Tight Bounds on Minimax Regret under Logarithmic Loss via Self-Concordance

Blair Bilodeau [1 2 3] Dylan J. Foster [4] Daniel M. Roy [1 2 3]

Abstract

We consider the classical problem of sequential probability assignment under logarithmic loss while competing against an arbitrary, potentially nonparametric class of experts. We obtain tight bounds on the minimax regret via a new approach that exploits the self-concordance property of the logarithmic loss. We show that for any expert class with (sequential) metric entropy $\mathcal{O}(\gamma^{-p})$ at scale γ, the minimax regret is $\mathcal{O}(n^{\frac{p}{p+1}})$, and that this rate cannot be improved without additional assumptions on the expert class under consideration. As an application of our techniques, we resolve the minimax regret for nonparametric Lipschitz classes of experts.

1. Introduction

Sequential probability assignment is a classical problem that has been studied intensely throughout domains including portfolio optimization (Cover, 1991; Cover & Ordentlich, 1996; Cross & Barron, 2003), information theory (Rissanen, 1984; Merhav & Feder, 1998; Xie & Barron, 2000), and—more recently—adversarial machine learning (Goodfellow et al., 2014; Grnarova et al., 2018; Liang & Modiano, 2018). The goal is for a *player* to assign probabilities to an arbitrary, potentially adversarially generated sequence of outcomes, and to do so nearly as well as a benchmark class of *experts*. More formally, consider the following protocol: for rounds $t = 1, \ldots, n$, the player receives a *context* $x_t \in \mathcal{X}$, predicts a probability $\widehat{p}_t \in [0,1]$ (using only the context x_t), observes a binary outcome $y_t \in \{0,1\}$, and incurs the *logarithmic loss* ("log loss"), defined by

$$\ell(\widehat{p}_t, y_t) = -y_t \log(\widehat{p}_t) - (1-y_t)\log(1-\widehat{p}_t).$$

The log loss penalizes the player based on how much probability mass they place on the actual outcome. Without distributional assumptions, one cannot control the total incurred loss, and so it is standard to study the *regret*; that is, the difference between the player's total loss and the total loss of the single best predictor in a (potentially uncountable) reference class of experts. Writing the vector of player predictions as $\widehat{\boldsymbol{p}} = (\widehat{p}_1, \ldots, \widehat{p}_n)$, and likewise defining $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$, the player's regret with respect to a class of experts $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ is defined as

$$\mathcal{R}_n(\mathcal{F}; \widehat{\boldsymbol{p}}, \boldsymbol{x}, \boldsymbol{y}) = \sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t).$$

Compared to similar sequential prediction problems found throughout the literature on online learning (Cesa-Bianchi & Lugosi, 2006; Hazan, 2016; Lattimore & Szepesvári, 2020), the distinguishing feature of the sequential probability assignment problem is the log loss, which amounts to evaluating the log-likelihood of the observed outcome under the player's predicted distribution. Typical results in online learning assume the loss function to be convex and smooth or Lipschitz (e.g., absolute loss or square loss on bounded predictions) or at least bounded (e.g., classification loss), while the log loss may have unbounded values and unbounded gradient. Consequently, beyond simple classes of experts, naively applying the standard tools of online learning leads to loose guarantees; instead, we exploit refined properties of the log loss to obtain tight regret bounds for sequential probability assignment.

Minimax Regret. We investigate the fundamental limits for sequential probability assignment through the lens of minimax analysis. We focus on *minimax regret*, defined by

$$\mathcal{R}_n(\mathcal{F}) = \sup_{x_1}\inf_{\widehat{p}_1}\sup_{y_1}\cdots\sup_{x_n}\inf_{\widehat{p}_n}\sup_{y_n}\mathcal{R}_n(\mathcal{F}; \widehat{\boldsymbol{p}}, \boldsymbol{x}, \boldsymbol{y})\,, \quad (1)$$

where $x_t \in \mathcal{X}_t$ (defined formally in Section 2), $\widehat{p}_t \in [0,1]$, and $y_t \in \{0,1\}$ for all $t \in [n]$. The minimax regret expresses worst-case performance of the best player across all adaptively chosen data sequences. For simple (e.g., parametric) classes of experts, the minimax regret is well-understood, including exact constants (Rissanen, 1986;

[1]Statistical Sciences, University of Toronto [2]Vector Institute [3]Institute for Advanced Study [4]Massachusetts Institute of Technology. Correspondence to: Blair Bilodeau <blair.bilodeau@mail.utoronto.ca>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

1996; Shtar'kov, 1987; Freund, 2003). For rich classes of experts, however, tight guarantees are not known, and hence our aim in this paper is to answer: *How does the complexity of $\mathcal{F}$ shape the minimax regret?*

A standard object used to control the minimax regret in statistical learning and sequential prediction is the *covering number*, which is a measure of the complexity of an expert class $\mathcal{F}$. The covering number for $\mathcal{F}$ is the size of the smallest subset of $\mathcal{F}$ such that every element of $\mathcal{F}$ is *close* to an element of the subset, where close is defined for appropriate notions of scale and distance. Early covering-based bounds for sequential probability assignment (Opper & Haussler, 1999; Cesa-Bianchi & Lugosi, 1999) use coarse notions of distance, but these bounds become vacuous for sufficiently rich expert classes.

More recently, Rakhlin & Sridharan (2015) gave sharper guarantees that use a finer notion of cover, referred to as a *sequential cover* (see Definition 1), which has previously been shown to characterize the minimax regret for simpler online learning problems with Lipschitz losses (Rakhlin et al., 2015a). Unfortunately, to deal with the fact that log loss is non-Lipschitz, this result (and all prior work in this line) approximates regret by truncating the allowed probabilities away from 0 and 1. This approximation forces the gradient of log loss to be bounded, but leads to suboptimal bounds for rich expert classes. Hence, Rakhlin & Sridharan (2015) posed the problem of whether the minimax regret for sequential probability assignment can be characterized using only the notion of sequential covering. This question is natural, as the answer is affirmative for the absolute loss (Rakhlin et al., 2015a), square loss (Rakhlin & Sridharan, 2014), and other common Lipschitz losses such as the hinge.

1.1. Overview of Results

Our main result is to show that for experts classes $\mathcal{F}$ for which the *sequential entropy* (the log of the sequential covering number) at scale γ grows as γ^{-p}, we have

$$\mathcal{R}_n(\mathcal{F}) \leq \mathcal{O}(n^{\frac{p}{p+1}}).$$

This upper bound recovers the best-known rates for all values of p, and offers strict improvement whenever $p > 1$ (i.e., whenever the class $\mathcal{F}$ is sufficiently complex). We further show that for certain expert classes—in particular, nonparametric Lipschitz classes over $[0,1]^p$—this rate cannot be improved. As a consequence, we resolve the minimax regret for these classes.

An important implication of our results is that for general classes $\mathcal{F}$, the optimal rate for regret cannot be characterized purely in terms of sequential covering numbers; this follows by combining our improved upper and lower bounds with an earlier observation from Rakhlin & Sridharan (2015).

Our upper bounds are obtained through a new technique that exploits the curvature of log loss (specifically, the property of *self-concordance*) to bound the regret. This allows us to handle the non-Lipschitzness of the log loss directly without invoking the truncation and approximation arguments that lead to suboptimal regret in previous approaches.

1.2. Related Work

For finite expert classes, it is well-known that the minimax regret $\mathcal{R}_n(\mathcal{F})$ is of order $\log|\mathcal{F}|$ (Vovk, 1998). Sharp guarantees are also known for countable expert classes (Banerjee, 2006) and parametric classes (Rissanen, 1986; 1996; Shtar'kov, 1987; Xie & Barron, 2000; Freund, 2003; Miyaguchi & Yamanishi, 2019); see also Chapter 9 of Cesa-Bianchi & Lugosi (2006).

In this work, we focus on obtaining tight guarantees for rich, nonparametric classes of experts. Previous work in this direction has obtained bounds for large expert classes using various notions of complexity for the class. Opper & Haussler (1999); Cesa-Bianchi & Lugosi (1999) bound the minimax regret under log loss using covering numbers for the expert class defined with respect to the sup-norm over the context space; that is, $d_{\sup}(f,g) = \sup_{x\in\mathcal{X}}|f(x) - g(x)|$. Covering with respect to all the elements in the domain is rather restrictive, and there are many cases for which the sup-norm covering number is infinite even though the class is learnable, or where the sup-norm cover has undesirable dependence on the dimension of the context space.

