Online Nonparametric Regression with General Loss Functions

Alexander Rakhlin
University of Pennsylvania

Karthik Sridharan
Cornell University

January 25, 2015

Abstract

This paper establishes minimax rates for online regression with arbitrary classes of functions and general losses.[1] We show that below a certain threshold for the complexity of the function class, the minimax rates depend on both the curvature of the loss function and the sequential complexities of the class. Above this threshold, the curvature of the loss does not affect the rates. Furthermore, for the case of square loss, our results point to the interesting phenomenon: whenever sequential and i.i.d. empirical entropies match, the rates for statistical and online learning are the same.

In addition to the study of minimax regret, we derive a generic forecaster that enjoys the established optimal rates. We also provide a recipe for designing online prediction algorithms that can be computationally efficient for certain problems. We illustrate the techniques by deriving existing and new forecasters for the case of finite experts and for online linear regression.

1 Introduction

We study the problem of predicting a real-valued sequence $y_1, \ldots, y_n$ in an on-line manner. At time $t = 1, \ldots, n$, the forecaster receives side information in the form of an element x_t of an abstract set $\mathscr{X}$. The forecaster then makes a prediction $\widehat{y}_t$ on the basis of the current observation x_t and the data $\{(x_i, y_i)\}_{i=1}^{t-1}$ encountered thus far, and then observes the response y_t.

Such a problem of sequence prediction is studied in the literature under two distinct settings: probabilistic and deterministic [18]. In the former setting, which falls within the purview of time series analysis, one posits a parametric form for the data-generating mechanism and estimates the model parameters based on past instances and input information in order to make the next prediction. In contrast, in the deterministic setting one assumes no such probabilistic mechanism. Instead, the goal is phrased as that of predicting as well as the best forecaster from a benchmark set of strategies. This latter setting—often termed *prediction of individual sequences*, or *online learning*—is the focus of the present paper.

We let the outcome y_t and the prediction $\widehat{y}_t$ take values in $\mathscr{Y} \subseteq \mathbb{R}$ and $\widehat{\mathscr{Y}} \subseteq \mathbb{R}$, respectively. Formally, a deterministic prediction strategy is a mapping $(\mathscr{X} \times \mathscr{Y})^{t-1} \times \mathscr{X} \to \widehat{\mathscr{Y}}$. We let the loss function $(\widehat{y}_t, y_t) \mapsto \boldsymbol{\ell}(\widehat{y}_t, y_t)$ score the quality of the prediction on a single round.

Assume that the time horizon $n \in \mathbb{Z}_+$, is known to the forecaster. The overall quality of the forecaster is then evaluated against the benchmark set of predictors, denoted as a class $\mathscr{F}$ of functions $\mathscr{X} \to \widehat{\mathscr{Y}}$. The cumulative *regret* of the forecaster on the sequence $(x_1, y_1), \ldots, (x_n, y_n)$ is defined as

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

The forecaster aims to keep the difference in (1) small for *all* sequences $(x_1, y_1), \ldots, (x_n, y_n)$.

The comparison class $\mathscr{F}$ encodes the prior belief about the family of predictors one expects to perform well. If a forecasting strategy guarantees small regret for all sequences, and if $\mathscr{F}$ is a good model for the sequences observed in reality, then the forecasting strategy will also perform well in terms of its cumulative error. In fact, we can take

[1] This paper builds upon the study of online regression with square loss, presented by the authors at the COLT 2014 conference.

$\mathcal{F}$ to be a class of solutions (that is, forecasting strategies) to a set of probabilistic sources one would obtain by positing a generative model of data. By doing so, we are modeling solutions to the prediction problem rather than modeling the data-generating mechanism. We refer to [18, 21] for further discussions on this "duality" between the probabilistic and deterministic approaches.

To ensure that $\mathcal{F}$ captures the phenomenon of interest, we would like $\mathcal{F}$ to be large. However, increasing the "size" of $\mathcal{F}$ likely leads to larger regret, as the comparison term in (1) becomes smaller. On the other hand, decreasing the "size" of $\mathcal{F}$ makes the regret minimization task easier, yet the prediction method is less likely to be successful in practice. This dichotomy is an analogue of the bias-variance tradeoff commonly studied in statistics. A contribution of this paper is an analysis of the growth of regret (with n) in terms of various notions of complexity of $\mathcal{F}$. The task was already accomplished in [24] for the case of absolute loss $\ell(a,b) = |a-b|$. In the present paper we obtain optimal guarantees for convex Lipschitz losses under very general assumptions.

To give the reader a sense of the results of this paper, we state the following informal corollary. Let complexity of $\mathcal{F}$ be measured via *sequential entropy* at scale β, to be defined below. (For the reader familiar with covering numbers, this is a sequential analogue—introduced in [24]—of the classical Koltchinskii-Pollard entropy).

Corollary 1 (Informal)**.** *Suppose sequential entropy at scale β behaves as $\mathcal{O}(\beta^{-p})$, $p>0$. Then optimal regret*

- *for prediction with absolute loss grows as $n^{1/2}$ if $p \in (0,2)$, and as $n^{1-1/p}$ for $p>2$;*
- *for prediction with square loss grows as $n^{1-2/(2+p)}$ if $p \in (0,2)$, and as $n^{1-1/p}$ for $p>2$.*

Moreover, these rates have matching, sometimes modulo a logarithmic factor, lower bounds.

The first part of this corollary is established in [24]. The second part requires new techniques that take advantage of the curvature of the loss function.

In an attempt to entice the reader, let us discuss two conclusions that can be drawn from Corollary 1. First, the rates of convergence match optimal rates for excess square loss in the realm of distribution-free Statistical Learning Theory with i.i.d. data, under the assumption on the behavior of empirical covering numbers [27]. Hence, in the absence of a gap between classical and sequential complexities (introduced later) *the regression problems in the two seemingly different frameworks enjoy the same rates of convergence.* A deeper understanding of this phenomenon is of a great interest.

