Sequential Probability Assignment with Binary Alphabets and Large Classes of Experts

Alexander Rakhlin
University of Pennsylvania

Karthik Sridharan
Cornell University

January 28, 2015

Abstract

We analyze the problem of sequential probability assignment for binary outcomes with side information and logarithmic loss, where regret—or, redundancy—is measured with respect to a (possibly infinite) class of experts. We provide upper and lower bounds for minimax regret in terms of sequential complexities of the class, introduced in [14, 13]. These complexities were recently shown to give matching (up to logarithmic factors) upper and lower bounds for sequential prediction with general convex Lipschitz loss functions [11, 12]. To deal with unbounded gradients of the logarithmic loss, we present a new analysis that employs a sequential chaining technique with a Bernstein-type bound. The introduced complexities are intrinsic to the problem of sequential probability assignment, as illustrated by our lower bound in terms of the offset Rademacher complexity.

We also consider an example of a large class of experts parametrized by vectors in a high-dimensional Euclidean ball (or a Hilbert ball). The typical discretization approach fails, while our techniques give a non-trivial bound. For this problem we also present an algorithm based on regularization with a self-concordant barrier. This algorithm is of an independent interest, as it requires a bound on the function values rather than gradients.

1 Introduction

In this paper we study the problem of sequential prediction of a string of bits $(y_1,\dots,y_n) \triangleq y_{1:n} \in \{0,1\}^n$. At each round $t = 1,\dots,n$, the forecaster observes side information $x_t \in \mathscr{X}_t$, decides on the probability $\widehat{y}_t \in [0,1]$ of the event $y_t = 1$, observes the outcome $y_t \in \{0,1\}$, and pays according to the logarithmic (or, *self-information*) loss function

$$\boldsymbol{\ell}(\widehat{y}_t, y_t) = -\mathbf{1}\left\{y = 1\right\}\log\widehat{y}_t - \mathbf{1}\left\{y_t = 0\right\}\log(1-\widehat{y}_t).$$

At each time instance t, the side-information set $\mathscr{X}_t$ is a subset of an abstract set $\mathscr{X}$. The subset $\mathscr{X}_t$ is allowed to depend on the history $h_{1:t-1} \triangleq (x_{1:t-1}, y_{1:t-1})$, and the functions $\mathscr{X}_t : (\mathscr{X}\times\mathscr{Y})^{t-1} \to 2^{\mathscr{X}}$ are assumed to be known to the forecaster.

The goal of the forecaster is to predict as well as a benchmark set $\mathscr{F}$ of functions—sometimes called "experts"—mapping $\mathscr{X}$ to $[0,1]$. More specifically, the goal is to keep *regret*

$$\sum_{t=1}^n \boldsymbol{\ell}(\widehat{y}_t, y_t) - \inf_{f\in\mathscr{F}} \sum_{t=1}^n \boldsymbol{\ell}(f(x_t), y_t)$$

as small as possible for all sequences $y_1,\dots,y_n$ and $x_1,\dots,x_n$ (satisfying $x_t \in \mathscr{X}_t(h_{1:t-1})$).

To illustrate the setting, consider a few examples. We may take $\mathscr{X}_t(h_{1:t-1}) = \left\{(y_1,\dots,y_{t-1})\right\} \subset \{0,1\}^{t-1}$ to be a singleton set containing the exact realization of the sequence so far. In this case, the choice $x_t = (y_1,\dots,y_{t-1})$ is enforced and $f(x_t) = p_f(1|y_1,\dots,y_{t-1})$ may be viewed as a conditional distribution; the normalized maximum likelihood forecaster is known to be minimax optimal in this extensively studied scenario (e.g. [4, Ch. 9]). Alternatively, we may define $\mathscr{X}_t(h_{1:t-1}) = \left\{y' \in \{0,1\}^{t-1} : d_H(y_{1:t-1}, y') \le r\right\}$ to be a set that contains histories with up to r flips of the bits. In this case, the forecaster is facing a situation where history can be slightly altered in an adversarial fashion. As another example, we may take $\mathscr{X}_t(h_{1:t-1}) = \left\{(y_{t-k},\dots,y_{t-1})\right\}$, in which case the forecaster competes with a set of kth-order stationary Markov experts. The set $\mathscr{X}_t$ may also be time-invariant, in which case f is a memoryless expert that acts on side information. In short, the formulation we presented subsumes a wide range of interesting problems. Our goal in this paper is to understand how "complexity" of $\mathscr{F}$ affects minimax rates of regret.

The minimax regret for the problem of sequential probability assignment can be written as

$$V_n(\mathscr{F}) = \left\langle\!\!\left\langle \sup_{x_t \in \mathscr{X}_t(x_{1:t-1}, y_{1:t-1})} \inf_{\widehat{y}_t \in [0,1]} \sup_{p_t \in [0,1]} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{y}_t, y_t) - \inf_{f \in \mathscr{F}} \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \right\} \tag{1}$$

where $\mathbb{E}_{y_t \sim p_t}$ is a shorthand for the expectation with respect to Bernoulli y_t with bias p_t. Following [10], the notation $\langle\!\langle \dots \rangle\!\rangle_{t=1}^{n}$ represents a repeated application of the operators inside the brackets and corresponds to the unrolled minimax value of the associated game between the forecaster and Nature. Any upper bound on $V_n(\mathscr{F})$ guarantees existence of a strategy that attains regret of at most that amount. In the last few years, new techniques with roots in empirical process theory have emerged for analyzing minimax values of the form (1). We bring these techniques to bear on the problem of sequential probability assignment with self-information loss.

Our point of comparison will be the study of rich classes in [4, Section 9.10]. Following [4], we employ the truncation method to deal with the unbounded loss function. To this end, fix $\delta \in (0, 1/2)$, to be chosen later. For $a \in [0,1]$, let $\tau_\delta(a)$ denote the thresholded value

$$\tau_\delta(a) \;=\; \begin{cases} \delta & \text{if } a < \delta \\ a & \text{if } a \in [\delta, 1-\delta] \\ 1-\delta & \text{if } a > 1-\delta. \end{cases}$$

For a class $\mathscr{F}$, let $\mathscr{F}^\delta = \{\tau_\delta(f) : f \in \mathscr{F}\}$ denote the class of truncated functions. It is easy to check (see [4, Lemma 9.5]) that

$$V_n(\mathscr{F}) \le V_n(\mathscr{F}^\delta) + 2n\delta, \tag{2}$$

and we can, therefore, focus on the minimax regret with respect to $\mathscr{F}^\delta$. We show that $V_n(\mathscr{F}^\delta)$ can be upper bounded via a modified (offset) sequential Rademacher complexity, which in turn can be controlled via sequential chaining in the spirit of [11, 12]. Unlike the latter two papers, however, we do not employ symmetrization and instead use the self-information property of the loss function. We are able to mitigate the adverse dependence of $V_n(\mathscr{F}^\delta)$ on δ by introducing chaining with Bernstein-style terms that control the sub-Gaussian and sub-exponential tail behaviors. As an example, we recover the $n^{3/5}$ rate for monotonically increasing experts presented in [4, Sec 9.10-9.11]. However, our technique goes well beyond such examples of "static" experts. In particular, we can obtain non-trivial rates even in the setting where discretization in the style of [4, Sec 9.10-9.11],[3] leads to vacuous bounds. One such example is when experts are indexed by a unit ball in a Hilbert space (or, a high-dimensional Euclidean space) and expert's prediction depends linearly on side information. A discretization in the supremum norm of this set of experts is not finite, and thus the typical approaches to this problem fail. In contrast, we employ the ideas from empirical process theory and its sequential generalization in [14] in order to define "data-dependent" notions of complexity.

Despite the improvement over the technique of [4], the rates attained in this paper are not always minimax optimal, as we demonstrate in Section 6. This is in contrast to other loss functions (such as absolute, square, q-power, and logistic) for which matching upper and lower bounds (to within logarithmic factors) have been established recently in [12]. As mentioned in [4], the truncation method is crude, and we leave it as an open question whether a different technique can be employed to attain optimal rates.

We finish this introduction with a brief mention that sequential probability assignment is extensively studied in Information Theory, where regret is known as *redundancy* with respect to a set of codes. The vast literature mostly investigates the case of parametric classes (see [17, 18, 5, 15, 16] and the references in [4, Ch. 9]), with exact constants available in certain cases. We refer to [7] for a discussion of approaches to dealing with large comparator classes. Given the well-known connection to compression, it would be interesting to employ the relaxation-based algorithmic recipe of [9, 12, 10] to come up with novel data compression methods.

2 Complexity of Large Classes of Experts

We focus on the minimax value for the thresholded class $\mathscr{F}^{\delta}$. To state the first technical lemma, we need the definition of a tree. For an abstract set $\mathcal{Z}$, a $\mathcal{Z}$-valued complete binary tree $\mathbf{z}$ of depth n is a collection of labeling functions $\mathbf{z}_t : \{0,1\}^{t-1} \to \mathcal{Z}$ for $t \in \{1,\dots,n\}$. For a sequence $y = (y_1,\dots,y_n) \in \{0,1\}^n$ (which we call *a path*), we write $\mathbf{z}_t(y)$ for $\mathbf{z}_t(y_1,\dots,y_{t-1})$. Once we take $y_1,\dots,y_n$ to be random variables, we may view $\{\mathbf{z}_t\}$ as a predictable process with respect to the filtration given by $\sigma(y_1,\dots,y_{t-1})$. [1]

We will say that an $\mathbf{x}$ tree is *consistent* with respect to the side information set mappings $h_{1:t-1} \mapsto \mathscr{X}_t(h_{1:t-1})$ if for any $y \in \{0,1\}^n$, it holds that for all t,

$$\mathbf{x}_t(y) \in \mathscr{X}_t(\mathbf{x}_1(y),\dots,\mathbf{x}_{t-1}(y), y_1,\dots,y_{t-1}).$$

A consistent tree respects the sets of constraints $\mathscr{X}_t$ imposed by the problem. For the purposes of analyzing complexity of $\mathscr{F}$, it is important that the constraints are reflected in the tree $\mathbf{x}$.

Theorem 1 below relates the minimax regret with respect to $\mathscr{F}^{\delta}$ to the supremum of a stochastic process of a form similar to *offset Rademacher complexity* introduced in [11]. The key difference with respect to [11] is that the stochastic process is defined with potentially biased coin flips. To prove Theorem 1, we avoid symmetrization and instead exploit the fact that the logarithmic loss has the self-information property: in the maximin dual, the optimal probability assignment is given precisely by the distribution of the y_t variable. We note that the symmetrization approach of [11] appears to give worse rates for the logarithmic loss function.

