual Conference on Learning Theory*, PMLR, pp. 5682–5696, 2023.

[88] T. Weissman, N. Merhav, and A. Somekh-Baruch, “Twofold universal prediction schemes for achieving the finite-state predictability of a noisy individual binary sequence”, *IEEE Transactions on Information Theory*, vol. 47, no. 5, 2001, pp. 1849–1866.

[89] S. Gopi, G. Kamath, J. Kulkarni, A. Nikolov, Z. S. Wu, and H. Zhang, “Locally private hypothesis selection”, in *Conference on Learning Theory*, PMLR, pp. 1785–1816, 2020.

[90] C. Wu, A. Grama, and W. Szpankowski, “Information-theoretic limits of online classification with noisy labels”, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.