The second conclusion concerns the same optimal rate $n^{-1/p}$ for both square and absolute loss for "rich" classes ($p>2$). Informally, strong convexity of the loss does not affect the rate of convergence for such massive classes. A geometric explanation of this interesting phenomenon requires further investigation.

We finish this introduction with a note about the generality of the setting proposed so far. Suppose $\mathcal{X} = \cup_{t\le n}\mathcal{Y}^t$, the space of all histories of $\mathcal{Y}$-valued outcomes. Denoting $x_t = (y_1,\dots,y_{t-1}) \triangleq y^{t-1}$, we may view each $f \in \mathcal{F}$ itself as a strategy that maps history y^{t-1} to a prediction. Ensuring that x_t is not arbitrary but consistent with history only makes the task of regret minimization easier; the analysis of this paper for this case follows along the same lines, but we omit the extra overhead of restrictions on x_t's and instead refer the reader to [14, 21].

The paper is organized as follows. Section 2 introduces the notation and then presents a brief overview of sequential complexities. Upper and lower bounds on minimax regret are established in Sections 3 and 4. We calculate minimax rates for various examples in Section 5. We then turn to the question of developing algorithms in Section 6. We first show that an algorithm based on the Rademacher relaxation is admissible (see [19]) and yields the rates derived in a non-constructive manner in the first part of the paper. We show that further relaxations in finite dimensional space lead to the famous Vovk-Azoury-Warmuth forecaster. We also derive a prediction method for finite class $\mathcal{F}$.

2 Preliminaries

2.1 Assumptions and Definitions

We assume that the set of outcomes $\mathcal{Y}$ is a bounded set, a restriction that can be removed by standard truncation arguments (see e.g. [12]). Let $\mathcal{X}$ be some set of covariates, and let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \widehat{\mathcal{Y}}$ for some

$\widehat{\mathscr{Y}} \subseteq \mathbb{R}$. Recall the protocol of the online prediction problem: On each round $t \in \{1,\dots,n\}$, $x_t \in \mathscr{X}$ is revealed to the learner who subsequently makes a prediction $\widehat{y}_t \in \widehat{\mathscr{Y}}$. The response $y_t \in \mathscr{Y}$ is revealed after the prediction is made.

The loss function $\boldsymbol{\ell}(\cdot, y)$ is assumed to be convex. Let $\partial_a \boldsymbol{\ell}(a,y)$ denote any element of the subdifferential set (with respect to first argument), and assume that

$$\sup_{a\in\widehat{\mathscr{Y}}, y\in\mathscr{Y}} |\partial_a \boldsymbol{\ell}(a,y)| \le G < \infty.$$

We assume that for any distribution of y supported on $\mathscr{Y}$, there is a minimizer of expected loss that is finite and belongs to $\widehat{\mathscr{Y}}$:

$$\widehat{\mathscr{Y}} \cap \operatorname*{argmin}_{\widehat{y}\in\mathbb{R}} \mathbb{E}\boldsymbol{\ell}(\widehat{y}, y) \neq \emptyset.$$

Given a $y \in \mathscr{Y}$, the error of a linear expansion at a to approximate function value at b is denoted by

$$\Delta^{y}_{a,b} \triangleq \boldsymbol{\ell}(b,y) - [\boldsymbol{\ell}(a,y) + \partial_a \boldsymbol{\ell}(a,y)\cdot(b-a)].$$

Let $\underline{\boldsymbol{\Delta}} : (\widehat{\mathscr{Y}} - \widehat{\mathscr{Y}}) \to \mathbb{R}_{\ge 0}$ be a function defined pointwise as

$$\underline{\boldsymbol{\Delta}}(x) = \inf_{a,b\in\widehat{\mathscr{Y}}, y\in\mathscr{Y} \text{ s.t. } b-a=x} \Delta^{y}_{a,b}, \tag{2}$$

a lower bound on the residual for any two values separated by x. For instance, an easy calculation shows that $\underline{\boldsymbol{\Delta}}(x) = x^2$ for $\boldsymbol{\ell}(\widehat{y}, y) = (\widehat{y} - y)^2$.

2.2 Minimax Formulation

Unlike most previous approaches to the study of online regression, we do not start from an algorithm, but instead work directly with minimax regret. We will be able to extract a (not necessarily efficient) algorithm after obtaining upper bounds on the minimax value. Let us introduce the notation that makes the minimax regret definition more concise. We use $\langle\!\langle \cdots \rangle\!\rangle_{t=1}^{n}$ to denote an interleaved application of the operators, repeated over $t = 1\dots n$ rounds. With this notation, the minimax regret of the online regression problem described earlier can be written as

$$V_n = \left\langle\!\!\left\langle \sup_{x_t} \inf_{\widehat{y}_t} \sup_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{y}_t, y_t) - \inf_{f\in\mathscr{F}} \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \right\} \tag{3}$$

where each x_t ranges over $\mathscr{X}$, $\widehat{y}_t$ ranges over $\widehat{\mathscr{Y}}$, and y_t ranges over $\mathscr{Y}$. An upper bound on V_n guarantees the existence of an algorithm (that is, a way to choose $\widehat{y}_t$'s) with at most that much regret against any sequence. A lower bound on V_n, in turn, guarantees the existence of a sequence on which no method can perform better than the given lower bound.

2.3 Sequential Complexities

One of the key tools in the study of estimators based on i.i.d. data is the symmetrization technique [13]. By introducing Rademacher random variables, one can study the supremum of an empirical process conditionally on the data. Conditioning facilitates the introduction of sample-based complexities of a function class, such as an empirical covering number. For a class of bounded functions, the covering number with respect to the empirical metric is necessarily finite and leads to a correct control of the empirical process even if discretization of the function class in a data-independent manner is impossible. We will return to this point when comparing our approach with discretization-based methods.