Let

$$\eta(p,a) \triangleq -\mathbf{1}\{a=1\}\, p^{-1} + \mathbf{1}\{a=0\}\,(1-p)^{-1} \tag{3}$$

and observe that η is zero-mean if a is Bernoulli random variable with bias p.

Theorem 1. *The following upper bound holds:*

$$V_n(\mathscr{F}^{\delta}) \le \sup_{\mathbf{x},\boldsymbol{\mu},\mathbf{p}} \mathbb{E} \sup_{f\in\mathscr{F}^{\delta}} \left[\sum_{t:\mathbf{p}_t(y)\in[\delta,1-\delta]} \eta(\mathbf{p}_t(y),y_t)\left(\boldsymbol{\mu}_t(y) - f(\mathbf{x}_t(y))\right) - \frac{1}{2}\left(\boldsymbol{\mu}_t(y) - f(\mathbf{x}_t(y))\right)^2 \right] + 2n\delta\log(1/\delta),$$

where $\mathbf{p},\boldsymbol{\mu}$ *range over all* $[0,1]$*-valued trees,* $\mathbf{x}$ *ranges over consistent trees, and the stochastic process* $y_1,\dots,y_n$ *is defined via* $y_t|y_1,\dots,y_{t-1} \sim \mathsf{Bernoulli}(\mathbf{p}_t(y_1,\dots,y_{t-1}))$.

To shorten the notation in Theorem 1, let $\mathcal{Z} = \mathscr{X} \times [0,1]$ and for every $f \in \mathscr{F}^{\delta}$, write $g_f(z) = g_f(x,a) = a - f(x)$. The upper bound of Theorem 1 can be written more succinctly as

$$V_n(\mathscr{F}^{\delta}) \le \sup_{\mathbf{z},\mathbf{p}} \mathbb{E} \sup_{f\in\mathscr{F}^{\delta}} \left[\sum_{t:\mathbf{p}_t(y)\in[\delta,1-\delta]} \eta(\mathbf{p}_t(y),y_t) g_f(\mathbf{z}_t(y)) - \frac{1}{2} g_f(\mathbf{z}_t(y))^2 \right] + 2n\delta\log(1/\delta). \tag{4}$$

We keep in mind that the $\mathbf{x}$ part of $\mathbf{z}$ is a consistent tree. Observe that the expression above is a supremum of a collection of random variables indexed by $f \in \mathscr{F}^{\delta}$, each with a nonpositive-mean. To analyze the supremum of this stochastic process, we first consider the case when the indexing set is finite.

Lemma 2. *For any set* V *consisting of* $[-1,1]$ *valued trees, any* $[\delta,1-\delta]$*-valued tree* $\mathbf{p}$*, and any* $c>0$*,*

$$\mathbb{E}_y \max_{\mathbf{v}\in V} \left[\sum_{t=1}^{n} \eta(\mathbf{p}_t(y),y_t)\mathbf{v}_t(y) - c\mathbf{v}_t(y)^2 \right] \le \frac{\log|V|}{\delta\log(1+\frac{c}{2})}$$

where $y_t|y_1,\dots,y_{t-1} \sim \mathsf{Bernoulli}(\mathbf{p}_t(y_1,\dots,y_{t-1}))$*. Furthermore, the same upper bound holds if* $\mathbf{p}$ *is any* $[0,1]$*-valued tree but the summation is restricted to* $\{t : \mathbf{p}_t(y) \in [\delta,1-\delta]\}$*.*

[1] We remark that in [14, 13, 10], the trees are defined with respect to $\{\pm 1\}$-valued sequences, whereas here we use the $\{0,1\}$-valued variables. The change is purely notational and all the definitions and results can be rephrased appropriately.

The tight control of the expectation is possible because of the negative quadratic term that acts as a compensator. On the downside, the upper bound displays the adverse $1/\delta$ dependence. We now show a maximal inequality when the quadratic term is not present. The bound is of a Bernstein type, with the sub-Gaussian and sub-exponential behaviors. Crucially, the sub-Gaussian term scales with $1/\sqrt{\delta}$.

Lemma 3. *For any set V consisting of $[-1,1]$ valued trees, any $[\delta,1-\delta]$-valued tree $\mathbf{p}$, and any $c>0$,*

$$\mathbb{E}_y \max_{\mathbf{v}\in V}\left[\sum_{t=1}^n \eta(\mathbf{p}_t(y),y_t)\mathbf{v}_t(y)\right] \le 5\bar{v}\sqrt{\frac{n\log|V|}{\delta}}+\frac{2v_{max}\log|V|}{\delta}$$

where $y_t|y_1,\dots,y_{t-1}\sim \mathsf{Bernoulli}(\mathbf{p}_t(y_1,\dots,y_{t-1}))$, $\bar{v}=\max_{\mathbf{v}\in V}\max_y(\frac{1}{n}\sum_{t=1}^n \mathbf{v}_t(y)^2)^{1/2}$, and $v_{max}=\max_{\mathbf{v}\in V}\max_y|\mathbf{v}_t(y)|$. The same upper bound holds if $\mathbf{p}$ is any $[0,1]$-valued tree but the summation is restricted to $\{t:\mathbf{p}_t(y)\in[\delta,1-\delta]\}$.

We now pass from a finite collection to an infinite one via the sequential chaining technique [14]. For this purpose, we recall the definition of ℓ_p sequential covering numbers.

Definition 1 ([14])**.** A set V of $\mathbb{R}$-valued trees of depth n is a (sequential) γ-cover (with respect to ℓ_p, $p\ge 1$) of $\mathscr{G}\subseteq\mathbb{R}^{\mathcal{Z}}$ on a $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n if

$$\forall g\in\mathscr{G},\ \ y\in\{0,1\}^n,\ \ \exists\mathbf{v}\in V,\ \text{ s.t. } \left(\frac{1}{n}\sum_{t=1}^n|\mathbf{v}_t(y)-g(\mathbf{z}_t(y))|^p\right)^{1/p}\le\gamma. \tag{5}$$

The size of the smallest γ-cover is denoted by $\mathcal{N}_p(\mathscr{G},\gamma,\mathbf{z})$. For $p=\infty$, (5) becomes $\max_t|\mathbf{v}_t(y)-g(\mathbf{z}_t(y))|\le\gamma$.

Theorem 4. *Let $\mathscr{G}$ be a class of functions $\mathcal{Z}\to[-1,1]$. For any $[0,1]$-valued tree $\mathbf{p}$, any $\mathcal{Z}$-valued tree $\mathbf{z}$, any $K>0$, and $\gamma>0$,*

$$\begin{aligned}\mathbb{E}\sup_{g\in\mathscr{G}}&\left[\sum_{t:\mathbf{p}_t(y)\in[\delta,1-\delta]}\eta(\mathbf{p}_t(y),y_t)g(\mathbf{z}_t(y))-Kg(\mathbf{z}_t(y))^2\right]\\ &\le\frac{1}{\delta}\frac{\log\mathcal{N}_\infty(\mathscr{G},\gamma,\mathbf{z})}{\log(1+\frac{K}{8})}+\inf_{\alpha(0,\gamma]}\left\{\frac{4n\alpha}{\delta}+30\sqrt{\frac{2n}{\delta}}\int_\alpha^\gamma\sqrt{\log\mathcal{N}_\infty(\mathscr{G},\rho,\mathbf{z})}\,d\rho+\frac{8}{\delta}\int_\alpha^\gamma\log\mathcal{N}_\infty(\mathscr{G},\rho,\mathbf{z})\,d\rho\right\}\end{aligned}$$

where the stochastic process $y_1,\dots,y_n$ is defined via $y_t|y_1,\dots,y_{t-1}\sim\mathsf{Bernoulli}(\mathbf{p}_t(y_1,\dots,y_{t-1}))$.

Theorem 4 is readily applied to the upper bound of Theorem 1 by identifying

$$\mathscr{G}=\{g_f(z)=g_f(x,\mu)=\mu-f(x):f\in\mathscr{F}^\delta,\mu\in\mathbb{R},x\in\mathscr{X}\}$$

and $\mathbf{z}_t(y)=(\mathbf{x}_t(y),\boldsymbol{\mu}_t(y))$. It is immediate from the definition of a cover that for any $\boldsymbol{\mu}$, $\mathbf{x}$, and $\mathbf{z}=(\mathbf{x},\boldsymbol{\mu})$,

$$\mathcal{N}_p(\mathscr{F}^\delta,\mathbf{x},\alpha)=\mathcal{N}_p(\mathscr{G},\mathbf{z},\alpha). \tag{6}$$

The lower bound of Lemma 10 (presented in Section 7) and the relation between the offset Rademacher complexity and sequential fat-shattering dimension [14, 12] yield the next theorem.

Theorem 5. *For the case of constant sets $\mathscr{X}_1=\mathscr{X}_2=\ldots=\mathscr{X}$, the following are equivalent:*

- *Minimax regret is sublinear: $\frac{1}{n}V_n(\mathscr{F})\to 0$ as $n\to\infty$*
- *Sequential dimension $\mathrm{fat}_\beta(\mathscr{F},\mathscr{X})$ is finite for all $\beta>0$*

A few remarks are in order. First, the theorem can be easily extended to non-constant sets $\mathscr{X}_t$, in which case fat_β is defined with respect to consistent trees (as in the next section). Second, one may also phrase the equivalence through sequential covering numbers, thanks to the relations outlined in [14, 12].

In summary, the sequential complexities we study are intrinsic to the problem of sequential probability assignment (unlike, for instance, covering numbers with respect to the supremum norm on $\mathscr{X}$ — see Section 4 for an example). Yet, the upper bounds we derive do not quite match the lower bounds, due to the hard thresholding approach and the need to balance $n\delta$ with $V_n(\mathscr{F}^\delta)$ at the end of the day. It is an open problem to close the gap between the upper and lower bounds.

The upper bound of Theorem 4 is quantified as soon as we have control of sequential covering numbers. While covering numbers could be computed directly in many situations, it is often simpler to upper bound a "scale-sensitive dimension" of the class, defined in the next section. In Section 5 we present an example of such a simple calculation.

3 Covering Numbers and Combinatorial Parameters

Suppose we can define a preorder $\preceq$ on the set $\mathscr{X}$ (that is, a binary relation that is reflexive and transitive). We say that an $\mathscr{X}$-valued tree $\mathbf{x}$ of depth n is *ordered* if for any path $y \in \{0,1\}^n$, it holds that $\mathbf{x}_t(y) \preceq \mathbf{x}_{t+1}(y)$ for all $t = 1, \ldots, n-1$. In this section we show that the combinatorial dimensions, covering numbers, and the associated upper bounds in [14] can be extended to "respect" the preorder (of course, one can always define a vacuous relation $\preceq$ and recover prior results).