Building on a line of work which characterizes minimax rates for Lipschitz losses (Rakhlin & Sridharan, 2014; Rakhlin et al., 2015a), Rakhlin & Sridharan (2015) gave improved upper bounds for sequential probability assignment based on *sequential covering numbers*, which require that covering elements are close only on finite sequences of contexts induced by binary trees. Sequential covering numbers can be much smaller than sup-norm covers. For example, infinite dimensional linear functionals do not admit a finite sup-norm cover, but Rakhlin & Sridharan (2015) show via sequential covering that they are learnable at a rate of $\tilde{\mathcal{O}}(n^{3/4})$. Moreover, Rakhlin & Sridharan (2015) show that sublinear regret is possible only for expert classes with bounded sequential covering numbers.

While the rates obtained by Rakhlin & Sridharan (2015) are nonvacuous for many expert classes, they have suboptimal order for even moderately complex classes. Indeed, in order to handle the unbounded gradient of log loss, Rakhlin & Sridharan (2015) rely on truncation of the probabilities allowed to be predicted: They restrict the probabilities to $[\delta, 1-\delta]$ for some $0 < \delta \leq 1/2$, and then bound the true minimax regret by the minimax regret subject to this restricted probability range, plus an error term of size $\mathcal{O}(n\delta)$. This strategy allows one to treat the log loss as uniformly bounded and $1/\delta$-Lipschitz, but leads to poor rates compared to more common Lipschitz/strongly convex losses such as the square loss. Subsequent work by Foster et al. (2018) gave improvements to this approach that exploit the *mixability* property of the log loss (Vovk, 1998). While their results lead to improved rates for classes of "moderate" complexity, they face similar suboptimality for high-complexity expert classes.

1.3. Organization

Section 3 presents our improved minimax upper bound for general, potentially nonparametric expert classes (Theorem 2). In Section 3.1, we instantiate this bound for concrete examples of expert classes, and present a lower bound that is tight for certain expert classes (Theorem 3). In Section 3.2 we give a detailed comparison between our rates and those of prior work as a function of the sequential entropy.

In Section 4, we prove our upper bound via a new approach based on self-concordance of the log loss. Section 5 proves our lower bound, completing our characterization of the minimax regret.

We conclude the paper with a short discussion in Section 6.

2. Preliminaries

Contexts. We allow for time-varying context sets. At time t, we take x_t to belong to a set $\mathcal{X}_t \subseteq \mathcal{X}$, whose value may depend on the *history*, defined by $h_{1:t-1} = (x_1, y_1, \ldots, x_{t-1}, y_{t-1})$, but not future observations. Formally, we have $\mathcal{X}_t : (\mathcal{X} \times \{0,1\})^{t-1} \to 2^{\mathcal{X}}$, so that $x_t \in \mathcal{X}_t(h_{1:t-1})$. An example is the observed outcomes up to a given round, given by $\mathcal{X}_t(h_{1:t-1}) = \{(y_{1:t-1})\}$, which covers the standard setting of Cesa-Bianchi & Lugosi (1999). Another example is time-independent information, for example, $\mathcal{X}_t(h_{1:t-1}) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq 1\}$, which can be viewed analogously to the covariates in a standard regression task.

Sequential Covers and Metric Entropy. Sequential covering numbers are defined using *binary trees* indexed by sequences of binary observations ("paths"). Formally, for a set $\mathcal{A}$, an $\mathcal{A}$-valued binary tree $\boldsymbol{a}$ of depth n is a sequence of mappings $a_t : \{0,1\}^{t-1} \to \mathcal{A}$ for $t \in [n]$.

For a sequence (path) $\boldsymbol{\varepsilon} \in \{0,1\}^n$ and a tree $\boldsymbol{a}$ of depth n, let $a_t(\boldsymbol{\varepsilon}) := a_t(\varepsilon_1, \ldots, \varepsilon_{t-1})$ for $t \in [n]$. Also, denote the sequence of values a tree $\boldsymbol{a}$ takes on a path $\boldsymbol{\varepsilon}$ by $\boldsymbol{a}(\boldsymbol{\varepsilon}) = (a_1(\boldsymbol{\varepsilon}), \ldots, a_n(\boldsymbol{\varepsilon}))$. For a function $f : \mathcal{A} \to \mathbb{R}$, let $f \circ \boldsymbol{a}$ denote the tree taking values $(f(a_1(\boldsymbol{\varepsilon})), \ldots, f(a_n(\boldsymbol{\varepsilon})))$ on the path $\boldsymbol{\varepsilon}$. We extend this notation for a set of functions $\mathcal{F}$ by defining $\mathcal{F} \circ \boldsymbol{a} = \{f \circ \boldsymbol{a} : f \in \mathcal{F}\}$. Further, we say an $\mathcal{X}$-valued binary tree $\boldsymbol{x}$ is *consistent* if for all rounds $t \in [n]$ and paths $\boldsymbol{y} \in \{0,1\}^n$, $x_t(\boldsymbol{y}) \in \mathcal{X}_t(h_{1:t-1})$. For the remainder of this paper we will only consider context trees $\boldsymbol{x}$ with this property.

The notion of trees allows us to formally define a *sequential cover*, which may be thought of as a generalization of the classical notion of empirical covering that encodes the dependency structure of the online game.

Definition 1 (Rakhlin et al. 2015a)**.** *Let $\mathcal{A}$ and V be collections of $\mathbb{R}$-valued binary trees of depth n. V is a sequential cover for $\mathcal{A}$ at scale γ if*

$$\max_{\boldsymbol{y} \in \{0,1\}^n} \sup_{\boldsymbol{a} \in \mathcal{A}} \inf_{\boldsymbol{v} \in V} \max_{t \in [n]} |a_t(\boldsymbol{y}) - v_t(\boldsymbol{y})| \leq \gamma.$$

Let $\mathcal{N}_\infty(\mathcal{A}, \gamma)$ be the size of the smallest such cover.[1]

For a function class $\mathcal{F}$, we define the *sequential entropy* of $\mathcal{F}$ at scale γ and depth n as the log of the worst-case sequential covering number:

$$\mathcal{H}_\infty(\mathcal{F}, \gamma, n) = \sup_{\boldsymbol{x}} \log \mathcal{N}_\infty(\mathcal{F} \circ \boldsymbol{x}, \gamma),$$

where the $\sup$ is taken over all context trees of depth n.

Sequential covering numbers incorporate the dependence structure of online learning, and consequently are never smaller than classical *empirical covers* found in statistical learning, which require that the covering elements are close only on a fixed sequence $x_{1:n}$. While the sequential covering number of $\mathcal{F} \circ \boldsymbol{x}$ for context trees of depth n will never be smaller than the empirical covering number for datasets of size n, it will—importantly—always be finite. Additionally, because the definition allows one to choose the covering element as a function of the path $\boldsymbol{y}$, the sequential covering number at depth n is typically much smaller than, for example, the empirical covering number of size 2^n, despite the context tree having $2^n - 1$ unique values.

Asymptotic Notation. We adopt standard big-oh notation. Consider two real-valued sequences (x_n) and (y_n). We write $x_n \leq \mathcal{O}(y_n)$ if there exists a constant $M > 0$ such that for all sufficiently large n, $|x_n| \leq M y_n$. Conversely, $x_n \geq \Omega(y_n)$ if $y_n \leq \mathcal{O}(x_n)$. We write $x_n \leq \tilde{\mathcal{O}}(y_n)$ if there is some $r > 0$ such that $x_n \leq \mathcal{O}(y_n(\log(n))^r)$. We also write $x_n = \Theta(y_n)$ if $\Omega(y_n) \leq x_n \leq \mathcal{O}(y_n)$, and similarly $x_n = \tilde{\Theta}(y_n)$ if $\Omega(y_n) \leq x_n \leq \tilde{\mathcal{O}}(y_n)$. Note that we do not specify a notion of $\tilde{\Omega}$. Instead, we say a sequence $x_n = \text{polylog}(y_n)$ if there exist some $0 < r < s$ such that $\Omega((\log(y_n))^r) \leq x_n \leq \mathcal{O}((\log(y_n))^s)$. Then, for any function g, $x_n \leq \mathcal{O}(g(\text{polylog}(y_n)))$ if there exists some sequence $y'_n = \text{polylog}(y_n)$ such that $x_n \leq \mathcal{O}(g(y'_n))$.