In the online prediction scenario, symmetrization is more subtle and involves the notion of a binary tree. The binary tree is, in some sense, the smallest entity that captures the sequential nature of the problem. More precisely, a $\mathscr{Z}$-valued tree $\mathbf{z}$ of depth n is a complete rooted binary tree with nodes labeled by elements of a set $\mathscr{Z}$. Equivalently, we think of $\mathbf{z}$ as n labeling functions, where $\mathbf{z}_1$ is a constant label for the root, $\mathbf{z}_2(-1), \mathbf{z}_2(+1) \in \mathscr{Z}$ are the labels for the left and right children of the root, and so forth. Hence, for $\epsilon = (\epsilon_1, \dots, \epsilon_n) \in \{\pm 1\}^n$, $\mathbf{z}_t(\epsilon) = \mathbf{z}_t(\epsilon_1, \dots, \epsilon_{t-1}) \in \mathcal{Z}$ is the label of the node on the t-th level of the tree obtained by following the path ϵ. For a function $g : \mathcal{Z} \to \mathbb{R}$, $g(\mathbf{z})$ is an $\mathbb{R}$-valued tree with labeling functions $g \circ \mathbf{z}_t$ for level t (or, in plain words, evaluation of g on $\mathbf{z}$).

We now define two tree-based complexity notions of a class of functions.

Definition 1 ([24])**.** Sequential Rademacher complexity of a class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ on a given $\mathcal{X}$-valued tree $\mathbf{x}$ of depth n, as well as its supremum, are defined as

$$\mathfrak{R}_n(\mathcal{F}; \mathbf{x}) \triangleq \mathbb{E} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right], \quad \mathfrak{R}_n(\mathcal{F}) \triangleq \sup_{\mathbf{x}} \mathfrak{R}_n(\mathcal{F}; \mathbf{x}) \tag{4}$$

where the expectation is over a sequence of independent Rademacher random variables $\epsilon = (\epsilon_1, \dots, \epsilon_n)$.

One may think of the functions $\mathbf{x}_1, \dots, \mathbf{x}_n$ as a predictable process with respect to the dyadic filtration $\{\sigma(\epsilon_1, \dots, \epsilon_t)\}_{t \ge 1}$. The following notion of a β-cover quantifies complexity of the class $\mathcal{F}$ evaluated on the predictable process.

Definition 2 ([24])**.** A set V of $\mathbb{R}$-valued trees of depth n forms a β-cover (with respect to the ℓ_q norm) of a function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ on a given $\mathcal{X}$-valued tree $\mathbf{x}$ of depth n if

$$\forall f \in \mathcal{F}, \forall \epsilon \in \{\pm 1\}^n, \exists \mathbf{v} \in V \quad \text{s.t.} \quad \frac{1}{n} \sum_{t=1}^{n} |f(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon)|^q \le \beta^q.$$

A β-cover in the ℓ_∞ sense requires that $|f(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon)| \le \beta$ for all $t \in [n]$. The size of the smallest β-cover is denoted by $\mathcal{N}_q(\beta, \mathcal{F}, \mathbf{x})$, and $\mathcal{N}_q(\beta, \mathcal{F}, n) \triangleq \sup_{\mathbf{x}} \mathcal{N}_q(\beta, \mathcal{F}, \mathbf{x})$.

We will refer to $\log \mathcal{N}_q(\beta, \mathcal{F}, n)$ as *sequential entropy* of $\mathcal{F}$. In particular, we will study the behavior of V_n when sequential entropy grows polynomially[2] as the scale β decreases:

$$\log \mathcal{N}_2(\beta, \mathcal{F}, n) \sim \beta^{-p}, \quad p > 0. \tag{5}$$

We also consider the parametric "$p = 0$" case when sequential covering itself behaves as

$$\mathcal{N}_2(\beta, \mathcal{F}, n) \sim \beta^{-d} \tag{6}$$

(e.g. linear regression in a bounded set in $\mathbb{R}^d$). We remark that the ℓ_∞ cover is necessarily n-dependent, so the forms we assume for nonparametric and parametric cases, respectively, are

$$\log \mathcal{N}_\infty(\beta, \mathcal{F}, n) \sim \beta^{-p} \log(n/\beta) \quad \text{or} \quad \mathcal{N}_\infty(\beta, \mathcal{F}, n) \sim (n/\beta)^d \tag{7}$$

3 Upper Bounds

The following theorem from [24] shows the importance of sequential Rademacher complexity for prediction with absolute loss.

Theorem 2 ([24])**.** *Let* $\mathcal{Y} = [-1, 1]$, $\mathcal{F} = [-1, 1]^{\mathcal{X}}$, *and* $\boldsymbol{\ell}(\widehat{y}, y) = |\widehat{y} - y|$. *It then holds that*

$$\mathfrak{R}_n(\mathcal{F}) \le V_n \le 2\mathfrak{R}_n(\mathcal{F}).$$

Furthermore, an upper bound of $2G\mathfrak{R}_n(\mathcal{F})$ holds for any G-Lipschitz loss. We observe, however, that as soon as $\mathcal{F}$ contains two distinct functions, sequential Radmeacher complexity of $\mathcal{F}$ scales as $\Omega(n^{1/2})$. Yet, it is known that minimax regret for prediction with square loss grows slower than this rate. Therefore, the direct analysis based on sequential Rademacher complexity (and a contraction lemma) gives loose upper bounds on minimax regret. The key contribution of this paper is an introduction of an offset Rademacher complexity that captures the correct behavior.

In the next lemma, we show that minimax value of the sequential prediction problem with any convex Lipschitz loss function can be controlled via offset sequential Rademacher complexity. As before, let $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ where each ϵ_i is an independent Rademacher random variable.

[2] It is straightforward to allow constants in this definition, and we leave these details out for the sake of simplicity.