Definition 2. A class $\mathscr{F} \subset \mathbb{R}^{\mathscr{X}}$ shatters (at scale $\beta > 0$) an ordered $\mathscr{X}'$-valued tree of depth d if there exists a $\mathbb{R}$-valued witness tree $\mathbf{s}$ of depth d such that

$$\forall y \in \{0,1\}^d, \ \exists f \in \mathscr{F}, \ \text{s.t.} \ (2y_t - 1)(f(\mathbf{x}_t(y)) - \mathbf{s}_t(y)) \geq \beta/2.$$

The largest depth of an ordered $\mathscr{X}'$-valued tree is denoted by $\text{fat}^o_\beta(\mathscr{F}, \mathscr{X}')$, where the superscript o stands for "ordered".

The notion of the Littlestone's dimension $\text{Ldim}(\mathscr{F}, \mathscr{X}')$ for $\{0, \ldots, k\}$-valued function classes extends in exactly the same way to the case of ordered trees.

The main step in obtaining upper bounds on sequential covering numbers is the analogue of the Vapnik-Chervonenkis-Sauer-Shelah lemma, proved in [14, 13]. We now show that if we ask for a β-cover on an ordered tree $\mathbf{x}$, the sequential covering numbers are controlled via the ordered version $\text{fat}^o_\beta(\mathscr{F}, \mathsf{Img}(\mathbf{x}))$ of the fat-shattering dimension in Definition 2.

Theorem 6 (Extension of Theorem 4 in [14])**.** *Let $\mathscr{F} \subseteq \{0, \ldots, k\}^{\mathscr{X}}$ be a class of functions with $\text{fat}^o_2(\mathscr{F}, \mathscr{X}) = d$. Then for any $n > d$ and any ordered $\mathscr{X}$-valued tree $\mathbf{x}$,*

$$\mathscr{N}_\infty(\mathscr{F}, 1/2, \mathbf{x}) \leq \sum_{i=0}^{d} \binom{n}{i} k^i.$$

Hence, for a class $\mathscr{G} \subseteq [-1,1]^{\mathscr{X}}$, for any $\beta > 0$,

$$\mathscr{N}_\infty(\mathscr{G}, \beta, \mathbf{x}) \leq \left(\frac{2en}{\beta}\right)^{\text{fat}^o_\beta(\mathscr{G}, \mathscr{X})}.$$

The following three sections are devoted to particular examples. We start by exhibiting a simple class for which sequential covering numbers are small, yet the discretization with respect to the supremum norm (typically performed to appeal to a finite-experts method) gives vacuous bounds.

4 Example: Consistent History

We would like to illustrate that sequential covering number can be much smaller than covering numbers with respect to the supremum norm over $\mathscr{X}$. Consider the particular case of $\mathscr{X}_t(h_{1:t-1}) = \{(y_1, \ldots, y_{t-1})\}$. Clearly, there is only one consistent tree, namely the one defined by $\mathbf{x}_t(y) = (y_1, \ldots, y_{t-1})$ for any t. In this case, the requirement (5) in Definition 1 with class $\mathscr{F}^\delta$, consistent tree $\mathbf{x}$, and $p = \infty$ reads as

$$\forall f \in \mathscr{F}^\delta, \ \ y \in \{0,1\}^n, \ \exists \mathbf{v} \in V, \ \text{s.t.} \ |\mathbf{v}_t(y_{1:t-1}) - f(y_{1:t-1})| \leq \gamma. \tag{7}$$

We contrast this with the definition in [4, Sec. 9.10], where the covering of $\mathscr{F}$ is done with respect to the following pointwise metric (which we normalized by $\sqrt{n}$ for uniformity):

$$d(f,g) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\sup_{y_{1:t}}\left(\boldsymbol{\ell}(f(y_{1:t-1}),y_t) - \boldsymbol{\ell}(g(y_{1:t-1}),y_t)\right)^2}. \tag{8}$$

To illustrate a gap in the two covering-number approaches, construct a particular class $\mathscr{F}$ as follows. For each element $b \in \{0,1\}^n$, define f_b by

$$f_b(y_{1:t-1}) = \frac{1}{4}\mathbf{1}\left\{b_{1:t-1} = y_{1:t-1}\right\} + \frac{1}{4}$$

and take $\mathscr{F} = \{f_b : b \in \{0,1\}^n\}$. In other words, on round t, expert f_b predicts probability 1/2 if history coincides with $b_{1:t-1}$, and 1/4 otherwise. For two elements $f_b, f_{b'} \in \mathscr{F}$, let $\kappa(b,b') = \max\{t : b_t = b'_t\}$ be the last time the two sequences agree (defined as 0 if $b_1 \neq b'_1$). Then

$$\sum_{t=1}^{n}\sup_{y_{1:t}}\left(\boldsymbol{\ell}(f_b(y_{1:t-1}),y_t) - \boldsymbol{\ell}(f_{b'}(y_{1:t-1}),y_t)\right)^2 \geq \sum_{t=1}^{n}\left(\boldsymbol{\ell}(f_b(b_{1:t-1}),1) - \boldsymbol{\ell}(f_{b'}(b_{1:t-1}),1)\right)^2 \geq (n-\kappa(b,b'))\log(2)^2$$

and thus there are at least $2^{n/2}$ functions at a constant distance $d(f,g) \geq c$.

In contrast, consider sequential covering in the sense of (7) (and Definition 1). Take any $y \in \{0,1\}^n$ and $f_b \in \mathscr{F}$. The sequence of n values $(f_b(\emptyset), f_b(y_1), \ldots, f_b(y_{1:t-1}), \ldots, f_b(y_{1:n-1}))$ is equal to 1/2 until $t = \kappa(b,y)$ and 1/4 afterwards. Let V be a set of n trees $\mathbf{v}^1, \ldots, \mathbf{v}^n$ labeled by $\{1/4, 1/2\}$. Each $\mathbf{v}^i$ is defined as

$$\forall y \in \{0,1\}^n, t \in \{1,\ldots,n\}, \qquad \mathbf{v}^i_t(y) = (1/4)\mathbf{1}\{t \leq i-1\} + 1/4.$$

It is immediate that this set of n trees provides an exact cover of $\mathscr{F}$ (at scale 0) in the sense of Definition 1. This leads to $\mathscr{O}(\log(n)/n)$ bounds on minimax regret, while the discretization with respect to the supremum norm (8) fails.

The above failure is endemic to approaches that attempt to discretize the set of experts before the prediction process even started. In contrast, sequential complexities can be viewed as an analogue of "data-based" discretization, which is known in statistical learning since the work of Vapnik and Chervonenkis in the 60's.

5 Example: Monotonically Nondecreasing Experts

We consider an example of a nonparametric class analyzed in [4, p. 270]. Let $f \in \mathscr{F}$ be a set of experts such that the forecasted probability does not decrease in time. To model this scenario in a general manner, we suppose that the side information $x_t = (t, x'_t) \in \mathbb{N} \times \mathscr{X}'_t(x_{1:t-1}, y_{1::t-1})$ contains the time stamp, and $f(t+1, x'_t) \geq f(t, x''_t)$ for any $f \in \mathscr{F}$. The particular case of *static* experts—with prediction depending only on t and no other side information—has been considered in [4].

To invoke the results of the previous section, define a preorder on $(t,x) \in \mathscr{X} = \mathbb{N} \times \mathscr{X}'$ according to the time stamp: $(t,u) \preceq (s,v)$ for any $t < s$ and $u, v \in \mathscr{X}'$. Suppose an ordered $\mathscr{X}$-valued tree $\mathbf{x}$ of depth d is shattered, according to Definition 2, with a witness tree $\mathbf{s}$. We claim that the values of the witness tree must be increasing by at least β along the path $y = (1,1,1,\ldots)$. Indeed, consider any $t \geq 1$, and let $y' = (y_{1:t}, 0, y_{t+2:d})$. By the definition of shattering, there must be a function that satisfies $f(\mathbf{x}_t(y')) \geq \mathbf{s}_t(y') + \beta/2$ and $f(\mathbf{x}_{t+1}(y')) \leq \mathbf{s}_{t+1}(y') - \beta/2$. Since $f(\mathbf{x}_t(y')) \leq f(\mathbf{x}_{t+1}(y'))$, we conclude that $\mathbf{s}_t(y) = \mathbf{s}_t(y') \leq \mathbf{s}_{t+1}(y') - \beta = \mathbf{s}_{t+1}(y) - \beta$. Hence, $\mathbf{s}_t$ increases by at least β along the path $(1,\ldots,1)$ and thus $d \leq 1/\beta$. This quick calculation gives $\mathrm{fat}^o_\beta(\mathscr{F},\mathscr{X}) \leq 1/\beta$.

In view of Theorem 6,

$$\log\mathscr{N}_\infty(\mathscr{F}^\delta, \beta, \mathbf{x}) \leq (1/\beta)\log\left(2en/\beta\right)$$

In view of (6), the same covering number estimate holds for $\mathscr{G}$. Then Theorem 4 with $\alpha = 1/n$ and $\gamma = n^{-a}$ (with a to be determined later) implies that

$$\int_\alpha^\gamma \log\mathscr{N}_\infty(\mathscr{G}, \rho, \mathbf{z})\, d\rho \leq C\log^2 n$$

is a lower order term, with C being an absolute constant. We also have

$$\int_{\alpha}^{\gamma} \sqrt{\log \mathcal{N}_\infty(\mathcal{G},\rho,\mathbf{z})} d\rho \le C' \sqrt{\log n} \cdot \gamma^{1/2}.$$

Now, ignoring constants and logarithmic terms, this gives the overall rate of

$$\mathcal{O}^*\left(\frac{1}{\delta\gamma} + \sqrt{\frac{n\gamma}{\delta}}\right) = \mathcal{O}^*\left(n^{1/3}\delta^{-2/3}\right)$$

for the minimax regret with respect to $\mathcal{F}^\delta$. The terms are balanced by choosing $\gamma = n^{-1/3}\delta^{-1/3}$. The rate with respect to $\mathcal{F}$ is then

$$\mathcal{O}^*\left(n\delta + n^{1/3}\delta^{-2/3}\right) = \mathcal{O}^*\left(n^{3/5}\right)$$

by choosing $\delta = n^{-2/5}$. This corresponds to the rate obtained by [4].