[1] The "∞" subscript reflects that the cover is defined with respect to the empirical L_∞ norm.

3. Minimax Regret Bounds

We now state our upper bound on the minimax regret for sequential probability assignment. Our result is nonconstructive; that is, we do not provide an explicit algorithm that achieves our upper bound. Rather, we characterize the fundamental limits of the learning problem for arbitrary expert classes, providing a benchmark for algorithm design going forward.

Theorem 2. *For any $\mathcal{X}$ and $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$,*

$$\mathcal{R}_n(\mathcal{F}) \leq \inf_{\gamma>0} \left\{ 4n\gamma + c\,\mathcal{H}_\infty(\mathcal{F},\gamma,n) \right\},$$

where $c = \frac{2-\log(2)}{\log(3)-\log(2)} \leq 4$.

For simple parametric classes where $\mathcal{H}_\infty(\mathcal{F},\gamma,n) = \Theta(d\log(1/\gamma))$, Theorem 2 recovers the usual fast rate of $\mathcal{O}(d\log(en/d))$. More interesting is the rich/high-complexity regime where $\mathcal{H}_\infty(\mathcal{F},\gamma,n) = \Theta(\gamma^{-p})$ for $p > 0$, for which Theorem 2 implies that

$$\mathcal{R}_n(\mathcal{F}) \leq \mathcal{O}(n^{\frac{p}{p+1}}). \tag{2}$$

As we discuss at length in Section 3.2, this rate improves over prior work for all $p > 1$. More importantly, this upper bound is tight for certain nonparametric classes (namely, the 1-Lipschitz experts). That is, if one wishes to bound regret only in terms of sequential entropy, Theorem 2 cannot be improved.

Theorem 3. *For any $p \in \mathbb{N}$, the class $\mathcal{F}$ of 1-Lipschitz (w.r.t. ℓ_∞) experts on $[0,1]^p$ satisfies $\mathcal{H}_\infty(\mathcal{F},\gamma,n) = \Theta(\gamma^{-p})$ and*

$$\mathcal{R}_n(\mathcal{F}) = \Theta(n^{\frac{p}{p+1}}).$$

While Theorem 3 shows that our new upper bound cannot be improved in a worst-case sense, there is still room for improvement for specific function classes of interest. Let $\mathbb{B}_2$ be the unit ball in a Hilbert space. Consider the class of infinite-dimensional linear predictors $\mathcal{F} = \{x \mapsto \frac{1}{2}[\langle w, x\rangle + 1] \mid w \in \mathbb{B}_2\}$, with $\mathcal{X} = \mathbb{B}_2$. This class has sequential entropy $\mathcal{H}_\infty(\mathcal{F},\gamma,n) = \tilde{\Theta}(\gamma^{-2})$, so $\mathcal{R}_n(\mathcal{F}) \leq \tilde{\mathcal{O}}(n^{2/3})$ by Theorem 2. However, Rakhlin & Sridharan (2015) describe an explicit algorithm that attains regret $\tilde{\mathcal{O}}(n^{1/2})$ for this class, meaning that our upper bound is loose for this example. Yet, since Theorem 3 shows that the upper bound cannot be improved without further assumptions, we draw the following conclusion.

Corollary 4. *The minimax rates for sequential probability assignment with the log loss cannot be characterized purely in terms of sequential entropy.*

We discuss a couple more features of Theorem 2 and Theorem 3 below.

- The proof strategy for Theorem 2 differs from previous approaches by discretizing $\mathcal{F}$ at a single scale rather than multiple scale levels (referred to as chaining). Surprisingly, this rather coarse approach achieves the previous best known results and improves on them for rich expert classes. Key to this improvement is the self-concordance of the log loss, which enables us to avoid truncation arguments.
- Theorem 3 in fact lower bounds the minimax regret when data is generated i.i.d. from a well-specified model, which implies that for Lipschitz classes, this apparently easier setting is in fact just as hard as the fully adversarial setting. This is in contrast to the case for square loss, where the rates for the i.i.d. well-specified and i.i.d. misspecified settings diverge once $p \geq 2$ (Rakhlin et al., 2017).

3.1. Further Examples

In order to place our new upper bound in the context of familiar expert classes, we walk through some additional examples below.

Example 1 (Sequential Rademacher Complexity)**.** *The sequential Rademacher complexity of an expert class $\mathcal{F}$ is given by*

$$\mathfrak{R}_n(\mathcal{F}) = \sup_{\boldsymbol{x}} \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \varepsilon_t\, f(x_t(\varepsilon)),$$

where $\sup_{\boldsymbol{x}}$ ranges over all $\mathcal{X}$-valued trees and $\varepsilon \in \{\pm 1\}^n$ are Rademacher random variables.[2] *Via Corollary 1 and Lemma 2 of Rakhlin et al. (2015b), we deduce that*

$$\mathcal{R}_n(\mathcal{F}) \leq \tilde{\mathcal{O}}\Big(\mathfrak{R}_n^{2/3}(\mathcal{F}) \cdot n^{1/3}\Big).$$

Example 2 (Smooth Nonparametric Classes)**.** *Let $\mathcal{F}$ be the class of all bounded functions over $[0,1]^d$ for which the first $k-1$ derivatives are Lipschitz. Then we may take $p = d/k$ (see, e.g., Example 5.11 of Wainwright, 2019), and hence Theorem 2 gives that $\mathcal{R}_n(\mathcal{F}) \leq \tilde{\mathcal{O}}(n^{\frac{d}{d+k}})$. One can show that this is optimal via a small modification to the proof of Theorem 3.*

Example 3 (Neural Networks)**.** *Rakhlin et al. (2015a) show that neural networks with Lipschitz activations and ℓ_1-bounded weights have $\mathfrak{R}_n(\mathcal{F}) \leq \tilde{\mathcal{O}}(\sqrt{n})$. We conclude from Example 1 that $\mathcal{R}_n(\mathcal{F}) \leq \tilde{\mathcal{O}}(n^{2/3})$ for these classes.*

3.2. Comparing to Previous Regret Bounds

We now compare the bound from Theorem 2 to the previous state of the art, Theorem 7 of Foster et al. (2018), which shows that for any $\mathcal{X}$ and $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$,

[2] Here we overload the definition of a tree in the natural way to allow arguments in $\{\pm 1\}$ rather than $\{0,1\}$.

$$\mathcal{R}_n(\mathcal{F}) \le \inf_{\substack{\gamma \ge \alpha > 0 \\ \delta > 0}} \left\{ \frac{4n\alpha}{\delta} + 30\sqrt{\frac{2n}{\delta}} \int_\alpha^\gamma \sqrt{\mathcal{H}_\infty(\mathcal{F}, \varepsilon, n)}\, \mathrm{d}\varepsilon \right. \\ + \frac{8}{\delta} \int_\alpha^\gamma \mathcal{H}_\infty(\mathcal{F}, \varepsilon, n)\, \mathrm{d}\varepsilon \\ \left. + \mathcal{H}_\infty(\mathcal{F}, \gamma, n) + 3n\delta \log(1/\delta) \right\}. \quad (3)$$

For any expert class $\mathcal{F}$ we will refer to the upper bound of Theorem 2 by $\mathrm{U}_n^{\mathrm{new}}(\mathcal{F})$ and the upper bound of Foster et al. (2018, Theorem 7) by $\mathrm{U}_n^{\mathrm{old}}(\mathcal{F})$. We observe the following relationship, proven in Appendix B.