Lemma 3. *Under the assumptions and definitions in Section 2.1, the minimax rate is bounded by*

$$V_n \le \sup_{\mathbf{x},\boldsymbol{\mu}} \mathbb{E} \sup_{f\in\mathcal{F}} \left[\sum_{t=1}^n 2G\epsilon_t (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - \underline{\boldsymbol{\Delta}}\Big(f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)\Big)\right] \tag{8}$$

where $\mathbf{x}$ *and* $\boldsymbol{\mu}$ *range over all* $\mathcal{X}$*-valued and* $\widehat{\mathcal{Y}}$*-valued trees of depth n, respectively.*

The right-hand side of (8) will be termed *offset Rademacher complexity* of a function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ with respect to a convex even *offset function* $\underline{\boldsymbol{\Delta}} : \mathbb{R} \to \mathbb{R}_{\ge 0}$ and a mean $\mathbb{R}$-valued tree $\boldsymbol{\mu}$. If $\underline{\boldsymbol{\Delta}} \equiv 0$, we recover the notion of sequential Rademacher complexity since $\mathbb{E}[\epsilon_t \boldsymbol{\mu}_t(\epsilon)] = 0$.

A matching lower bound on the minimax value will be presented in Section 4, and the two results warrant a further study of offset Rademacher complexity. To this end, a natural next question is whether the chaining technique can be employed to control the supremum of this modified stochastic process. As a point of comparison, we first recall that sequential Rademacher complexity of a class $\mathcal{G}$ of $[-1,1]$-valued functions on $\mathcal{Z}$ can be upper bounded via the Dudley integral-type bound

$$\mathbb{E}\sup_{g\in\mathcal{G}} \left[\sum_{t=1}^n \epsilon_t g(\mathbf{z}_t(\epsilon))\right] \le \inf_{\rho\in(0,1]} \left\{4\rho n + 12\sqrt{n}\int_\rho^1 \sqrt{\log \mathcal{N}_2(\delta,\mathcal{G},\mathbf{z})}\, d\delta\right\}, \tag{9}$$

for any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n, as shown in [26]. We aim to obtain tighter upper bounds on the offset Rademacher by taking advantage of the negative offset term.

To initiate the study of offset Rademacher complexity with functions $\underline{\boldsymbol{\Delta}}$ other than quadratic, we recall the notion of a convex conjugate.

Definition 3. For a convex function $\psi : \mathcal{D} \to \mathbb{R}$ with domain $\mathcal{D} \subseteq \mathbb{R}$, the convex conjugate $\psi^* : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is defined as

$$\psi^*(a) = \sup_{d\in\mathcal{D}} \{ad - \psi(d)\}.$$

The chaining technique for controlling a supremum of a stochastic process requires a statement about the behavior of the process over a finite collection. The next lemma provides such a statement for the offset Rademacher process.

Lemma 4. *Let* Δ *be a convex, nonnegative, even function on* $\mathbb{R}$ *and let* Γ^* *denote the convex conjugate of the function* $x \mapsto \Delta(\sqrt{|x|})$*. Assume* Γ^* *is nondecreasing. For any finite set* W *of* $\mathbb{R}$*-valued trees of depth n and any constant* $C > 0$*,*

$$\mathbb{E}\max_{\mathbf{w}\in W} \left\{\sum_{t=1}^n 2C\epsilon_t \mathbf{w}_t(\epsilon) - \Delta(\mathbf{w}_t(\epsilon))\right\} \le \inf_{\lambda>0} \left\{\frac{1}{\lambda}\log|W| + n\,\Gamma^*\left(2C^2\lambda\right)\right\}. \tag{10}$$

Further, for any $[-G,G]$*-valued tree* $\boldsymbol{\eta}$*,*

$$\mathbb{E}\max_{\mathbf{w}\in W} \left[\sum_{t=1}^n \epsilon_t \boldsymbol{\eta}_t(\epsilon)\mathbf{w}_t(\epsilon)\right] \le G\sqrt{2\log|W| \cdot \max_{\mathbf{w}\in W,\epsilon_{1:n}} \sum_{t=1}^n \mathbf{w}_n(\epsilon)^2}\,. \tag{11}$$

As an example, if $\Delta(x) = x^2$, an easy calculation shows that $\Gamma^*(1) = 0$ and $\Gamma^*(y) = +\infty$ for any $y \ne 1$. Hence, the infimum in (10) is achieved at $\lambda = 1/(2C^2)$, and the upper bound becomes $2C^2\log|W|$.

We can now employ the chaining technique to extend the control of the stochastic process beyond the finite collection.

Lemma 5. *Let* Δ *and* Γ^* *be as in Lemma 4. For any* $\mathcal{Z}$*-valued tree* $\mathbf{z}$ *of depth n and a class* $\mathcal{G}$ *of functions* $\mathcal{Z} \to \mathbb{R}$ *and any constant* $C > 0$*,*

$$\mathbb{E}\sup_{g\in\mathcal{G}} \left[\sum_{t=1}^n 2C\epsilon_t g(\mathbf{z}_t(\epsilon)) - \Delta\big(g(\mathbf{z}_t(\epsilon))\big)\right] \le \inf_{\gamma>0}\left\{C \inf_{\rho\in(0,\gamma)}\left\{4\rho n + 12\sqrt{n}\int_\rho^\gamma \sqrt{\log\mathcal{N}_\infty(\delta,\mathcal{G},\mathbf{z})}\,d\delta\right\}\right.$$
$$\left. + \inf_{\lambda>0}\left\{\frac{1}{\lambda}\log\mathcal{N}_\infty\left(\tfrac{\gamma}{2},\mathcal{G},\mathbf{z}\right) + n\,\Gamma^*\left(2C^2\lambda\right)\right\}\right\}.$$

Remark 1. *For the case of $\Delta(x) = x^2$, it is possible to prove the upper bound of Lemma 5 in terms of ℓ_2 sequential covering numbers rather than ℓ_∞ (see [22]).*

Lemma 5, together with Lemma 3, yield upper bounds on minimax regret under assumptions on the growth of sequential entropy. Before detailing the rates, we present lower bounds on the minimax value in terms of the offset Rademacher complexity and combinatorial dimensions.