6 Example: Linear Prediction

In this section we consider the special case of $\mathcal{X}_1 = \ldots = \mathcal{X}_n = \mathcal{X} = B_2$ and

$$\mathcal{F} = \{f(x) = (\langle w, x\rangle + 1)/2 : w \in B_2\} \tag{9}$$

where B_2 is a unit Euclidean (or Hilbert) ball. Written as a function of w, the loss at time t is (up to an additive constant log(2))

$$g_t(w) = -\mathbf{1}\left\{y_t = 1\right\}\log(1+\langle w, x_t\rangle) - \mathbf{1}\left\{y_t = 0\right\}\log(1-\langle w, x_t\rangle). \tag{10}$$

It is possible to estimate the sequential fat_β dimension of a unit Hilbert ball as $\text{fat}_\beta = \mathcal{O}^*(1/\beta^2)$, where the $\mathcal{O}^*$ notation ignores logarithmic factors. Then Theorem 4 gives an upper bound of

$$V_n(\mathcal{F}^\delta) = \mathcal{O}^*\left(n^{1/2}\delta^{-1}\right),$$

and thus

$$V_n(\mathcal{F}) = \mathcal{O}^*\left(n^{3/4}\right).$$

Below, we exhibit an algorithm that attains regret of $\mathcal{O}^*\left(n^{1/2}\right)$, implying that the upper bounds obtained with our technique are not always tight.

6.1 Algorithm: Regularization with Self-Concordant Barrier

To develop an algorithm for the problem, we turn to the field of online convex optimization. We observe that functions g_t defined in (10) are convex, but not strongly convex. Moreover, the gradients of $g_t(w)$ are not bounded. We may consider a restricted set to mitigate the exploding gradient; however, a δ-shrinkage of the ball B_2 still leaves the gradient to be of size $O(1/\delta)$. A direct gradient descent method will give the suboptimal $O(n^{3/4})$ upper bound derived above in a non-constructive way. We also mention that while the functions are exp-concave, the upper bounds for the Online Newton Step method [6] scale with the dimension of the space, which we assume to be large or infinite.

We now present an algorithm based on self-concordant barrier regularization, which appears to be of an independent interest. The algorithm answers the following question: *can one obtain regret bounds for online convex optimization in terms of the maximum of function values rather than gradients?*

Consider the Follow-the-Regularized-Leader method

$$w_{t+1} = \underset{w\in B_2}{\operatorname{argmin}} \sum_{s=1}^{t} \left\langle \nabla g_s(w_s), w\right\rangle + \eta^{-1}R(w) \tag{11}$$

with the self-concordant barrier $R(w) = -\log(1 - \|w\|^2)$. In accordance with the protocol of the probability assignment problem, we predict $\langle w_t, x_t \rangle$ at round t after observing x_t. It is shown in [2] that regret of (11) against any $w^* \in B_2$ is

$$\sum_{t=1}^{n} g_t(w_t) - g_t(w^*) \leq 2\eta \sum_{t=1}^{n} \|\nabla g_t(w_t)\|_{w_t}^{*2} + \eta^{-1} R(w^*) \tag{12}$$

as long as η satisfies $\eta \|\nabla g_t(w_t)\|_{w_t}^* \leq 1/4$. Here, the *local norm* is defined as

$$\|h\|_w^* = \sqrt{h^\mathsf{T} (\nabla^2 R(w))^{-1} h}.$$

According to the lemma below, the local norm is bounded by a constant that is independent of the dimension:

Lemma 7. *For any t, the local norm of $\nabla g_t(w_t)$ is upper bounded by a constant:*

$$\|\nabla g_t(w_t)\|_{w_t}^* \leq 3.$$

Together with (12), Lemma 7 implies a regret bound of $18\eta n + \eta^{-1} R(w^*)$. Instead of taking w^* at the boundary of the ball where $R(w^*)$ is infinite, we can evaluate regret against $w = (1 - 1/n) w^*$. For such a comparator, $R(w) = \mathcal{O}(\log n)$. By choosing η appropriately and using an argument similar to (2), we conclude that regret against any $w^* \in B_2$ is upper bounded by

$$C\sqrt{n \log n}.$$

Importantly, C is an absolute constant that does not depend on the dimension of the problem. This rate is optimal up to polylogarithmic factors. The optimality follows from Lemma 10 below and an estimate on sequential covering number of a Hilbert ball [11, 12].

Lemma 8. *For the linear class in (9),*

$$V_n(\mathcal{F}) = \Theta^*(n^{1/2}).$$

The proof of Lemma 7 relied heavily on the ability to calculate the gradient of the loss function and match it to the inverse Hessian of the self-concordant barrier. We now give an alternative proof based on a simple and charming, yet unexpected lemma due to Nesterov (see Appendix for the short proof):

Lemma 9 (Lemma 4 in [8])**.** *Let ψ be concave and positive on int $\mathcal{K}$. Then for any $x \in$ int $\mathcal{K}$ we have*

$$\|\nabla \psi(x)\|_x^* \leq \psi(x).$$

The lemma allows us to upper bound regret in an online convex optimization problem if we only know that the values of the functions (and not the gradients) are bounded. Consider the FTRL algorithm (11), but over the shrunk ball $(1 - 1/n)B_2$. Suppose we can ensure $0 < g_t < A$. Then $A - g_t$ is concave and positive. Hence, by above lemma

$$\|\nabla g_t(w_t)\|_{w_t}^* = \|\nabla (A - g_t(w_t))\|_{w_t}^* \leq A - g_t(w_t) \leq A$$

which provides an alternative to the bound of Lemma 7. Regret is then upper bounded by

$$\sum_{t=1}^{n} g_t(w_t) - \sum_{t=1}^{n} g_t(w^*) \leq 2\eta n A^2 + \eta^{-1} R(w^*)$$

Crucially, by employing self-concordant regularization, we avoid paying for a large gradient of cost functions at the boundary of the set. Over the shrunk set $(1 - 1/n)B_2$, we ensure that the values of functions g_t are upper bounded by $A = O(\log n)$ even if the gradients blow up linearly with n. This surprising observations leads to a dimension-independent $O(\sqrt{n} \log n)$ regret bound for the Euclidean ball, and can also be used for other convex bodies and non-logarithmic loss functions when the closed-form analysis of Lemma 7 is not available.

7 A Lower Bound

In this section, we show that the offset sequential Rademacher complexity serves as a lower bound on the minimax regret. Hence, the complexities of the class $\mathscr{F}$ of experts are intrinsic to the problem. We refer to [12] for further lower bounds on the offset Rademacher complexity via the scale-sensitive dimension and sequential covering numbers.

Lemma 10. *The following lower bound holds:*

$$V_n(\mathscr{F}) + 1 \geq \sup_{\mathbf{x}} \mathbb{E}_y \left[\sup_{f \in \mathscr{F}^{1/n}} \left\{ \sum_{t=1}^{n} 2(2y_t - 1)(f(\mathbf{x}_t(y)) - 1/2) - 4(\log n)(f(\mathbf{x}_t(y)) - 1/2)^2 \right\} \right]$$

where $y_1, \ldots, y_n$ are independent with distribution $\mathsf{Bernoulli}(1/2)$ *and the supremum is taken over consistent trees with respect to constraints $\mathscr{X}_t$.*

Proof of Lemma 10. To prove the lower bound, we proceed as in [12]. First, we observe that (2) holds in the other direction too:

$$V_n(\mathscr{F}^{\delta}) \leq V_n(\mathscr{F}) + n\delta. \tag{13}$$

To see this, note that any f only loses from thresholding when either $f(x_t) > 1-\delta$ and $y_t = 1$, or when $f(x_t) < \delta$ and $y_t = 0$. In both cases, the difference in logarithmic loss is at most $-\log(1-\delta) \leq \delta$ for $\delta < 1/2$. For the purposes of a lower bound, we take $\delta = 1/n$ and turn to lower-bounding $V_n(\mathscr{F}^{1/n})$.

As in the development leading to (25) in the proof of the upper bound, the minimax value $V_n(\mathscr{F}^{1/n})$ is *equal* to

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in [0,1]} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}^{1/n}} \left\{ \sum_{t=1}^{n} \inf_{\widehat{y}_t \in [0,1]} \mathbb{E}_{y_t}\left[\boldsymbol{\ell}(\widehat{y}_t, y_t)\right] - \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \right\} \right] \tag{14}$$

which, by the self-information property of the loss equal to

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in [0,1]} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}^{1/n}} \left\{ \sum_{t=1}^{n} \mathbb{E}_{y_t}\left[\boldsymbol{\ell}(p_t, y_t)\right] - \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \right\} \right] \tag{15}$$

By the linearity of expectation (and since the terms $\mathbb{E}_{y_t}\left[\boldsymbol{\ell}(p_t, y_t)\right]$ do not involve f), we have

$$V_n(\mathscr{F}^{1/n}) = \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in [0,1]} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}^{1/n}} \left\{ \sum_{t=1}^{n} \boldsymbol{\ell}(p_t, y_t) - \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \right\} \right]. \tag{16}$$

We now pass to the first lower bound by choosing $p_t = 1/2$ for all t.

Consider the case $y_t = 1$ and expand the loss function around $p_t = 1/2$ for $z \in [1/n, 1]$:

$$\boldsymbol{\ell}(1/2, 1) - \boldsymbol{\ell}(z, 1) = -\log(1/2) - (-\log(z)) = 2(z - 1/2) - R(z) \tag{17}$$

where $R(z)$ is the remainder. We claim that the remainder can be upper bounded by a quadratic over the interval $[1/n, 1]$. To this end, consider the function

$$g(z) = -2z + (1 + \log(2)) + 4(\log n)(z - 1/2)^2$$

and note that the derivative and the value of this function at $1/2$ coincide with the derivative and the value of $-\log(z)$ at the same point. We claim that $g(z)$ dominates $-\log(z)$ on $[1/n, 1]$. For $z > 1/2$, this follows from $g' > (-\log)'$. The same argument holds for the interval $[1/\log(n), 1/2]$. Now, at $z = 1/n$, $g(z) > -\log(z)$ and $|g'(z)| < |\log(z)'|$. The derivative relation continues to hold on the interval $[1/n, c/\log(n)]$ for large enough c, establishing $g > -\log$ on this interval too. The remaining interval $[c/\log(n), 1/\log(n)]$ is easily checked by the direct computation of function value. In sum, the remainder in (17) can be upper bounded by $R(z) \leq 4(\log n)(z - 1/2)^2$.