Proposition 5. *For any $\mathcal{X}$ and $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$, the following hold:*

(i) *If* $\mathcal{H}_\infty(\mathcal{F}, \gamma, n) = \Theta(\log(1/\gamma))$,

$$\frac{\mathrm{U}_n^{\mathrm{new}}(\mathcal{F})}{\mathrm{U}_n^{\mathrm{old}}(\mathcal{F})} = \Theta(1).$$

(ii) *If* $\mathcal{H}_\infty(\mathcal{F}, \gamma, n) = \Theta(\gamma^{-p})$ *for $p \le 1$,*

$$\frac{\mathrm{U}_n^{\mathrm{new}}(\mathcal{F})}{\mathrm{U}_n^{\mathrm{old}}(\mathcal{F})} = \Theta\left(\frac{1}{\mathrm{polylog}(n)}\right).$$

(iii) *If* $\mathcal{H}_\infty(\mathcal{F}, \gamma, n) = \Theta(\gamma^{-p})$ *for $p > 1$,*

$$\frac{\mathrm{U}_n^{\mathrm{new}}(\mathcal{F})}{\mathrm{U}_n^{\mathrm{old}}(\mathcal{F})} = \Theta\left(\frac{1}{n^{\frac{p-1}{2p(p+1)}}\mathrm{polylog}(n)}\right).$$

4. Proof of Theorem 2

We now prove our main result. The proof has three parts. First, we use a minimax theorem to move to the dual of the online learning game, where we can evaluate the optimal strategy for the learner. This allows us to express the value of the minimax regret as a dependent empirical processes. For the next step, we move to a simpler, linearized upper bound on this process using the self-concordance property of the log loss, leading to a particular "offset" process. For finite classes, we can directly bound the value of the offset process by $\log|\mathcal{F}|$; the final bound in Theorem 2 follows by applying this result along with a discretization argument.

Before proceeding, we elaborate on the second point above. Let us take a step back and consider the simpler problem of bounding the minimax regret for square loss. Rakhlin & Sridharan (2014) show that via a similar minimax theorem, it is possible to bound the regret by a dependent random process called the *offset sequential Rademacher complexity*, which, informally, takes the form

$$\mathbb{E} \sup_{f \in \mathcal{F}} [X_{\mathrm{emp}}(f) - Y_{\mathrm{offset}}(f)]. \quad (4)$$

Here, $X_{\mathrm{emp}}(f)$ is a zero-mean Rademacher process indexed by $\mathcal{F}$ and $Y_{\mathrm{offset}}(f)$ is a quadratic offset. The offset component arises due to the strong convexity of the square loss, and penalizes large fluctuations in the Rademacher process, leading to fast rates.

For the log loss, the issue faced if one attempts to apply the same strong convexity-based argument is that the process $X_{\mathrm{emp}}(f)$, which involves the derivative of the loss, becomes unbounded as f approaches the boundary of $[0,1]$, and the quadratic offset $Y_{\mathrm{offset}}(f)$ does not grow fast enough to neutralize it. The simplest way to address this issue, and the one taken by Rakhlin & Sridharan (2015), is to truncate predictions. Our main insight is that using self-concordance of the log loss rather than strong convexity leads to an offset that can neutralize the derivative, removing the need for truncation and resulting in faster rates. The inspiration for using this property came from Rakhlin & Sridharan (2015, Section 6), who design a variant of mirror descent using a self-concordant barrier as a regularizer to obtain fast rates for linear prediction with the log loss, though our use of the property here is technically quite different.

4.1. Minimax Theorem and Dual Game

As our first step, we move to the *dual game* in which the order of max and min at each time step is swapped. Moving to the dual game is a now-standard strategy (Abernethy et al., 2009; Rakhlin & Sridharan, 2014; Rakhlin et al., 2015a; Rakhlin & Sridharan, 2015; Foster et al., 2018), and is a useful tool for analysis because the optimal strategy for the learner is much more tractable to compute in the dual.

In particular, for our sequential probability assignment setting, the following minimax theorem (Appendix A.1) holds.

Lemma 6. *For any $\mathcal{X}$ and $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$,*

$$\mathcal{R}_n(\mathcal{F}) = \sup_{x_1} \sup_{p_1 \in [0,1]} \mathop{\mathbb{E}}_{y_1 \sim p_1} \cdots \sup_{x_n} \sup_{p_n \in [0,1]} \mathop{\mathbb{E}}_{y_n \sim p_n} \sup_{f \in \mathcal{F}} \\ \sum_{t=1}^n \left\{ \inf_{\widehat{p}_t \in [0,1]} \mathop{\mathbb{E}}_{y_t \sim p_t} [\ell(\widehat{p}_t, y_t)] - \ell(f(x_t), y_t) \right\}.$$

The parameter $p_t \in [0,1]$ represents a distribution over the adversary's outcome $y_t \in \{0,1\}$, which the player can observe before they select $\widehat{p}_t$. For log loss, it is easy to see that the infimum of the interior expectation in Lemma 6 is achieved at $\widehat{p}_t = p_t$, so by the linearity of expectation the minimax regret can be written as

$$\sup_{x_1} \sup_{p_1 \in [0,1]} \mathop{\mathbb{E}}_{y_1 \sim p_1} \cdots \sup_{x_n} \sup_{p_n \in [0,1]} \mathop{\mathbb{E}}_{y_n \sim p_n} \mathcal{R}_n(\mathcal{F}; \boldsymbol{p}, \boldsymbol{x}, \boldsymbol{y}).$$

We simplify this statement using the tree notation from Section 2. In particular, writing $\mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{p}}$ to denote the nested conditional expectations $\mathbb{E}_{y_t \sim p_t(\boldsymbol{y})}$ for each $t \in [n]$, we can write the minimax regret as

$$\mathcal{R}_n(\mathcal{F}) = \sup_{\boldsymbol{x},\boldsymbol{p}} \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}} \mathcal{R}_n(\mathcal{F}; \boldsymbol{p}(\boldsymbol{y}), \boldsymbol{x}(\boldsymbol{y}), \boldsymbol{y}),$$

where $\boldsymbol{x}$ and $\boldsymbol{p}$ are respectively $\mathcal{X}$- and $[0,1]$-valued binary trees of depth n. We now fix an arbitrary context tree $\boldsymbol{x}$ and probability tree $\boldsymbol{p}$, and show that the bound of Theorem 2 holds for $\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}} \mathcal{R}_n(\mathcal{F}; \boldsymbol{p}(\boldsymbol{y}), \boldsymbol{x}(\boldsymbol{y}), \boldsymbol{y})$. Recall that there is a $\sup_{f\in\mathcal{F}}$ inside $\mathcal{R}_n(\mathcal{F}; \boldsymbol{p}(\boldsymbol{y}), \boldsymbol{x}(\boldsymbol{y}), \boldsymbol{y})$, so we must control the expected supremum of a dependent empirical process.

4.2. Self-Concordance and Offset Process

As sketched earlier, the key step in our proof is to upper bound $\mathcal{R}_n(\mathcal{F}; \boldsymbol{p}(\boldsymbol{y}), \boldsymbol{x}(\boldsymbol{y}), \boldsymbol{y})$ in terms of a new type of offset process using self-concordance. Let us first introduce the property formally.

Definition 7. *A function $F : \mathbb{R}^d \to \mathbb{R}$ is self-concordant on $S \subseteq \mathbb{R}^d$ if for all $s \in \operatorname{interior}(S)$ and $h \in \mathbb{R}^d$,*

$$\frac{d}{d\alpha}\nabla^2 F(s+\alpha h)\Big|_{\alpha=0} \preccurlyeq 2\nabla^2 F(s)\sqrt{h^\top \nabla^2 F(s) h}.$$

If $F : \mathbb{R} \to \mathbb{R}$, this can be written as

$$|F'''(s)| \leq 2F''(s)^{3/2}.$$

The class of self-concordant functions was first introduced by Nesterov & Nemirovski (1994) to study interior point methods. The logarithm is in fact the defining self-concordant function, satisfying equality in Definition 7. Consequently, we are able to apply the following result about self-concordance to log loss (viewed as a function of the predictions).

Lemma 8 (Nesterov 2004, Theorem 4.1.7)**.** *If $F : S \to \mathbb{R}$ is self-concordant on a convex set S, then for all $s, t \in \operatorname{interior}(S)$,*

$$F(t) \geq F(s) + \langle \nabla F(s), t-s\rangle + w\left(\|t-s\|_{F,s}\right),$$

where $w(z) = z - \log(1+z)$ and $\|h\|_{F,s} = \sqrt{h^\top \nabla^2 F(s) h}$ is the local norm with respect to F.

We use Lemma 8 to linearize the log loss, leading to a decomposition similar to (4); note that we only require the scalar version of the lemma. This decomposition allows us to exploit the fact that, while both the logarithm's value and its derivative tend to infinity near the boundary, the value does so at a much slower rate.