4 Lower Bounds

The function $\underline{\mathbf{\Delta}}$, arising from uniform (or strong) convexity of the loss function, enters the upper bounds on minimax regret. For proving lower bounds, we consider the dual property, that of (restricted) smoothness. To this end, let $S \subseteq \widehat{\mathcal{Y}}$ be a subset satisfying the following condition:

$$\forall s \in S,\ \exists y_1(s), y_2(s) \in \mathcal{Y} \quad \text{s.t.} \quad s \in \operatorname*{argmin}_{\widehat{y} \in \widehat{\mathcal{Y}}} \frac{1}{2}\left(\boldsymbol{\ell}(\widehat{y}, y_1(s)) + \boldsymbol{\ell}(\widehat{y}, y_2(s))\right). \tag{12}$$

For any such subset S, let $\overline{\mathbf{\Delta}}_S : (\widehat{\mathcal{Y}} - S) \to \mathbb{R}_{\geq 0}$ be defined as

$$\overline{\mathbf{\Delta}}_S(x) = \sup_{s \in S, b \in \widehat{\mathcal{Y}} \text{ s.t. } b-s=x} \max\left\{\Delta_{s,b}^{y_1(s)}, \Delta_{s,b}^{y_2(s)}\right\}. \tag{13}$$

We write $\overline{\mathbf{\Delta}}_\kappa$ for the singleton set $S = \{\kappa\}$.

The lower bounds in this section will be constructed from symmetric distributions supported on two carefully chosen points. Crucially, we do not require a uniform notion of smoothness, but rather a condition on the loss that holds for a restricted subset S and a two-point distribution.

As an example, consider square loss and $\mathcal{Y} = \widehat{\mathcal{Y}} = (-B, B)$. For any $s \in \widehat{\mathcal{Y}}$, we may choose the two points as $s \pm \delta \in \mathcal{Y}$, for small enough δ, with the desired property. Then $S = \widehat{\mathcal{Y}}$ and $\overline{\mathbf{\Delta}}_S(x) = x^2$.

Lemma 6. *Fix $R > 0$. Suppose $S \neq \emptyset$ satisfies condition (12), and suppose that for any $s \in S$,*

$$\partial\boldsymbol{\ell}(s, y_1(s)) = +R, \quad \partial\boldsymbol{\ell}(s, y_2(s)) = -R.$$

Then for any S-valued tree $\boldsymbol{\mu}$ of depth n,

$$V_n \geq \sup_{\mathbf{x}} \mathbb{E} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^n \epsilon_t R\left(f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)\right) - \overline{\mathbf{\Delta}}_{\boldsymbol{\mu}_t(\epsilon)}(f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon))\right]. \tag{14}$$

The lower bound in (14) is an offset Rademacher complexity that matches the upper bound of Lemma 3 up to constants, as long as functions $\underline{\mathbf{\Delta}}$ and $\overline{\mathbf{\Delta}}$ exhibit the same behavior. In particular, the upper and lower bounds match up to a constant for the case of square loss.

Our next step is to quantify the lower bound in terms of n according to "size" of $\mathcal{F}$. In contrast to the more common statistical approaches based on covering numbers and Fano inequality, we turn to a notion of a combinatorial dimension as the main tool.

Definition 4. An $\mathcal{X}$-valued tree of depth d is said to be β-shattered by $\mathcal{F}$ if there exists an $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d such that

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f^\epsilon \in \mathcal{F} \quad \text{s.t.} \quad \epsilon_t(f^\epsilon(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) \geq \beta/2$$

for all $t \in \{1, \ldots, d\}$. The tree $\mathbf{s}$ is called a *witness*. The largest d for which there exists a β-shattered $\mathcal{X}$-valued tree is called the (sequential) fat-shattering dimension, denoted by $\text{fat}_\beta(\mathcal{F})$.

The reader will notice that the upper bound of Lemma 5 is in terms of sequential entropies rather than combinatorial dimensions. The two notions, however, are closely related.

Theorem 7 ([26])**.** *Let $\mathscr{F}$ be a class of functions $\mathscr{X} \to [-1,1]$. For any $\beta > 0$,*

$$\mathscr{N}_2(\beta,\mathscr{F},n) \le \mathscr{N}_\infty(\beta,\mathscr{F},n) \le \left(\frac{2en}{\beta}\right)^{\mathrm{fat}_\beta(\mathscr{F})} .$$

As a consequence of the above theorem, if $\log \mathscr{N}_2(\beta,\mathscr{F},n) \ge (c/\beta)^p$ and $\beta \ge 1/n$, then $\mathrm{fat}_\beta(\mathscr{F}) \ge (c'/\beta)^p/\log(n)$ where c, c' may depend on the range of functions in $\mathscr{F}$.

The lower bounds will now be obtained assuming $\mathrm{fat}_\beta(\mathscr{F}) \ge \beta^{-p}$ behavior of the fat-shattering dimension, and the corresponding statements in terms of the sequential entropy growth will involve extra logarithmic factors, hidden in the $\widetilde{\Omega}(\cdot)$ notation.

Lemma 8. *Suppose the statement of Lemma 6 holds for some $R > 0$, and suppose*

$$\overline{\boldsymbol{\Delta}}_{\boldsymbol{\mu}_t(\epsilon)}(\boldsymbol{\mu}_t(\epsilon) - f(\mathbf{x}_t(\epsilon))) \le \frac{R}{2}|\boldsymbol{\mu}_t(\epsilon) - f(\mathbf{x}_t(\epsilon))| \tag{15}$$

for any $f \in \mathscr{F}$ and $\boldsymbol{\mu}, \mathbf{x}$ in the statement of Lemma 6. Then it holds that for any $\beta > 0$ and $n = \mathrm{fat}_\beta(\mathscr{F})$,

$$V_n \ge (R/2)n\beta.$$

In particular, if $\mathrm{fat}_\beta(\mathscr{F}) \ge \beta^{-p}$ for $p > 0$, we have

$$\frac{1}{n}V_n \ge (R/2)n^{-1/p} .$$

As an example, consider the case of square loss with $\mathscr{Y} = [-B,B]$. Then we may take $S = \{0\}$, $y_1 = B$, $y_2 = -B$, and hence $R = 2B$. We verify that (15) holds for $\widehat{\mathscr{Y}} = [-B/2, B/2]$.