The case of $y_t = 0$ is exactly analogous, and we obtain

$$\ell(p_t, y_t) - \sum_{t=1}^{n} \ell(f(x_t), y_t) \geq 2\left[\mathbf{1}\left\{y_t = 1\right\}(f(x_t) - 1/2) + \mathbf{1}\left\{y_t = 0\right\}(-f(x_t) + 1/2)\right] - 4(\log n)(f(x_t) - 1/2)^2 \tag{18}$$

$$= 2\left[y_t(f(x_t) - 1/2) + (1 - y_t)(-f(x_t) + 1/2)\right] - 4(\log n)(f(x_t) - 1/2)^2 \tag{19}$$

$$= 2(2y_t - 1)(f(x_t) - 1/2) - 4(\log n)(f(x_t) - 1/2)^2. \tag{20}$$

The lower bound in (21) then becomes

$$V_n(\mathscr{F}^{1/n}) \geq \left\langle\!\!\left\langle \sup_{x_t \in \mathscr{X}_t(x_{1:t-1}, y_{1:t-1})} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}^{1/n}} \left\{ \sum_{t=1}^{n} 2(2y_t - 1)(f(x_t) - 1/2) - 4(\log n)(f(x_t) - 1/2)^2 \right\} \right] \tag{21}$$

$$= \sup_{\mathbf{x}} \mathbb{E}_y \left[\sup_{f \in \mathscr{F}^{1/n}} \left\{ \sum_{t=1}^{n} 2(2y_t - 1)(f(\mathbf{x}_t(y)) - 1/2) - 4(\log n)(f(\mathbf{x}_t(y)) - 1/2)^2 \right\} \right] \tag{22}$$

where $y_1, \ldots, y_n$ are independent with distribution $\mathsf{Bernoulli}(1/2)$.

□

8 Discussion and Open Questions

At the very first step, the analysis in this paper thresholds the class $\mathscr{F}$ to avoid dealing with the exploding gradient of the loss function. The authors believe that this "hard thresholding" approach is the source of sub-optimality, and that "smooth" approaches should be possible. When the class of functions has a specific structure, such as in the example of Section 6, the exploding gradient can be mitigated in a "smooth way" by a regularization technique. It is not clear to the authors how to perform the "smooth thresholding" analysis when such a structure is not available.

Another interesting venue of investigation is the development of algorithms. It has been shown that the minimax analysis, of the type performed in this paper, can be made constructive [9, 12, 10]. It appears that the relaxation approach may yield new (and possibly computationally efficient) methods for sequential probability assignment and data compression.

A Proofs

Proof of Theorem 1. Let us use the shorthand $\mathscr{D} = [\delta, 1 - \delta]$. The value $V_n(\mathscr{F}^{\delta})$ can be upper bounded by

$$\left\langle\!\!\left\langle \sup_{x_t} \inf_{\widehat{y}_t \in \mathscr{D}} \sup_{p_t \in [0,1]} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathscr{F}^{\delta}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right\} \tag{23}$$

simply because each infimum is taken over a smaller set. Henceforth, it will be understood that x_t ranges over $\mathscr{X}_t(x_{1:t-1}, y_{1:t-1})$. The expression in (23) is equal to

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in [0,1]} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sum_{t=1}^{n} \inf_{\widehat{y}_t \in \mathscr{D}} \mathbb{E}_{y_t}\left[\ell(\widehat{y}_t, y_t)\right] - \inf_{f \in \mathscr{F}^{\delta}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right\} \tag{24}$$

by an argument that can be found in [1, 13]. Here, it is understood that y_t is a Bernoulli random variable with distribution p_t. Taking the infimum outside the negative sign, the above quantity is equal to

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in [0,1]} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}^{\delta}} \left\{ \sum_{t=1}^{n} \inf_{\widehat{y}_t \in \mathscr{D}} \mathbb{E}_{y_t}\left[\ell(\widehat{y}_t, y_t)\right] - \sum_{t=1}^{n} \ell(f(x_t), y_t) \right\} \right] \tag{25}$$

We now claim that each infimum in (25) is achieved at $\widehat{y}_t = \tau_{\delta}(p_t)$. Indeed, this follows because the unconstrained minimizer over $[0,1]$ is p_t by the well-known property of entropy:

$$\operatorname*{argmin}_{\widehat{y}_t \in [0,1]} \mathbb{E}_{y_t}\left[\ell(\widehat{y}_t, y_t)\right] = \operatorname*{argmin}_{\widehat{y}_t \in [0,1]} \left\{ -p_t \log(\widehat{y}_t) - (1 - p_t)\log(1 - \widehat{y}_t) \right\} = p_t.$$

We conclude that (25) is equal to

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in [0,1]} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}^\delta} \left\{ \sum_{t=1}^{n} \mathbb{E}_{y_t} \left[\boldsymbol{\ell}(\tau_\delta(p_t), y_t) \right] - \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \right\} \right]. \tag{26}$$

Now, the terms in the first sum do not depend on $f \in \mathscr{F}$, and thus can pass through the multiple infima and suprema. By linearity of expectation, (26) is equal to

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in [0,1]} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}^\delta} \left\{ \sum_{t=1}^{n} \boldsymbol{\ell}(\tau_\delta(p_t), y_t) - \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \right\} \right] \tag{27}$$

We now separately deal with the case that $p_t \notin \mathscr{D}$. To this end, observe that

$$\begin{aligned}
\mathbf{1}\left\{p_t < \delta\right\}(\boldsymbol{\ell}(\tau_\delta(p_t), y_t) - \boldsymbol{\ell}(f(x_t), y_t)) &= \mathbf{1}\left\{p_t < \delta, y_t = 0\right\}(\boldsymbol{\ell}(\delta, y_t) - \boldsymbol{\ell}(f(x_t), y_t)) \\
&\quad + \mathbf{1}\left\{p_t < \delta, y_t = 1\right\}(\boldsymbol{\ell}(\delta, y_t) - \boldsymbol{\ell}(f(x_t), y_t)) \\
&\le \mathbf{1}\left\{p_t < \delta, y_t = 1\right\}(\boldsymbol{\ell}(\delta, 1) - \boldsymbol{\ell}(f(x_t), 1)) \\
&\le -\mathbf{1}\left\{p_t < \delta, y_t = 1\right\}\log\delta
\end{aligned}$$

The first inequality is obtained by dropping the non-positive term. Indeed, $p_t < \delta$ gives higher odds to the outcome $y_t = 0$ than $f(x_t) \ge \delta$. Positivity of $\boldsymbol{\ell}$ gives the second inequality. A similar calculation gives

$$\mathbf{1}\left\{p_t > 1-\delta\right\}(\boldsymbol{\ell}([p_t], y_t) - \boldsymbol{\ell}(f(x_t), y_t)) \le -\mathbf{1}\left\{p_t > 1-\delta, y_t = 0\right\}\log\delta$$

Substituting into (27), we obtain an upper bound of

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in [0,1]} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}^\delta} \left\{ \sum_{t=1}^{n} \mathbf{1}\left\{p_t \in \mathscr{D}\right\}(\boldsymbol{\ell}(\tau_\delta(p_t), y_t) - \boldsymbol{\ell}(f(x_t), y_t)) - \mathbf{1}\left\{p_t < \delta, y_t = 1\right\}\log\delta - \mathbf{1}\left\{p_t > 1-\delta, y_t = 0\right\}\log\delta \right\} \right]$$

Since

$$\mathbb{E}_{y_t \sim p_t}\mathbf{1}\left\{p_t < \delta, y_t = 1\right\}\log(1/\delta) \le \delta\log(1/\delta),$$

and since $\mathbf{1}\left\{p_t \in \mathscr{D}\right\}(\boldsymbol{\ell}(\tau_\delta(p_t), y_t) = \mathbf{1}\left\{p_t \in \mathscr{D}\right\}(\boldsymbol{\ell}(p_t, y_t)$, we conclude that the minimax value $V_n(\mathscr{F}^\delta)$ is upper bounded by

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in [0,1]} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}^\delta} \left\{ \sum_{t=1}^{n} \mathbf{1}\left\{p_t \in \mathscr{D}\right\}(\boldsymbol{\ell}(p_t, y_t) - \boldsymbol{\ell}(f(x_t), y_t)) \right\} \right] + 2n\delta\log(1/\delta). \tag{28}$$

We now linearize the terms $\boldsymbol{\ell}(p_t, y_t) - \boldsymbol{\ell}(f(x_t), y_t)$. The derivative of $\boldsymbol{\ell}(\cdot, y_t)$ at p_t is

$$\boldsymbol{\ell}'(p_t, y_t) = -\mathbf{1}\left\{y_t = 1\right\}\frac{1}{p_t} + \mathbf{1}\left\{y_t = 0\right\}\frac{1}{1-p_t}.$$

Observe that the second derivative $\boldsymbol{\ell}''(\cdot, y_t) \ge 1$, and hence $\boldsymbol{\ell}(\cdot, y_t)$ is strongly convex, for either value of y_t. Strong convexity implies that

$$\boldsymbol{\ell}(p_t, y_t) - \boldsymbol{\ell}(f(x_t), y_t) \le \boldsymbol{\ell}'(p_t, y_t) \cdot (p_t - f(x_t)) - \frac{1}{2}(p_t - f(x_t))^2$$

and thus (28) is upper bounded by

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in [0,1]} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}^\delta} \sum_{t : p_t \in \mathscr{D}} \boldsymbol{\ell}'(p_t, y_t) \cdot (p_t - f(x_t)) - \frac{1}{2}(p_t - f(x_t))^2 \right] + 2n\delta\log(1/\delta). \tag{29}$$

Observe that the derivatives are mean-zero:

$$\mathbb{E}_{y_t\sim p_t}\boldsymbol{\ell}'(p_t,y_t)=\mathbb{E}\left[-\mathbf{1}\left\{y_t=1\right\}\frac{1}{p_t}+\mathbf{1}\left\{y_t=0\right\}\frac{1}{1-p_t}\right]=0, \tag{30}$$

which suggests that we can symmetrize these terms as in [11, 12]. The key observation is that tighter control on the supremum over $\mathscr{F}$ will be obtained if we keep the derivatives to have a non-uniform distribution given by p_t.