Lemma 9. *Let $\eta(p,y) = \frac{d}{dp}\ell(p,y)$ for $p \in [0,1]$ and $y \in \{0,1\}$, and define $\varphi(z) = z - |z| + \log(1+|z|)$. Then, $\mathcal{R}_n(\mathcal{F}; \boldsymbol{p}(\boldsymbol{y}), \boldsymbol{x}(\boldsymbol{y}), \boldsymbol{y})$ is bounded above almost surely by*

$$\sup_{f\in\mathcal{F}} \sum_{t=1}^n \varphi\Big(\eta(p_t(\boldsymbol{y}), y_t)[p_t(\boldsymbol{y}) - f(x_t(\boldsymbol{y}))]\Big)$$

under $\boldsymbol{y} \sim \boldsymbol{p}$.

In the language of (4), we can interpret the linear term z in $\varphi(z) = z - (|z| - \log(1+|z|))$ as giving rise to a mean-zero process, while the term $-(|z| - \log(1+|z|))$ is a (negative) offset that behaves like a quadratic for small values of z and like the absolute value for large values.

Proof. Taking derivatives of $\ell(p,y)$ with respect to p,

$$\begin{aligned}
\ell'(p,y) &= \frac{-y}{p} + \frac{1-y}{1-p}, \\
\ell''(p,y) &= \frac{y}{p^2} + \frac{1-y}{(1-p)^2}, \quad \text{and} \\
\ell'''(p,y) &= \frac{-2y}{p^3} + \frac{2(1-y)}{(1-p)^3}.
\end{aligned}$$

Since $y \in \{0,1\}$, $|\ell'''(p,y)| = 2\ell''(p,y)^{3/2}$, so the log loss is indeed self-concordant in p on $(0,1)$. Now, fix $y \in \{0,1\}$ and $t \in [n]$, and consider $F(a) = \ell(a,y)$. We can then apply Lemma 8 to F evaluated at $p = p_t(\boldsymbol{y}) \in (0,1)$ and $f = f(x_t(\boldsymbol{y})) \in (0,1)$. This gives

$$F(p) - F(f) \leq (p-f)F'(p) - w(\|p-f\|_{F,p}). \tag{5}$$

By definition, $(p-f)F'(p) = (p-f)\eta(p,y)$. Further,

$$\|p-f\|_{F,p} = \sqrt{(p-f)^2 F''(p)}.$$

Finally, since $y \in \{0,1\}$, $\ell''(p,y) = \eta(p,y)^2$, so

$$\|p-f\|_{F,p} = |(p-f)\eta(p,y)|.$$

Applying the definition of $w(z)$ gives the result on $(0,1)$. For the boundary points $p \in \{0,1\}$ and $f \in \{0,1\}$, it is easy to check the inequality holds by observing that $p=0$ implies $y=0$ a.s. and $p=1$ implies $y=1$ a.s.; we complete this calculation in Lemma 16. ■

4.3. Applying Sequential Covering

We now follow the standard strategy of covering the expert class $\mathcal{F}$, bounding the supremum for the cover, and then paying a penalty for approximation.

Consider the class of trees $\mathcal{F}_{\boldsymbol{p},\boldsymbol{x}} = \{\boldsymbol{p} - (f\circ\boldsymbol{x}) : f \in \mathcal{F}\}$. Our goal is to obtain a bound in terms of the sequential entropy of this class, which we observe is the same as the sequential entropy of $\mathcal{F}\circ\boldsymbol{x}$. Fix some $\gamma > 0$, and let $V_{\boldsymbol{p},\boldsymbol{x}}$ be a sequential cover of $\mathcal{F}_{\boldsymbol{p},\boldsymbol{x}}$ at scale γ. Then, by adding and subtracting terms after applying Lemma 9,

$\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}\,\mathcal{R}_n(\mathcal{F};\boldsymbol{p}(\boldsymbol{y}),\boldsymbol{x}(\boldsymbol{y}),\boldsymbol{y})$ is bounded above by

$$\begin{aligned}&\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}}\sup_{\boldsymbol{g}\in\mathcal{F}_{\boldsymbol{p},\boldsymbol{x}}}\min_{\boldsymbol{v}\in V_{\boldsymbol{p},\boldsymbol{x}}}\sum_{t=1}^{n}\\ &\qquad\left\{\varphi\Big(\eta(p_t(\boldsymbol{y}),y_t)g_t(\boldsymbol{y})\Big)-\varphi\Big(\eta(p_t(\boldsymbol{y}),y_t)v_t(\boldsymbol{y})\Big)\right\} \quad (6)\\ &\quad+\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}}\max_{\boldsymbol{v}\in V_{\boldsymbol{p},\boldsymbol{x}}}\sum_{t=1}^{n}\varphi\Big(\eta(p_t(\boldsymbol{y}),y_t)v_t(\boldsymbol{y})\Big). \quad (7)\end{aligned}$$

We have now reduced the problem to controlling the approximation error (6) and the finite class process (7). Controlling the approximation error is handled by the following property of the function φ, which we prove in Appendix A.3.

Lemma 10. *For any $s,t\in\mathbb{R}$, $\varphi(s)-\varphi(t)\le 2|s-t|$.*

Applying Lemma 10, the approximation error term (6) is bounded above by

$$\begin{aligned}&2\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}}\sup_{\boldsymbol{g}\in\mathcal{F}_{\boldsymbol{p},\boldsymbol{x}}}\min_{\boldsymbol{v}\in V_{\boldsymbol{p},\boldsymbol{x}}}\sum_{t=1}^{n}\Big|\eta(p_t(\boldsymbol{y}),y_t)[g_t(\boldsymbol{y})-v_t(\boldsymbol{y})]\Big|\\ &\quad\le 2\gamma\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}}\sum_{t=1}^{n}\Big|\eta(p_t(\boldsymbol{y}),y_t)\Big|, \quad (8)\end{aligned}$$

where we have used the fact that $V_{\boldsymbol{p},\boldsymbol{x}}$ is a sequential cover of $\mathcal{F}_{\boldsymbol{p},\boldsymbol{x}}$ at scale γ.

For any particular realization of $\boldsymbol{y}$, the value of $\eta(p_t(\boldsymbol{y}),y_t)$ in (8) depends inversely on p_t and $1-p_t$, and so can be arbitrarily large. Luckily, we recognize that the large values of η are exactly controlled by the small probability of paths that generate them. That is, adopting the shorthand $\boldsymbol{y}_t=y_{1:t}$,

$$\begin{aligned}&\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}}\sum_{t=1}^{n}\Big|\eta(p_t(\boldsymbol{y}),y_t)\Big|\\ &\quad=\mathop{\mathbb{E}}_{\boldsymbol{y}_{n-1}\sim\boldsymbol{p}}\mathop{\mathbb{E}}_{y_n\sim p_n(\boldsymbol{y})}\left[\sum_{t=1}^{n}\left(\frac{y_t}{p_t(\boldsymbol{y})}+\frac{1-y_t}{1-p_t(\boldsymbol{y})}\right)\right]\\ &\quad=\mathop{\mathbb{E}}_{\boldsymbol{y}_{n-1}\sim\boldsymbol{p}}\left[\sum_{t=1}^{n-1}\left(\frac{y_t}{p_t(\boldsymbol{y})}+\frac{1-y_t}{1-p_t(\boldsymbol{y})}\right)\right.\\ &\qquad\qquad\left.+\mathop{\mathbb{E}}_{y_n\sim p_n(\boldsymbol{y})}\left[\left(\frac{y_n}{p_n(\boldsymbol{y})}+\frac{1-y_n}{1-p_n(\boldsymbol{y})}\right)\right]\right]\\ &\quad=\mathop{\mathbb{E}}_{\boldsymbol{y}_{n-1}\sim\boldsymbol{p}}\sum_{t=1}^{n-1}\Big|\eta(p_t(\boldsymbol{y}),y_t)\Big|+2.\end{aligned}$$

Iterating this argument down to $t=1$ gives

$$\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}}\sum_{t=1}^{n}\Big|\eta(p_t(\boldsymbol{y}),y_t)\Big|=2n. \quad (9)$$

It remains to control the value of the finite-class process in (7). For this we use the offset property, and again exploit the fact that the η term only takes large values on paths with low probability.

For a $[0,1]$-valued tree $\boldsymbol{p}$, we say that a $[-1,1]$-valued tree $\boldsymbol{v}$ is a $[\boldsymbol{p}-1,\boldsymbol{p}]$-valued tree if for all $t\in[n]$ and $\boldsymbol{y}\in\{0,1\}^n$, $v_t(\boldsymbol{y})\in[p_t(\boldsymbol{y})-1,p_t(\boldsymbol{y})]$. We have the following bound.