Lemma 9. *Suppose the statement of Lemma 6 holds for some $R > 0$. For any class $\mathscr{F}'$ and $\beta > 0$, there exists a modified class $\mathscr{F}$ such that for all $\beta' < \beta$, $\mathrm{fat}_{\beta'}(\mathscr{F}') \le \mathrm{fat}_{\beta'}(\mathscr{F}) \le 2\mathrm{fat}_{\beta'}(\mathscr{F}') + 4$ and for $n > \mathrm{fat}_\beta(\mathscr{F})$,*

$$\frac{1}{n}V_n \ge \sup_{\beta,\kappa}\left\{\frac{R\,\beta}{2}\sqrt{\frac{\mathrm{fat}_\beta(\mathscr{F})}{2n}} - \boldsymbol{\Delta}_\kappa\left(\frac{\beta}{2}\right)\right\}.$$

Armed with the upper bounds of Section 3 and the lower bounds of Section 4, we are ready to detail specific minimax rates of convergence for various classes of regression functions $\mathscr{F}$ and a range of loss functions $\boldsymbol{\ell}$.

5 Minimax Rates

Combining Lemma 3 and Lemma 5, we can detail the behavior of minimax regret under an assumption about the growth rate of sequential entropy.

Theorem 10. *Let $r \ge 2$, $p > 0$ and suppose the loss function and the function class are such that*

$$\underline{\boldsymbol{\Delta}}(t) \ge K t^r, \quad \log \mathscr{N}_\infty(\beta,\mathscr{F},n) \le \beta^{-p}\log(n/\beta).$$

Then for $p \in (0,2)$,

$$\frac{1}{n}V_n \le \min\left\{c_{r,p}\; n^{-\frac{r}{2(r-1)+p}}\, G^{\frac{2r}{2(r-1)+p}}\, K^{-\frac{2-p}{2(r-1)+p}} \log(n)\;,\;\; c_{\mathscr{F}}G\log^{1/2}(n)n^{-1/2}\right\}. \tag{16}$$

and for $p > 2$

$$\frac{1}{n}V_n \le c_p G\log^{1/2}(n)n^{-1/p} \tag{17}$$

Here, $c_{\mathscr{F}}$ depends on $\sup_{f\in\mathscr{F}}|f|_\infty$. At $p = 2$, the bound (17) gains an extra $\log(n)$ factor.

We match the above upper bounds with lower bounds under the assumption on the growth of the combinatorial dimension.

Theorem 11. *Suppose the statement of Lemma 6 holds for some $R > 0$ and $\kappa \in S \subseteq \widehat{\mathscr{Y}}$. Let $r \geq 2$, $p \in (0,2)$, and assume*

$$\overline{\mathbf{\Delta}}_\kappa\left(\beta/2\right) \leq K\beta^r, \quad \mathrm{fat}_\beta \geq \beta^{-p}.$$

Then there exists a function class such that for some constant $c_{p,r} > 0$,

$$\frac{1}{n}V_n \geq c_{p,r}\min\left\{n^{-\frac{r}{2(r-1)+p}}R^{\frac{2r}{2(r-1)+p}}K^{-\frac{2-p}{2(r-1)+p}}\ ,\ Rn^{-1/2}\right\}.$$

for $p \in (0,2)$. Furthermore, for $p > 2$, for any $\mathscr{F}$ with $\mathrm{fat}_\beta \geq \beta^{-p}$,

$$\frac{1}{n}V_n \geq (R/2)n^{-1/p}$$

under the assumption (15).

The lower bound of Theorem 11 matches (up to polylogarithmic in n factors) the upper bound of Theorem 10 in its dependence on n, the dependence on the constant K, and in dependence on the size of the gradients G (respectively, R). The rest of this section is devoted to the discussion of the derived upper and lower bounds for particular loss functions or particular classes of functions.

5.1 Absolute loss

We verify that the general statements recover the correct rates for the case of $\boldsymbol{\ell}(\widehat{y}, y) = |\widehat{y} - y|$. Since the absolute loss is not strongly convex, we take $K = 0$ (and $\underline{\mathbf{\Delta}} \equiv 0$). Theorem 10 then yields the $\widetilde{\mathscr{O}}(n^{-1/2})$ rate for $p \in (0,2)$ and $\widetilde{\mathscr{O}}(n^{-1/p})$ for $p > 2$, up to logarithmic factors. These rates are matched, again up to logarithmic factors, in Theorem 11. Of course, the result already follows from Theorem 2.

It is also instructive to check the case of $r \to \infty$. In this case, if K is scaled properly by the range of function values, the function $\underline{\mathbf{\Delta}}$ approaches the zero function, indicating absence of strong convexity of the loss. Examining the power $\frac{r}{2(r-1)+p}$ in Theorem 10, we see that it approaches 1/2, matching the discussion of the preceding paragraph.

5.2 Square loss

The case of square loss $\boldsymbol{\ell}(\widehat{y}, y) = (\widehat{y} - y)^2$ has been studied in [22]. In view of Remark 1, we state the corollary below in terms of ℓ_2 covering numbers, thus removing some logarithmic terms of Theorem 10.