Let us drop the term $2n\delta\log(1/\delta)$ in (29) and concentrate on the first term. Consider the following upper bound:

$$\begin{aligned}&\left\langle\!\!\left\langle \sup_{x_t}\sup_{p_t\in[0,1]}\mathbb{E}_{y_t}\right\rangle\!\!\right\rangle_{t=1}^{n}\left[\sup_{f\in\mathscr{F}^\delta}\sum_{t:p_t\in\mathscr{D}}\boldsymbol{\ell}'(p_t,y_t)\cdot\left(p_t-f(x_t)\right)-\frac{1}{2}(p_t-f(x_t))^2\right]\\ &\le\left\langle\!\!\left\langle \sup_{x_t}\sup_{p_t,p'_t\in[0,1]}\mathbb{E}_{y_t\sim p'_t}\right\rangle\!\!\right\rangle_{t=1}^{n}\left[\sup_{f\in\mathscr{F}^\delta}\sum_{t:p_t\in\mathscr{D}}\boldsymbol{\ell}'(p_t,y_t)\cdot\left(p'_t-f(x_t)\right)-\frac{1}{2}(p_t-f(x_t))^2\right]\end{aligned} \tag{31}$$

This upper bound holds because the supremum allows the choice $p_t=p'_t$ in addition to distinct choices for the two distributions.

We now pass to the tree notation. Observe that the optimal choice of x_t,p_t,p'_t depends on $(y_1,\dots,y_{t-1})\in\{0,1\}^{t-1}$. In the functional form, let $\mathbf{x}$ be a sequence of mappings $\mathbf{x}_1,\dots,\mathbf{x}_n$ with the consistency property $\mathbf{x}_t(y_1,\dots,y_{t-1})\in\mathscr{X}_t(\mathbf{x}_1(y),\dots,\mathbf{x}_{t-1}(y),y_{1:t-1})$ for all $y_{1:t-1}$. Similarly, let $\boldsymbol{\mu}$ and $\mathbf{p}$ be sequences of mappings with $\boldsymbol{\mu}_t,\mathbf{p}_t:\{0,1\}^{t-1}\to[0,1]$. With the same reasoning as in [13], we can write (31) as

$$\sup_{\mathbf{x},\boldsymbol{\mu},\mathbf{p}}\mathbb{E}\sup_{f\in\mathscr{F}^\delta}\left[\sum_{t:\mathbf{p}_t(y)\in\mathscr{D}}\boldsymbol{\ell}'(\mathbf{p}_t(y),y_t)\left(\boldsymbol{\mu}_t(y)-f(\mathbf{x}_t(y))\right)-\frac{1}{2}\left(\boldsymbol{\mu}_t(y)-f(\mathbf{x}_t(y))\right)^2\right]$$

where y_t's in $\{0,1\}$ are drawn from $\mathbf{p}$. More specifically, $y_1\sim\mathbf{p}_1$ and subsequently $y_t\sim\mathbf{p}_t(y_{1:t-1})$. □

Proof of Lemma 2.

$$\begin{aligned}\mathbb{E}\sup_{\mathbf{v}\in V}\left[\sum_{t=1}^{n}\eta(\mathbf{p}_t(y),y_t)\mathbf{v}_t(y)-c\mathbf{v}_t(y)^2\right]&=\mathbb{E}\inf_{\lambda>0}\frac{1}{\lambda}\log\left(\sum_{\mathbf{v}\in V}\exp\left(\lambda\sum_{t=1}^{n}\eta(\mathbf{p}_t(y),y_t)\mathbf{v}_t(y)-c\mathbf{v}_t(y)^2\right)\right)\\ &\le\inf_{\lambda>0}\frac{1}{\lambda}\log\left(\sum_{\mathbf{v}\in V}\mathbb{E}\prod_{t=1}^{n}\exp\left(\lambda\left(\eta(\mathbf{p}_t(y),y_t)\mathbf{v}_t(y)-c\mathbf{v}_t(y)^2\right)\right)\right).\end{aligned} \tag{32}$$

Let X be a zero-mean random variable taking on a value $-v/p$ with probability p and $v/(1-p)$ with probability $(1-p)$, where $\delta<p<1/2$ and $|v|\le1$. From the fact that $(e^x-x-1)/x^2$ is a non-decreasing function and $|X|<1/\delta$ almost surely, it follows that

$$e^{\lambda X}-\lambda X-1\le\delta^2X^2\left(e^{\lambda/\delta}-\lambda/\delta-1\right).$$

Taking expectation over X and upper bounding the variance $\mathbb{E}X^2\le2p(v/p)^2\le2v^2/\delta$,

$$\mathbb{E}e^{\lambda X}-1\le2v^2\delta\left(e^{\lambda/\delta}-\lambda/\delta-1\right).$$

Using $1+x\le e^x$,

$$\mathbb{E}e^{\lambda X}\le\exp\left\{2v^2\delta\left(e^{\lambda/\delta}-\lambda/\delta-1\right)\right\}.$$

Applying the above derivation,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\left(\eta(\mathbf{p}_t(y),y_t)\mathbf{v}_t(y)-c\mathbf{v}_t(y)^2\right)\right)\mid y_1,\dots,y_{t-1}\right]&=\exp\left(-\lambda c\mathbf{v}_t(y)^2\right)\times\mathbb{E}\left[\exp\left(\lambda\eta(\mathbf{p}_t(y),y_t)\mathbf{v}_t(y)\right)\mid y_1,\dots,y_{t-1}\right]\\ &\le\exp\left(-\lambda c\mathbf{v}_t(y)^2\right)\times\exp\left\{2\mathbf{v}_t(y)^2\delta\left(e^{\lambda/\delta}-\lambda/\delta-1\right)\right\}\\ &=\exp\left(2\delta\,\mathbf{v}_t(y)^2\left(e^{\frac{\lambda}{\delta}}-1-\left(1+\frac{c}{2}\right)\frac{\lambda}{\delta}\right)\right).\end{aligned}$$

Choosing $\lambda = \log(1+\frac{c}{2})\delta$ we ensure that $\left(e^{\frac{\lambda}{\delta}} - 1 - \frac{(2+c)\lambda}{2\delta}\right) < 0$ and

$$\mathbb{E}\left[\exp\left(\lambda\left(\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y) - c\mathbf{v}_t(y)^2\right)\right) \mid y_1, \dots, y_{t-1}\right] \le 1.$$

Iterating the argument from $t = n$ down to $t = 1$ in (32), we obtain

$$\mathbb{E}\sup_{\mathbf{v}\in V}\left[\sum_{t=1}^{n}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y) - c\mathbf{v}_t(y)^2\right] \le \frac{\log|V|}{\delta\log(1+\frac{c}{2})}.$$

The case when $\mathbf{p}$ is $[0,1]$-valued, but the summation is taken only over $\{t : \mathbf{p}_t(y) \in [\delta, 1-\delta]\}$, follows immediately through the same argument. □

Proof of Lemma 3. Both sides of the inequality in the statement of the Lemma are homogenous with respect to $v_{\max}$, and so we can assume $v_{\max} = 1$ and rescale the problem. We have

$$\begin{aligned}\mathbb{E}_y\max_{\mathbf{v}\in V}\left[\sum_{t=1}^{n}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y)\right] &\le \inf_{\lambda>0}\left\{\frac{1}{\lambda}\log\sum_{\mathbf{v}\in V}\mathbb{E}\exp\left(\lambda\sum_{t=1}^{n}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y)\right)\right\} \\ &\le \inf_{\lambda>0}\left\{\frac{\log|V|}{\lambda} + \max_{\mathbf{v}\in V}\frac{1}{\lambda}\log\mathbb{E}\exp\left(\lambda\sum_{t=1}^{n}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y)\right)\right\}. \qquad (33)\end{aligned}$$

As shown in the proof of Lemma 2, if X is a zero-mean random variable taking on a value $-v/p$ with probability p and $v/(1-p)$ with probability $(1-p)$, where $\delta < p < 1/2$ and $|v| \le 1$, then

$$\log\mathbb{E}e^{\lambda X} \le 2v^2\delta\phi(\lambda/\delta)$$

where $\phi(x) = e^x - x - 1$. Hence,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{n}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y)\right)\ \Big|y_1, \dots, y_{n-1}\right] &\le \exp\left(\lambda\sum_{t=1}^{n-1}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y)\right) \times \mathbb{E}\left[\exp\left(\lambda\eta(\mathbf{p}_n(y), y_n)\mathbf{v}_n(y)\right)\ \big|y_1, \dots, y_{n-1}\right] \\ &\le \exp\left(\lambda\sum_{t=1}^{n-1}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y)\right) \times \exp\left\{2\delta\phi(\lambda/\delta)\max_{y_{n-1}}\mathbf{v}_n(y)^2\right\}\end{aligned}$$

For y_{n-1}, we proceed in a similar fashion:

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{n-1}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y)\right) \times \exp\left\{2\delta\phi(\lambda/\delta)\max_{y_{n-1}}\mathbf{v}_n(y)^2\right\}\ \Big|y_1, \dots, y_{n-2}\right] \\ &\le \exp\left(\lambda\sum_{t=1}^{n-2}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y)\right) \times \mathbb{E}\left[\exp\left\{\lambda\eta(\mathbf{p}_{n-1}(y), y_{n-1})\mathbf{v}_{n-1}(y) + 2\delta\phi(\lambda/\delta)\max_{y_{n-1}}\mathbf{v}_n(y)^2\right\}\ \Big|y_1, \dots, y_{n-2}\right] \\ &\le \exp\left(\lambda\sum_{t=1}^{n-2}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y)\right) \times \exp\left\{2\delta\phi(\lambda/\delta)\mathbf{v}_{n-1}(y)^2 + 2\delta\phi(\lambda/\delta)\max_{y_{n-1}}\mathbf{v}_n(y)^2\right\} \\ &\le \exp\left(\lambda\sum_{t=1}^{n-2}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y)\right) \times \exp\left\{2\delta\phi(\lambda/\delta)\max_{y_{n-2}, y_{n-1}}\left\{\mathbf{v}_{n-1}(y)^2 + \mathbf{v}_n(y)^2\right\}\right\}\end{aligned}$$

Unrolling the expression to $t = 1$ we obtain

$$\log\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{n}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y)\right)\right] \le 2\delta\phi(\lambda/\delta)n\bar{v}^2$$

where $\bar{v}^2 = \max_{\mathbf{v}\in V}\max_y\frac{1}{n}\sum_{t=1}^{n}\mathbf{v}_t(y)^2$. In view of (33), we get

$$\mathbb{E}_y\max_{\mathbf{v}\in V}\left[\sum_{t=1}^{n}\eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y)\right] \le \inf_{\lambda>0}\left\{\frac{\log|V|}{\lambda} + \frac{2\delta\phi(\lambda/\delta)n\bar{v}^2}{\lambda}\right\}. \qquad (34)$$