Lemma 11. *Consider a $[0,1]$-valued binary tree $\boldsymbol{p}$ and a finite class V of $[\boldsymbol{p}-1,\boldsymbol{p}]$-valued trees. Then*

$$\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}}\max_{\boldsymbol{v}\in V}\sum_{t=1}^{n}\varphi\Big(\eta(p_t(\boldsymbol{y}),y_t)v_t(\boldsymbol{y})\Big)\le c\log|V|,$$

where $c=\frac{2-\log(2)}{\log(3)-\log(2)}$.

Proof. First, for all $\lambda>0$, we have

$$\begin{aligned}&\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}}\max_{\boldsymbol{v}\in V}\sum_{t=1}^{n}\varphi\Big(\eta(p_t(\boldsymbol{y}),y_t)v_t(\boldsymbol{y})\Big)\\ &=\log\left(\exp\left\{\lambda\frac{1}{\lambda}\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}}\max_{\boldsymbol{v}\in V}\sum_{t=1}^{n}\varphi\Big(\eta(p_t(\boldsymbol{y}),y_t)v_t(\boldsymbol{y})\Big)\right\}\right)\\ &\le\frac{1}{\lambda}\log\left(\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}}\exp\left\{\lambda\max_{\boldsymbol{v}\in V}\sum_{t=1}^{n}\varphi\Big(\eta(p_t(\boldsymbol{y}),y_t)v_t(\boldsymbol{y})\Big)\right\}\right)\\ &\le\frac{1}{\lambda}\log\left(\sum_{\boldsymbol{v}\in V}\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}}\exp\left\{\lambda\sum_{t=1}^{n}\varphi\Big(\eta(p_t(\boldsymbol{y}),y_t)v_t(\boldsymbol{y})\Big)\right\}\right),\end{aligned}$$

where the first inequality is Jensen's and the second follows because the maximum is contained in the sum. Now, for any fixed tree $\boldsymbol{v}$,

$$\begin{aligned}&\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\boldsymbol{p}}\exp\left\{\lambda\sum_{t=1}^{n}\varphi\Big(\eta(p_t(\boldsymbol{y}),y_t)v_t(\boldsymbol{y})\Big)\right\}\\ &=\mathop{\mathbb{E}}_{\boldsymbol{y}_{n-1}\sim\boldsymbol{p}}\mathop{\mathbb{E}}_{y_n\sim p_n(\boldsymbol{y})}\left[\exp\left\{\lambda\sum_{t=1}^{n}\varphi\Big(\eta(p_t(\boldsymbol{y}),y_t)v_t(\boldsymbol{y})\Big)\right\}\right]\\ &=\mathop{\mathbb{E}}_{\boldsymbol{y}_{n-1}\sim\boldsymbol{p}}\left[\exp\left\{\lambda\sum_{t=1}^{n-1}\varphi\Big(\eta(p_t(\boldsymbol{y}),y_t)v_t(\boldsymbol{y})\Big)\right\}\times\right.\\ &\qquad\qquad\left.\psi_{p_n(\boldsymbol{y}),\lambda}\Big(v_n(\boldsymbol{y})\Big)\right], \quad (10)\end{aligned}$$

where, for any $p\in[0,1]$ and $\lambda>0$, we define $\psi_{p,\lambda}:[-1,1]\to\mathbb{R}$ by

$$\begin{aligned}\psi_{p,\lambda}(v)&=\mathbb{E}_{y\sim p}\exp\Big\{\lambda\varphi\Big(\eta(p,y)v\Big)\Big\}\\ &=p\left(1+\frac{|v|}{p}\right)^{\lambda}\exp\left\{-\lambda\left(\frac{v+|v|}{p}\right)\right\}\\ &\quad+(1-p)\left(1+\frac{|v|}{1-p}\right)^{\lambda}\exp\left\{\lambda\left(\frac{v-|v|}{1-p}\right)\right\}.\end{aligned}$$

Then, we observe the following.

Lemma 12. *Whenever* $\lambda \le \frac{\log(3)-\log(2)}{2-\log(2)}$,

$$\sup_{p\in[0,1]}\ \sup_{v\in[p-1,p]} \psi_{p,\lambda}(v) \le 1.$$

The proof of Lemma 12 is a tedious calculation, and we leave it for Appendix A.4, but we provide a brief sketch of the argument here. First, $\psi_{p,\lambda}(v)$ can be simplified by fixing v to be positive or negative. This allows us to show that if λ is smaller than some function of p and v, $\psi_{p,\lambda}(v)$ is increasing when $v<0$ and decreasing when $v>0$. Then, we observe that this function of p and v (which must upper bound λ) is lower bounded by $\frac{\log(3)-\log(2)}{2-\log(2)}$. Finally, since $\psi_{p,\lambda}(0)=1$ for any $p\in[0,1]$ and $\lambda>0$, the result holds.

Thus, when $\lambda \le \frac{\log(3)-\log(2)}{2-\log(2)}$, (10) is bounded above by

$$\mathop{\mathbb{E}}_{\boldsymbol{y}_{n-1}\sim\boldsymbol{p}} \exp\left\{\lambda\sum_{t=1}^{n-1}\varphi\left(\eta(p_t(\boldsymbol{y}),y_t)v_t(\boldsymbol{y})\right)\right\}.$$

Iterating this argument through $t\in[n]$ and taking λ as large as possible gives the result. ■

We can apply Lemma 11 directly to (7) by observing that each tree $\boldsymbol{g}\in\mathcal{F}_{\boldsymbol{p},\boldsymbol{x}}$ can be written as $\boldsymbol{p}-(f\circ\boldsymbol{x})$ for some $f\in\mathcal{F}$, and consequently $g_t(\boldsymbol{y})\in[p_t(\boldsymbol{y})-1,p_t(\boldsymbol{y})]$ for all times $t\in[n]$ and paths $\boldsymbol{y}\in\{0,1\}^n$. Thus, without loss of generality, any cover $V_{\boldsymbol{p},\boldsymbol{x}}$ of $\mathcal{F}_{\boldsymbol{p},\boldsymbol{x}}$ can also be assumed to satisfy $v_t(\boldsymbol{y})\in[p_t(\boldsymbol{y})-1,p_t(\boldsymbol{y})]$, as clipping its value to this range will only decrease the approximation error.

Theorem 2 now follows by applying (8) and (9) to (6) and applying Lemma 11 to (7). ■

5. Proof of Theorem 3

We now prove Theorem 3. Lemma 19 in Appendix C shows that for $\mathcal{F}$ defined to be the 1-Lipschitz experts on $[0,1]^p$, $\mathcal{H}_\infty(\mathcal{F},\gamma,n)=\Theta(\gamma^{-p})$, so (2) applies for the upper bound. It remains to show that the lower bound holds. To begin, we lower bound the minimax regret in our adversarial setting by the minimax risk (the analogue of regret in batch learning) for the simpler i.i.d. batch setting with a well-specified model, which admits a simple expression in terms of KL divergence.

Let $\widehat{f}$ denote an arbitrary prediction strategy for the player that, for each t, outputs a predictor $\widehat{f}_t:\mathcal{X}\to[0,1]$ using only the history $h_{1:t-1}$. Then, let $\mathcal{P}$ be the set of all distributions on $(\mathcal{X},[0,1])$, and define the set

$$\mathcal{P}_{\mathcal{F}}=\left\{\mathcal{D}\in\mathcal{P}:\exists f^*_{\mathcal{D}}\in\mathcal{F}\ \forall x\in\mathcal{X}\ f^*_{\mathcal{D}}(x)=\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|x]\right\}.$$

Using these new objects, and letting $\mathrm{KL}\left(p\parallel q\right)$ denote the KL divergence between $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$, we obtain the following result (proven in Appendix C.1).

Lemma 13. *For any $\mathcal{X}$ and $\mathcal{F}\subseteq[0,1]^{\mathcal{X}}$,*

$$\frac{1}{n}\mathcal{R}_n(\mathcal{F})\ge\inf_{\widehat{f}}\sup_{\mathcal{D}\in\mathcal{P}_{\mathcal{F}}}\mathbb{E}\left[\mathrm{KL}\left(f^*_{\mathcal{D}}(x)\parallel\widehat{f}_n(x)\right)\right],$$

where $\mathbb{E}$ denotes expectation over $(x_{1:n-1},y_{1:n-1})\sim\mathcal{D}^{\otimes n-1}$ and $(x,y)\sim\mathcal{D}$.