Corollary 12. *For a class $\mathscr{F}$ with sequential entropy growth $\log\mathscr{N}_2(\beta, \mathscr{F}, n) \leq \beta^{-p}$,*

- *For $p > 2$, the minimax regret[3] is bounded as $\frac{1}{n}V_n \leq Cn^{-1/p}$*
- *For $p \in (0,2)$, the minimax regret is bounded as $\frac{1}{n}V_n \leq Cn^{-2/(2+p)}$*
- *For the parametric case (6), $\frac{1}{n}V_n \leq Cdn^{-1}\log(n)$*
- *For finite set $\mathscr{F}$, $\frac{1}{n}V_n \leq Cn^{-1}\log|\mathscr{F}|$*

Corollary 13. *The upper bounds of Corollary 12 are tight[4]:*

- *For $p \geq 2$, for any class $\mathscr{F}$ of uniformly bounded functions with a lower bound of β^{-p} on sequential entropy growth, $\frac{1}{n}V_n \geq \widetilde{\Omega}(n^{-1/p})$*
- *For $p \in (0,2]$, for any class $\mathscr{F}$ of uniformly bounded functions, there exists a slightly modified class $\mathscr{F}'$ with the same sequential entropy growth such that $\frac{1}{n}V_n \geq \widetilde{\Omega}(n^{-2/(2+p)})$*
- *There exists a class $\mathscr{F}$ with the covering number as in (6), such that $\frac{1}{n}V_n \geq \Omega(dn^{-1}\log(n))$*

[3] For $p = 2$, $\frac{1}{n}V_n \leq C\log(n)n^{-1/2}$.

[4] The $\widetilde{\Omega}(\cdot)$ notation suppresses logarithmic factors

5.3 q-loss for $q \in (1,2)$

Consider the case of $\ell(\hat{y}, y) = |y - \hat{y}|^q$, for $q \in (1,2)$, which interpolates between the absolute value and square losses.

Corollary 14. *Suppose $\mathscr{Y} = \widehat{\mathscr{Y}} = [-1,1]$ and $\ell(\hat{y}, y) = |y - \hat{y}|^q$ for $q \in (1,2)$. Assume complexity of $\mathscr{F}$ as in Theorems 10 and 11 for some $p > 0$. Then*

$$\frac{1}{n} V_n = \Theta\left(\min\left\{(q-1)^{-\frac{2-p}{2+p}} n^{-\frac{2}{2+p}}, n^{-1/2}\right\}\right)$$

5.4 q-loss for $q \geq 2$

It is easy to check that for $q > 2$, $\boldsymbol{\ell}(\cdot, y) = |\cdot - y|^q$ is q-uniformly convex, and thus

$$\underline{\boldsymbol{\Delta}}(t) \geq C t^q$$

The upper bound of

$$\frac{1}{n} V_n \leq C n^{-\frac{q}{2(q-1)+p}}$$

then follows from Theorem 10.

5.5 Logistic loss

The loss function $\boldsymbol{\ell}(\hat{y}, y) = \log(1 + \exp\{-\hat{y}y\})$ is strongly convex and smooth if the sets $\mathscr{Y}, \widehat{\mathscr{Y}}$ are bounded. This can be seen by computing the second derivative with respect to the first argument:

$$\boldsymbol{\ell}''(\hat{y}, y) = y^2 \frac{\exp\{\hat{y}y\}}{(1 + \exp\{\hat{y}y\})^2}$$

We conclude that

$$\frac{1}{n} V_n = \widetilde{\Theta}\left(\min\left\{n^{-\frac{2}{2+p}}, n^{-1/2}\right\}\right)$$

Logistic loss is an example of a function with third derivative bounded by a multiple of the second derivative. Control of the remainder term in Taylor approximation for such functions is given in [5, Lemma 1]. Other examples of strongly convex and smooth losses are the exponential loss and truncated quadratic loss. These enjoy the same minimax rate as given above.

5.6 Logarithmic loss

The technique developed in this paper is not universal. In particular, it does not yield correct rates for rich classes of functions under the loss

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

for the problem of probability assignment and a binary alphabet $\mathscr{Y} = \{0,1\}$. The suboptimality of Lemma 3 is due to the exploding Lipschitz constant. However, a modified approach is possible, and will be carried out in a separate paper.

5.7 Sparse linear predictors and square loss

We now focus on quadratic loss and instead detail minimax rates for specific classes of functions. Consider the following parametric class. Let $\mathscr{G} = \{g_1, \ldots, g_M\}$ be a set of M functions such that each $g_i : \mathscr{X} \mapsto [-1,1]$. Define $\mathscr{F}$ to be the convex combination of at most s out of these M functions. That is

$$\mathscr{F} = \left\{\sum_{j=1}^{s} \alpha_j g_{\sigma_j} : \sigma_{1:s} \subset [M], \forall j, \alpha_j \geq 0, \sum_{j=1}^{s} \alpha_j = 1\right\}$$

For this example note that the sequential covering number can be easily upper bounded: we can choose s out of M functions in $\binom{M}{s}$ ways and observe that pointwise metric entropy for convex combination of s bounded functions at scale β is bounded as β^{-s}. We conclude that

$$\mathcal{N}_2(\beta, \mathcal{F}, n) \le \left(\frac{eM}{s}\right)^s \beta^{-s}$$

From the main theorem, for the case of square loss, the upper bound is

$$\tfrac{1}{n} V_n \le O\left(\frac{s \log(M/s)}{n}\right) .$$

The extension to other loss functions follows immediately from the general statements.

5.8 Besov spaces and square loss

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$. Let $\mathcal{F}$ be a ball in Besov space $B^s_{p,q}(\mathcal{X})$. When $s > d/p$, pointwise metric entropy bounds at scale β scales as $\Omega(\beta^{-d/s})$ [31, p. 20]. On the other hand, when $s < d/p$, and $p > 2$, one can show that the space is a p-uniformly convex Banach space. From [26], it can be shown that sequential Rademacher can be upper bounded by $O(n^{1-1/p})$, yielding a bound on minimax rate. These two controls together give the bound on the minimax rate. The generic forecaster with Rademacher complexity as relaxation (see Section 6), enjoys the best of both of these rates. More specifically, we may identify the following regimes:

- If $s \ge d/2$, the minimax rate is $\frac{1}{n} V_n \le O\left(n^{-\frac{2s}{2s+d}}\right)$.
- If $s < d/2$, the minimax rate depends on the interaction of p and d, s:
  - if $p > \frac{d}{s}$, the minimax rate $\frac{1}{n} V_n \le O\left(n^{-\frac{s}{d}}\right)$, otherwise, the rate is $\frac{1}{n} V_n \le O\left(n^{-\frac{1}{p}}\right)$