First, consider the case $\delta \ge \frac{\log|V|}{4n\bar{v}^2}$. Then the choice $\lambda = \frac{1}{2}\sqrt{\frac{\delta\log|V|}{n\bar{v}^2}}$ ensures $\lambda \le \delta$. In this case, $\phi(\lambda/\delta)$ can be upper bounded by a quadratic $\phi(\lambda/\delta) \le (\lambda/\delta)^2 \cdot e$. The upper bound in (34) becomes

$$\frac{\log|V|}{\lambda} + \frac{2e(\lambda/\delta)^2\delta n\bar{v}^2}{\lambda} \le (2+e)\sqrt{\frac{n\bar{v}^2\log|V|}{\delta}}.$$

On the other hand, if $\delta < \frac{\log|V|}{4n\bar{v}^2}$, the upper bound in (34) becomes

$$\inf_{\lambda>0}\left\{\frac{\log|V|}{\lambda} + \frac{2\log|V|\phi(\lambda/\delta)}{4\lambda}\right\}.$$

Choosing $\lambda = \delta$ yields an upper bound of $\frac{2\log|V|}{\delta}$. Combining the two cases, we arrive at the statement of the Lemma. □

Proof of Theorem 4. Let us use the shorthand $\mathscr{D} = [\delta, 1-\delta]$. Let V' be a sequential γ-cover of $\mathscr{G}$ on $\mathbf{z}$ in the ℓ_∞ sense, i.e.

$$\forall y \in \{0,1\}^n,\ \ \forall g \in \mathscr{G},\ \ \exists \mathbf{v} \in V' \ \ \text{s.t.}\ \ |g(\mathbf{z}_t(y)) - \mathbf{v}_t(y)| \le \gamma.$$

Of course, an ℓ_∞ cover is also an ℓ_2 cover at the same scale. Let us augment V' to include the all-zero tree, and denote the resulting set by $V = V' \cup \{\mathbf{0}\}$. Denote by $\mathbf{v}[\epsilon, g]$ a γ-close tree promised above. We have

$$\begin{aligned}
&\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t:\mathbf{p}_t(y)\in\mathscr{D}} \eta(\mathbf{p}_t(y), y_t) g(\mathbf{z}_t(y)) - Kg(\mathbf{z}_t(y))^2\right] \qquad (35)\\
&= \mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t:\mathbf{p}_t(y)\in\mathscr{D}} \eta(\mathbf{p}_t(y), y_t)\Big(g(\mathbf{z}_t(y)) - \mathbf{v}[y,g]_t(y)\Big) - K\Big(g(\mathbf{z}_t(y))^2 - \frac{1}{4}\mathbf{v}[y,g]_t(y)^2\Big)\right. \qquad (36)\\
&\qquad\qquad \left. + \Big(\eta(\mathbf{p}_t(y), y_t)\mathbf{v}[y,g]_t(y) - \frac{K}{4}\mathbf{v}[y,g]_t(y)^2\Big)\right] \qquad (37)\\
&\le \mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t:\mathbf{p}_t(y)\in\mathscr{D}} \eta(\mathbf{p}_t(y), y_t)\Big(g(\mathbf{z}_t(y)) - \mathbf{v}[y,g]_t(y)\Big) - K\Big(g(\mathbf{z}_t(y))^2 - \frac{1}{4}\mathbf{v}[y,g]_t(y)^2\Big)\right] \qquad (38)\\
&+ \mathbb{E}\max_{\mathbf{v}\in V'}\left[\sum_{t:\mathbf{p}_t(y)\in\mathscr{D}} \eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y) - \frac{K}{4}\mathbf{v}_t(y)^2\right] \qquad (39)
\end{aligned}$$

We now claim that for any y and g there exists an element $\mathbf{v}[y,g] \in V$ such that

$$\sum_{t=1}^n g(\mathbf{z}_t(y))^2 \ge \frac{1}{4}\sum_{t=1}^n \mathbf{v}[y,g]_t(y)^2 \tag{40}$$

and so we can drop the corresponding negative term in the supremum over $\mathscr{G}$. First consider the easy case $\frac{1}{n}\sum_{t=1}^n g(\mathbf{z}_t(\epsilon))^2 \le \gamma^2$. Then we may choose $\mathbf{0} \in V$ as a tree that provides a sequential γ-cover in the ℓ_2 sense. Clearly, (40) is then satisfied with this choice of $\mathbf{v}[\epsilon, g] = \mathbf{0}$. Now, assume $\frac{1}{n}\sum_{t=1}^n g(\mathbf{z}_t(\epsilon))^2 > \gamma^2$. Fix any tree $\mathbf{v}[\epsilon, g] \in V$ that is γ-close in the ℓ_2 sense to g on the path ϵ. Denote $u = (\mathbf{v}[\epsilon,g]_1(\epsilon), \ldots, \mathbf{v}[\epsilon,g]_n(\epsilon))$ and $h = (g(\mathbf{z}_1(\epsilon)), \ldots, g(\mathbf{z}_n(\epsilon)))$. Thus, we have that $\|u-h\| \le \gamma$ and $\|h\| \ge \gamma$ for the norm $\|h\|^2 = \frac{1}{n}\sum_{t=1}^n h_t^2$. Then

$$\|u\| \le \|u-h\| + \|h\| \le \gamma + \|h\| \le 2\|h\|$$

and thus $\|h\| \ge \frac{1}{2}\|u\|$ as desired. We conclude that

$$\begin{aligned}
&\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t:\mathbf{p}_t(y)\in\mathscr{D}} \eta(\mathbf{p}_t(y), y_t) g(\mathbf{z}_t(y)) - Kg(\mathbf{z}_t(y))^2\right] \qquad (41)\\
&\le \mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t:\mathbf{p}_t(y)\in\mathscr{D}} \eta(\mathbf{p}_t(y), y_t)\Big(g(\mathbf{z}_t(y)) - \mathbf{v}[y,g]_t(y)\Big)\right] + \mathbb{E}\max_{\mathbf{v}\in V'}\left[\sum_{t:\mathbf{p}_t(y)\in\mathscr{D}} \eta(\mathbf{p}_t(y), y_t)\mathbf{v}_t(y) - \frac{K}{4}\mathbf{v}_t(y)^2\right]
\end{aligned}$$

By Lemma 2, the second term is upper bounded by

$$\frac{\log \mathcal{N}_\infty(\mathcal{G},\gamma,\mathbf{z})}{\delta \log(1+\frac{K}{8})}$$

As for the second term, we note that conditionally on $y_1,\ldots,y_{t-1}$, the random variable $\eta(\mathbf{p}_t(y),y_t)$ is zero-mean. Let us proceed with the chaining technique. To this end, let $\mathbf{v}[g,y]^j \in V^j$ be an element of a $\gamma 2^{-j}$-cover of $g \in \mathcal{G}$.

$$\begin{aligned}
&\mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t:\mathbf{p}_t(y)\in\mathcal{D}}\eta(\mathbf{p}_t(y),y_t)\Big(g(\mathbf{z}_t(y))-\mathbf{v}[y,g]_t(y)\Big)\right] && (42)\\
&\le \sum_{j=1}^{N}\mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t:\mathbf{p}_t(y)\in\mathcal{D}}\eta(\mathbf{p}_t(y),y_t)\Big(\mathbf{v}[y,g]_t^j(y)-\mathbf{v}[y,g]_t^{j-1}(y)\Big)\right] && (43)\\
&+\mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t:\mathbf{p}_t(y)\in\mathcal{D}}\eta(\mathbf{p}_t(y),y_t)\Big(g(\mathbf{z}_t(y))-\mathbf{v}[y,g]_t^N(y)\Big)\right] && (44)
\end{aligned}$$

For the last term we use the Cauchy-Schwartz inequality: for any y and $g\in\mathcal{G}$,

$$\begin{aligned}
\sum_{t:\mathbf{p}_t(y)\in\mathcal{D}}\eta(\mathbf{p}_t(y),y_t)\Big(g(\mathbf{z}_t(y))-\mathbf{v}[y,g]_t^N(y)\Big) &\le \left(\sum_{t:\mathbf{p}_t(y)\in\mathcal{D}}\eta(\mathbf{p}_t(y),y_t)^2\right)^{1/2}\left(\sum_{t:\mathbf{p}_t(y)\in\mathcal{D}}\Big(g(\mathbf{z}_t(y))-\mathbf{v}[y,g]_t^N(y)\Big)^2\right)^{1/2} && (45)\\
&\le \frac{1}{\delta}n\gamma 2^{-N} && (46)
\end{aligned}$$

Further, for any $j=1,\ldots,N$,

$$\mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t:\mathbf{p}_t(y)\in\mathcal{D}}\eta(\mathbf{p}_t(y),y_t)\Big(\mathbf{v}[y,g]_t^j(y)-\mathbf{v}[y,g]_t^{j-1}(y)\Big)\right]\le \mathbb{E}\max_{\mathbf{w}\in W^j}\left[\sum_{t:\mathbf{p}_t(y)\in\mathcal{D}}\eta(\mathbf{p}_t(y),y_t)\mathbf{w}_t(y)\right]$$

where W^j is defined as the set of difference trees, defined as follows. For each pair $\mathbf{v}'\in V^j,\mathbf{v}''\in V^{j-1}$, let $\mathbf{w}$ be defined for each path $(y_1,\ldots,y_n)\in\{0,1\}^n$ and $t\in\{1,\ldots,n\}$ as

$$\mathbf{w}_t(y)=\begin{cases}\mathbf{v}'_t(y)-\mathbf{v}''_t(y), & \text{if exists } (y'_t,\ldots,y'_n) \text{ s.t. } \exists g\in\mathcal{G} \text{ s.t. } \mathbf{v}'=\mathbf{v}[g,\bar{y}]^j,\mathbf{v}''=\mathbf{v}[g,\bar{y}]^{j-1},\bar{y}=(y_1,\ldots,y_{t-1},y'_t,\ldots,y'_n)\\ 0 & \text{otherwise}\end{cases}.$$

In other words, $\mathbf{w}$ is defined for each element of the tree as the difference between two trees if there is continuation of the path on which the two trees are indeed covering elements for some $g\in\mathcal{G}$, and 0 if no such continuation exists. Then W^j is defined as the collection of all such trees $\mathbf{w}$ obtained by pairing up all choices of trees from V^j and V^{j-1}. Clearly, the size $|W^j|\le|V^j|\times|V^{j-1}|\le|V^j|^2$.