Thus, we have reduced the problem to lower-bounding the minimax risk for $\mathcal{F}$ under a well-specified model, which is a more standard problem. To proceed, we use an argument along the lines of Assouad's lemma (Assouad, 1983), applied to our class $\mathcal{F}$ of 1-Lipschitz functions on $[0,1]^p$.

First, fix $\varepsilon\in(0,1/8)$, divide the space $[0,1]^p$ into $N=(\frac{1}{4\varepsilon})^p$ bins of width 4ε, and without loss of generality suppose that N is an integer. Denote the centers of each bin by $x^{(1)},\dots,x^{(N)}$. Define the set $\mathcal{V}=\{\pm1\}^N$ and the class $\mathcal{F}_{\mathcal{V}}\subseteq\mathcal{F}$ as follows: for each $v\in\mathcal{V}$, define the function f_v such that $f_v(x^{(i)})=4\varepsilon\mathbb{1}\{v_i=1\}+\varepsilon\mathbb{1}\{v_i=-1\}$ for $i\in[N]$. Define the rest of f_v by some linear interpolation between these points, and observe that f_v is 1-Lipschitz. Finally, for any $v\in\mathcal{V}$, define the distribution $\mathcal{D}_v$ on $([0,1]^p,[0,1])$ by $x\sim\mathrm{Unif}(\{x^{(1)},\dots,x^{(N)}\})$ and $y|x\sim\mathrm{Ber}(f_v(x))$.

Pick $v\in\mathcal{V}$ and $f:[0,1]^p\to[0,1]$. By definition of $\mathcal{D}_v$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_v}\mathrm{KL}\left(f_v(x)\parallel f(x)\right)=\frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}\left(f_v(x^{(i)})\parallel f(x^{(i)})\right).$$

Next, we use Lemma 18 in Appendix C to lower bound the KL divergence. Specifically, if $v_i=1$ then $f_v(x^{(i)})=4\varepsilon$, so

$$\mathrm{KL}\left(f_v(x^{(i)})\parallel f(x^{(i)})\right)\ge\frac{2\varepsilon}{3}\mathbb{1}\{f(x^{(i)})\le2\varepsilon\},$$

and if $v_i=-1$ then $f_v(x^{(i)})=\varepsilon$, so

$$\mathrm{KL}\left(f_v(x^{(i)})\parallel f(x^{(i)})\right)\ge\frac{\varepsilon}{4}\mathbb{1}\{f(x^{(i)})\ge2\varepsilon\}.$$

That is, for all $i\in[N]$,

$$\begin{aligned}&\mathrm{KL}\left(f_v(x^{(i)})\parallel f(x^{(i)})\right)\\&\quad\ge\frac{\varepsilon}{4}\Big[\mathbb{1}\{v_i=1\wedge f(x^{(i)})<2\varepsilon\}\\&\qquad\qquad+\mathbb{1}\{v_i=-1\wedge f(x^{(i)})\ge2\varepsilon\}\Big].\end{aligned}$$

Now, since the expression in Lemma 13 involves the supremum over all $\mathcal{D}\in\mathcal{P}_{\mathcal{F}}$, we can obtain a lower bound by taking an expectation over v uniformly chosen from $\mathcal{V}$ and setting $\mathcal{D}=\mathcal{D}_v$. In particular, for each $i\in N$, we define the distributions $\mathcal{D}_{+i}^{\otimes n-1}=2^{-(N-1)}\sum_{v\in\mathcal{V}:v_i=1}\mathcal{D}_v^{\otimes n-1}$

and $\mathcal{D}_{-i}^{\otimes n-1} = 2^{-(N-1)}\sum_{v\in\mathcal{V}:v_i=-1}\mathcal{D}_v^{\otimes n-1}$. Using the shorthand $\mathcal{D}^{\otimes n-1}(\cdot)$ to denote $\mathbb{P}_{(x_{1:n-1},y_{1:n-1})\sim\mathcal{D}^{\otimes n-1}}(\cdot)$, we obtain the lower bound for any $\widehat{f}$ of

$$\begin{aligned}
&\sup_{\mathcal{D}\in\mathcal{P}_{\mathcal{F}}} \mathbb{E}\left[\mathrm{KL}\left(f_{\mathcal{D}}^*(x) \,\|\, \widehat{f}_n(x)\right)\right] \\
&\geq \frac{1}{2^N}\sum_{v\in\mathcal{V}} \mathbb{E}\left[\mathrm{KL}\left(f_v(x) \,\|\, \widehat{f}_n(x)\right)\right] \\
&\geq \frac{1}{2^N}\sum_{v\in\mathcal{V}} \frac{\varepsilon}{4N}\sum_{i=1}^N \Big[\mathbb{1}\{v_i=1\}\mathcal{D}_v^{\otimes n-1}(\widehat{f}_n(x^{(i)})<2\varepsilon) \\
&\qquad\qquad + \mathbb{1}\{v_i=-1\}\mathcal{D}_v^{\otimes n-1}(\widehat{f}_n(x^{(i)})\geq 2\varepsilon)\Big] \\
&= \frac{\varepsilon}{8N}\sum_{i=1}^N \Big[\mathcal{D}_{+i}^{\otimes n-1}(\widehat{f}_n(x^{(i)})<2\varepsilon) \\
&\qquad\qquad + \mathcal{D}_{-i}^{\otimes n-1}(\widehat{f}_n(x^{(i)})\geq 2\varepsilon)\Big].
\end{aligned}$$

Then, we observe that for each $i\in[N]$,

$$\begin{aligned}
&\mathcal{D}_{+i}^{\otimes n-1}(\widehat{f}_n(x^{(i)})<2\varepsilon) + \mathcal{D}_{-i}^{\otimes n-1}(\widehat{f}_n(x^{(i)})\geq 2\varepsilon) \\
&= 1 + \mathcal{D}_{+i}^{\otimes n-1}(\widehat{f}_n(x^{(i)})<2\varepsilon) - \mathcal{D}_{-i}^{\otimes n-1}(\widehat{f}_n(x^{(i)})<2\varepsilon) \\
&\geq 1 - |\mathcal{D}_{+i}^{\otimes n-1}(\widehat{f}_n(x^{(i)})<2\varepsilon) - \mathcal{D}_{-i}^{\otimes n-1}(\widehat{f}_n(x^{(i)})<2\varepsilon)| \\
&\geq 1 - \left\|\mathcal{D}_{+i}^{\otimes n-1} - \mathcal{D}_{-i}^{\otimes n-1}\right\|_{\mathrm{TV}}.
\end{aligned}$$

Next, for each $v\in\mathcal{V}$, we define $\mathcal{D}_{v,+i}^{\otimes n-1}$ to be the distribution $\mathcal{D}_v^{\otimes n-1}$ with v_i forced to 1, and similarly define $\mathcal{D}_{v,-i}^{\otimes n-1}$ to be the distribution $\mathcal{D}_v^{\otimes n-1}$ with v_i forced to -1. Then, following the standard argument, we observe that

$$\begin{aligned}
\left\|\mathcal{D}_{+i}^{\otimes n-1} - \mathcal{D}_{-i}^{\otimes n-1}\right\|_{\mathrm{TV}} &= \left\|\frac{1}{2^N}\sum_{v\in\mathcal{V}}[\mathcal{D}_{v,+i}^{\otimes n-1} - \mathcal{D}_{v,-i}^{\otimes n-1}]\right\|_{\mathrm{TV}} \\
&\leq \frac{1}{2^N}\sum_{v\in\mathcal{V}}\left\|\mathcal{D}_{v,+i}^{\otimes n-1} - \mathcal{D}_{v,-i}^{\otimes n-1}\right\|_{\mathrm{TV}} \\
&\leq \max_{v,i}\left\|\mathcal{D}_{v,+i}^{\otimes n-1} - \mathcal{D}_{v,-i}^{\otimes n-1}\right\|_{\mathrm{TV}}.
\end{aligned}$$

Thus, we can apply this to Lemma 13 to get

$$\mathcal{R}_n(\mathcal{F}) \geq n\frac{\varepsilon}{8}\left[1 - \max_{v,i}\left\|\mathcal{D}_{v,+i}^{\otimes n-1} - \mathcal{D}_{v,-i}^{\otimes n-1}\right\|_{\mathrm{TV}}\right]. \quad (11)$$