We now use the result of Lemma 3:

$$\mathbb{E}\max_{\mathbf{w}\in W^j}\left[\sum_{t:\mathbf{p}_t(y)\in\mathcal{D}}\eta(\mathbf{p}_t(y),y_t)\mathbf{w}_t(y)\right]\le 5\bar{v}\sqrt{\frac{\log|W^j|}{\delta}}+\frac{2v_{\max}\log|W^j|}{\delta}. \tag{47}$$

with $\bar{v}=\max_{\mathbf{w},y}(\sum_{t=1}^n\mathbf{w}_t(y)^2)^{1/2}$ and $v_{\max}=\max_{\mathbf{w},y}|\mathbf{w}_t(y)|$. We over-bound $\bar{v}$ by $v_{\max}$ in the arguments below. By construction of each $\mathbf{w}\in W^j$, the ℓ_2 norm along any path is upper bounded by $3\sqrt{n}\gamma 2^{-j}$ (see [14]). We conclude that

$$\mathbb{E}\max_{\mathbf{w}\in W^j}\left[\sum_{t:\mathbf{p}_t(y)\in\mathcal{D}}\eta(\mathbf{p}_t(y),y_t)\mathbf{w}_t(y)\right]\le 15\sqrt{\frac{2n}{\delta}}(\gamma 2^{-j})\sqrt{\log|V^j|}+\frac{4(\gamma 2^{-j})\log|V^j|}{\delta}.$$

Observe that

$$\begin{aligned}
\sum_{j=1}^{N}\gamma 2^{-j}\sqrt{\log|V^j|} &= 2\sum_{j=1}^{N}(\gamma 2^{-j}-\gamma 2^{-(j+1)})\sqrt{\log\mathcal{N}_\infty(\mathcal{G},\gamma 2^{-j},\mathbf{z})} && (48)\\
&\le 2\int_{\gamma 2^{-(N+1)}}^{\gamma}\sqrt{\log\mathcal{N}_\infty(\mathcal{G},\rho,\mathbf{z})}d\rho && (49)
\end{aligned}$$

and similarly

$$\sum_{j=1}^{N} \gamma 2^{-j} \log |V^j| \le 2 \int_{\gamma 2^{-(N+1)}}^{\gamma} \log \mathcal{N}_\infty(\mathcal{G}, \rho, \mathbf{z}) d\rho. \tag{50}$$

Fix $\alpha \in (0,\gamma)$ and let $N = \max\{j : \gamma 2^{-j} > 2\alpha\}$. Then $\gamma 2^{-(N+1)} \le 2\alpha$ and $\gamma 2^{-N} \le 4\alpha$. Combining all the bounds,

$$\mathbb{E} \sup_{g \in \mathcal{G}} \left[\sum_{t : \mathbf{p}_t(y) \in \mathcal{D}} \eta(\mathbf{p}_t(y), y_t) \Big(g(\mathbf{z}_t(y)) - \mathbf{v}[y,g]_t(y) \Big) \right] \tag{51}$$

$$\le \inf_{\alpha (0,\gamma]} \left\{ \frac{4n\alpha}{\delta} + 30 \sqrt{\frac{2n}{\delta}} \int_{\alpha}^{\gamma} \sqrt{\log \mathcal{N}_\infty(\mathcal{G}, \rho, \mathbf{z})} d\rho + \frac{8}{\delta} \int_{\alpha}^{\gamma} \log \mathcal{N}_\infty(\mathcal{G}, \rho, \mathbf{z}) d\rho \right\} \tag{52}$$

The statement of the theorem follows by combining the two upper bounds for (41). □

Proof of Theorem 6. The proof closely follows the one in [14, Thm. 4], and we refer to that paper for the missing details. Define the function $g_k(d,n) = \sum_{i=0}^{d} \binom{n}{i} k^i$ for $n \ge 1$ and $d \ge 0$, and note the recursion

$$g_k(d, n-1) + k g_k(d-1, n-1) = g_k(d,n).$$

We proceed by induction on (n,d). The base of the induction is the same as in the proof of [14, Thm. 4]. For the induction step, fix an *ordered* $\mathcal{X}$-valued tree $\mathbf{x}$ of depth n and suppose $\text{fat}_2^o(\mathcal{F}, \mathcal{X}) = d$. Define the partition $\mathcal{F} = \cup_{i=1}^{k} \mathcal{F}_i$ according to $\mathcal{F}_i = \{f : f(\mathbf{x}_1) = i\}$. For the sake of contradiction, suppose $\text{fat}_2^o(\mathcal{F}_i, \mathsf{Img}(\mathbf{x})) = \text{fat}_2^o(\mathcal{F}_j, \mathsf{Img}(\mathbf{x})) = d$ for some $j - i \ge 2$. Then there exists two $\mathsf{Img}(\mathbf{x})$-valued ordered trees $\mathbf{w}$ and $\mathbf{v}$ of depth d that are 2-shattered by $\mathcal{F}_i$ and $\mathcal{F}_j$, respectively. Crucially, $\mathbf{x}_1$ cannot appear in either of these trees (that is, $\mathbf{x}_1 \notin \mathsf{Img}(\mathbf{w}) \cup \mathsf{Img}(\mathbf{v})$) because functions in $\mathcal{F}_i$ (resp., $\mathcal{F}_j$) are constant on $\mathbf{x}_1$. Furthermore, $\mathbf{x}_1 \preceq a$ for any $a \in \mathsf{Img}(\mathbf{w}) \cup \mathsf{Img}(\mathbf{v})$. Hence, by joining $\mathbf{w}$ and $\mathbf{v}$ with $\mathbf{x}_1$ at the root, we obtain an ordered tree which is now 2-shattered. The witness of this shattering is constructed by joining the two witnesses (for $\mathbf{w}$ and $\mathbf{v}$) and $(i+j)/2$ at the root. This leads to a contradiction. The rest of the proof follows exactly as in [14, Thm. 4]. □

Proof of Lemma 7. The gradient of g_t at w_t is

$$\nabla g_t(w_t) = -\mathbf{1}\{y_t = 1\} \frac{x_t}{1 + \langle w_t, x_t \rangle} + \mathbf{1}\{y_t = 0\} \frac{x_t}{1 - \langle w_t, x_t \rangle}$$

and the Hessian of the barrier as

$$\nabla^2 R(w_t) = \frac{2}{1 - \|w_t\|^2} I + \frac{4}{(1 - \|w_t\|^2)^2} w_t w_t^\top.$$

By rotational invariance, for the following calculation we may assume without loss of generality that $w_t = a\mathbf{e}_1$ is in the direction of the basis vector $\mathbf{e}_1$ and $a > 0$. We can then write the inverse (see [2]) as

$$\nabla^2 R(w_t)^{-1} = \frac{1}{2}(1-a^2)(I - \mathbf{e}_1\mathbf{e}_1^\top) + \frac{(1-a^2)^2}{2(1-a^2)+4} \mathbf{e}_1\mathbf{e}_1^\top \preceq (1-a)(I - \mathbf{e}_1\mathbf{e}_1^\top) + \frac{2}{3}(1-a)^2 \mathbf{e}_1\mathbf{e}_1^\top.$$

Consider the case $y_t = 0$ (the analysis for $y_t = 1$ follows the same lines). Let us write $x_t = b\mathbf{e}_1 + \mathbf{y}$ with $\langle \mathbf{y}, \mathbf{e}_1 \rangle = 0$ and $\|y\|^2 \le 1 - b^2$. We have

$$\nabla g_t(w_t) = \frac{x_t}{1 - \langle w_t, x_t \rangle} = \frac{b\mathbf{e}_1 + \mathbf{y}}{1 - ab}.$$

and

$$\nabla g_t(w_t)^\top \nabla^2 R(w_t)^{-1} \nabla g_t(w_t) \le \frac{b^2}{(1-ab)^2} \cdot (1-a)^2 + \frac{1-b^2}{(1-ab)^2} \cdot (1-a)$$

If $b \le 0$, the above expression is upper bounded by 2, and for $b > 0$, the expression is upper bounded by 3 (we did not optimize the constants). □

Proof of Lemma 9. We reproduce the proof from [8] for completeness. Let $x \in \operatorname{int} \mathcal{K}$ and $r \in [0,1)$. Let

$$y = x - \frac{r}{\|\nabla\psi(x)\|_x^*}[\nabla^2 F(x)]^{-1}\nabla\psi(x).$$

Then $y \in \operatorname{int}\mathcal{K}$ because the Dikin ellipsoid is contained in the set. Hence,

$$0 \le \psi(y) \le \psi(x) + \left\langle \nabla\psi(x), y - x \right\rangle = \psi(x) - r\|\nabla\psi(x)\|_x^*.$$

Statement follows because r is arbitrary in $[0,1)$. □

Acknowledgements

We gratefully acknowledge the support of NSF under grants CAREER DMS-0954737 and CCF-1116928, as well as Dean's Research Fund.

References

[1] J. Abernethy, A. Agarwal, P. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[2] J. Abernethy and A. Rakhlin. Beating the adaptive bandit with high probability. In *COLT*, 2009.

[3] N. Cesa-Bianchi and G. Lugosi. Minimax regret under log loss for general classes of experts. In *Proceedings of the Twelfth annual conference on computational learning theory*, pages 12–18. ACM, 1999.

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

[5] Y. Freund. Predicting a binary sequence almost as well as the optimal biased coin. In *Proceedings of the ninth annual conference on Computational learning theory*, pages 89–98. ACM, 1996.

[6] E. Hazan, A. Agarwal, and S Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[7] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147, 1998.

[8] Y. Nesterov. Barrier subgradient method. *Mathematical programming*, 127(1):31–56, 2011.

[9] A. Rakhlin, O. Shamir, and K. Sridharan. Relax and randomize: From value to algorithms. In *Advances in Neural Information Processing Systems 25*, pages 2150–2158, 2012.

[10] A. Rakhlin and K. Sridharan. Statistical learning and sequential prediction, 2012. Available at `http://stat.wharton.upenn.edu/~rakhlin/courses/stat928/stat928_notes.pdf`.

[11] A. Rakhlin and K. Sridharan. Online nonparametric regression. In *Conference on Learning Theory*, 2014.

[12] A. Rakhlin and K. Sridharan. Online nonparametric regression with general loss functions, 2015. Available at `http://arxiv.org/abs/1501.06598`.

[13] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *Advances in Neural Information Processing Systems 23*, pages 1984–1992, 2010.

[14] A. Rakhlin, K. Sridharan, and A. Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, February 2014.

[15] J. Rissanen. Complexity of strings in the class of markov sources. *Information Theory, IEEE Transactions on*, 32(4):526–532, 1986.

[16] J. Rissanen. Fisher information and stochastic complexity. *Information Theory, IEEE Transactions on*, 42(1):40–47, 1996.

[17] Y. M. Shtarkov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987.

[18] Q. Xie and A.R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *Information Theory, IEEE Transactions on*, 46(2):431–445, 2000.