To further lower bound this, consider a fixed $v\in\mathcal{V}$ and $i\in[N]$, and use $f_{v,+i}$ to denote f_v with v_i forced to 1, with the analogous definition for $f_{v,-i}$. By Pinsker's inequality and chain rule for KL,

$$\begin{aligned}
&\left\|\mathcal{D}_{v,+i}^{\otimes n-1} - \mathcal{D}_{v,-i}^{\otimes n-1}\right\|_{\mathrm{TV}}^2 \\
&\leq \frac{1}{2}\mathrm{KL}\left(\mathcal{D}_{v,+i}^{\otimes n-1} \,\|\, \mathcal{D}_{v,-i}^{\otimes n-1}\right) \\
&= \frac{n-1}{2N}\sum_{j=1}^N \mathrm{KL}\left(f_{v,+i}(x^{(j)}) \,\|\, f_{v,-i}(x^{(j)})\right) \\
&= \frac{n-1}{2N}\cdot\mathrm{KL}\left(4\varepsilon \,\|\, \varepsilon\right),
\end{aligned}$$

where the last step uses that $f_{v,+i}$ and $f_{v,-i}$ agree everywhere except $x^{(i)}$. Finally, we observe that

$$\begin{aligned}
\mathrm{KL}\left(4\varepsilon \,\|\, \varepsilon\right) &= 4\varepsilon\log(4) + (1-4\varepsilon)\log\left(\frac{1-4\varepsilon}{1-\varepsilon}\right) \\
&\leq 4\varepsilon\log(4) \\
&\leq 8\varepsilon.
\end{aligned}$$

We conclude from the definition of N that

$$\begin{aligned}
\left\|\mathcal{D}_{v,+i}^{\otimes n-1} - \mathcal{D}_{v,-i}^{\otimes n-1}\right\|_{\mathrm{TV}}^2 &\leq 4\frac{(n-1)\varepsilon}{N} \\
&= 4(n-1)\varepsilon(4\varepsilon)^p \\
&\leq n(4\varepsilon)^{1+p}.
\end{aligned}$$

Setting $\varepsilon = \frac{1}{8}n^{-\frac{1}{p+1}}$ gives $n(4\varepsilon)^{1+p} = 2^{-(1+p)} \leq 1/4$, and plugging this into (11) gives the lower bound

$$\mathcal{R}_n(\mathcal{F}) \geq n\frac{n^{-\frac{1}{p+1}}}{(8)(8)}[1-1/2] = \frac{n^{\frac{p}{p+1}}}{128}. \qquad \blacksquare$$

6. Discussion

We have shown that the self-concordance property of log loss leads to improved bounds on the minimax regret for sequential probability assignment with rich classes of experts, and that the rates we provide cannot be improved further without stronger structural assumptions on the expert class. An important open problem is to develop more refined complexity measures (e.g., variants of sequential entropy tailored directly to the log loss rather than the L_∞ norm) that lead to matching upper and lower bounds for *all* classes of experts; we intend to pursue this in future work.

On the technical side, it would be interesting to extend our guarantees to infinite outcome spaces; that is, adversarial online density estimation. To the best of our knowledge, very little progress has been made on this problem without stochastic assumptions.

Acknowledgements

BB is supported by an NSERC Canada Graduate Scholarship. The authors thank Jeffrey Negrea for many helpful discussions. DF thanks Sasha Rakhlin and Karthik Sridharan for many useful discussions over the years about this problem. This work was begun while all three authors were visiting the Simons Institute for Theoretical Computing in Berkeley, California for the Foundations of Deep Learning program, and completed while BB and DMR were visiting the Institute for Advanced Study in Princeton, New Jersey for the Special Year on Statistics, Optimization, and Theoretical Machine Learning. DMR acknowledges the support of an NSERC Discovery Grant and a stipend provided by the Charles Simonyi Endowment. BB's travel to the Institute for Advanced Study was funded by an NSERC Michael Smith Foreign Study Supplement. DF acknowledges the support of TRIPODS award #1740751.

References

Abernethy, J., Agarwal, A., Bartlett, P., and Rakhlin, A. A Stochastic View of Optimal Regret through Minimax Duality. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

Assouad, P. Deux Remarques sur l'Estimation. *C. R. Academy Scientifique Paris Séries I Mathematics*, 296 (23):1021–1024, 1983.

Banerjee, A. On Bayesian Bounds. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Cesa-Bianchi, N. and Lugosi, G. Minimax Regret under Log Loss for General Classes of Experts. In *Proceedings of the 12th Conference on Computational Learning Theory*, 1999.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. On the Generalization Ability of On-Line Learning Algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Cover, T. Universal Portfolios. *Mathematical Finance*, 1(1): 1–29, 1991.

Cover, T. and Ordentlich, E. Universal Portfolios with Side Information. *IEEE Transactions on Information Theory*, 42(2):348–363, 1996.

Cross, J. and Barron, A. Efficient Universal Portfolios for Past-Dependent Target Classes. *Mathematical Finance*, 13(2):245–276, 2003.

Foster, D., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Logistic Regression: The Importance of Being Improper. In *Proceedings of the 31st Conference on Learning Theory*, 2018.

Freund, Y. Predicting a Binary Sequence almost as well as the Optimal Biased Coin. *Information and Computation*, 182(2):73–94, 2003.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, 2014.

Grnarova, P., Levy, K., Lucchi, A., Hofmann, T., and Krause, A. An Online Learning Approach to Generative Adversarial Networks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Hazan, E. Introduction to Online Convex Optimization. *Foundations and Trends in Optimization*, 2(3–4):157–325, 2016.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Liang, Q. and Modiano, E. Minimizing Queue Length Regret under Adversarial Network Models. In *ACM on Measurement and Analysis of Computing Systems*. Association for Computing Machinery, 2018.

Merhav, N. and Feder, M. Universal Prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

Miyaguchi, K. and Yamanishi, K. Adaptive Minimax Regret against Smooth Logarithmic Losses over High-Dimensional ℓ_1-Balls via Envelope Complexity. In *Proceedings of Machine Learning Research*, volume 89, pp. 3440–3448, 2019.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Nesterov, Y. and Nemirovski, A. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia, USA, 1994.

Opper, M. and Haussler, D. Worst Case Prediction over Sequences under Log Loss. In Cybenko, G., O'Leary, D., and Rissanen, J. (eds.), *The Mathematics of Information Coding, Extraction and Distribution*, pp. 81–90, New York, NY, 1999. Springer New York.

Rakhlin, A. and Sridharan, K. Statistical Learning and Sequential Prediction, 2012. URL http://www.mit.edu/~rakhlin/courses/stat928/stat928_notes.pdf.

Rakhlin, A. and Sridharan, K. Online Nonparametric Regression. In *Proceedings of the 27th Conference on Learning Theory*, 2014.

Rakhlin, A. and Sridharan, K. Sequential Probability Assignment with Binary Alphabets and Large Classes of Experts, 2015. arXiv:1501.07340.

Rakhlin, A., Sridharan, K., and Tewari, A. Online Learning via Sequential Complexities. *Journal of Machine Learning Research*, 16(6):155–186, 2015a.

Rakhlin, A., Sridharan, K., and Tewari, A. Sequential Complexities and Uniform Martingale Laws of Large Numbers. *Probability Theory and Related Fields*, 161(1): 111–153, 2015b.

Rakhlin, A., Sridharan, K., and Tsybakov, A. Empirical Entropy, Minimax Regret and Minimax Risk. *Bernoulli*, 23(2):789–824, 2017.

Rissanen, J. Universal Coding, Information, Prediction, and Estimation. *IEEE Transactions on Information Theory*, 30:629–636, 1984.

Rissanen, J. Complexity of Strings in the Class of Markov Sources. *IEEE Transactions on Information Theory*, 32 (4):526–532, 1986.

Rissanen, J. Fisher Information and Stochastic Complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.

Shtar'kov, Y. Universal Sequential Coding of Single Messages. *Problems of Information Transmission*, 23(3):3–17, 1987.

Sion, M. On General Minimax Theorems. *Pacific Journal of Mathematics*, 8:171–176, 1958.

Vovk, V. A Game of Prediction with Expert Advice. *Journal of Computer and System Sciences*, 56:153–173, 1998.

Wainwright, M. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Xie, Q. and Barron, A. Asymptotic Minimax Regret for Data Compression, Gambling, and Prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, 2000.