5.9 Remarks: Experts, Mixability, and Discretization

The problem of prediction with expert advice has been central in the online learning literature [9]. One can phrase the experts problem in our setting by taking a finite class $\mathcal{F} = \{f^1, \dots, f^N\}$ of functions. It is possible to ensure sublinear regret by following the "advice" $f^{I_t}(x_t)$ of a randomly chosen "expert" I_t from an appropriate distribution over experts. The randomized approach, however, effectively linearizes the problem and does not take advantage of the curvature of the loss. The precise way in which the loss enters the picture has been investigated thoroughly by Vovk [28] (see also [15]). Vovk defines a mixability curve that parametrizes achievable regret of a form slightly different than (1). Specifically, Vovk allows a constant other than 1 in front of the infimum in the regret definition. Such regret bounds are called "inexact oracle inequalities" in statistics. Audibert [2] shows that the mixability condition on the loss function leads to a variance-type bound in his general PAC-based formulation, yet the analysis is restricted to the case of finite experts. While it is possible to repeat the analysis in the present paper with a constant other than 1 in front of the comparator, this goes beyond the scope of the paper. Importantly, our techniques go beyond the finite case and can give correct regret bounds even if discretization to a finite set of experts yields vacuous bounds.

Let us emphasize the above point again by comparing the upper bound of Lemma 5 to the bound we may obtain via a metric entropy approach, as in the work of [31]. Assume that $\mathcal{F}$ is a compact subset of $C(\mathcal{X})$ equipped with supremum norm. The metric entropy, denoted by $\mathcal{H}(\epsilon, \mathcal{F})$, is the logarithm of the smallest ϵ-net with respect to the sup norm on $\mathcal{X}$. An aggregating procedure over the elements of the net gives an upper bound (omitting constants and logarithmic factors)

$$n\epsilon + \mathcal{H}(\epsilon, \mathcal{F}) \tag{18}$$

on regret (1). Here, $n\epsilon$ is the amount we lose from restricting the attention to the ϵ-net, and the second term appears from aggregation over a finite set. The balance (18) fails to capture the optimal behavior for large nonparametric sets of functions. Indeed, for an $O(\epsilon^{-p})$ behavior of metric entropy, Vovk concludes the rate of $O\left(n^{\frac{p}{p+1}}\right)$. For

$p \le 2$, this is slower than the $O\left(n^{\frac{p}{p+2}}\right)$ rate one obtains from Lemma 5 by trivially upper bounding the sequential entropy by metric entropy. The gain is due to the chaining technique, a phenomenon well-known in statistical learning theory. Our contribution is to introduce the same concepts to the domain of online learning.

6 Relaxations and Algorithms

To design generic forecasters for the problem of online non-parametric regression we follow the recipe provided in [19]. It was shown in that paper that if one can find a relaxation $\mathbf{Rel}_n$ (a sequence of mappings from observed data to reals) that satisfies certain conditions, then one can define prediction strategies based on such relaxations. Specifically, we look for relaxations that satisfy the initial condition

$$\mathbf{Rel}_n\left(x_{1:n}, y_{1:n}\right) \ge -\inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t)$$

and the recursive admissibility condition that requires

$$\inf_{\widehat{y}_t}\sup_{y_t}\left\{\boldsymbol{\ell}(\widehat{y}_t, y_t) + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)\right\} \le \mathbf{Rel}_n\left(x_{1:t-1}, y_{1:t-1}\right) \tag{19}$$

for any $t \in [n]$ and any $x_t \in \mathcal{X}$. A relaxation $\mathbf{Rel}_n$ satisfying these two conditions is said to be admissible, and it leads to an algorithm

$$\widehat{y}_t = \underset{\widehat{y}\in\widehat{\mathcal{Y}}}{\operatorname{argmin}} \sup_{y_t\in\mathcal{Y}} \left\{\boldsymbol{\ell}(\widehat{y}, y_t) + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)\right\}. \tag{20}$$

For this forecast the associated bound on regret is

$$\mathbf{Reg}_n := \sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \le \mathbf{Rel}_n(\emptyset) \tag{21}$$

(see [19] for details). We now claim that the following conditional version of (8) gives an admissible relaxation and leads to a method that enjoys the regret bounds shown in the first part of the paper.

Lemma 15. *The following relaxation is admissible:*

$$\mathfrak{R}_n(x_{1:t}, y_{1:t}) = \sup_{\mathbf{x},\boldsymbol{\mu}} \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left[\sum_{j=t+1}^{n} 2G\epsilon_j(f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon)) - \underline{\boldsymbol{\Delta}}\Big(f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon)\Big) - \sum_{j=1}^{t} \ell(f(x_j), y_j)\right].$$

The algorithm (20) *with this relaxation enjoys the regret bound of offset Rademacher complexity*

$$\mathbf{Reg}_n \le \sup_{\mathbf{x},\boldsymbol{\mu}} \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left[\sum_{t=1}^{n} 2G\epsilon_t(f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - \underline{\boldsymbol{\Delta}}\Big(f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)\Big)\right].$$

The proof of Lemma 15 follows closely the proof of Lemma 3 and we omit it (see [19, 20]). Since the regret bound for the above forecaster is exactly the one given in (8), the upper bounds in Corollary 12 hold for the above algorithm. Therefore, the algorithm based on $\mathfrak{R}_n(x_{1:t}, y_{1:t})$ is optimal up to the tightness of the upper and lower bounds in Section 4 and Section 3.

For the rest of this section, we restrict our attention to the case when $\mathcal{Y} = \widehat{\mathcal{Y}} = [-B, B]$. We further assume that $\ell(\widehat{y}, y_t) + \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, y_t)\right)$ is a convex function of y_t. In this case, the prediction $\widehat{y}_t$ takes a simple form, as the supremum over y_t is attained either at B or $-B$. More precisely, the prediction can be written as

$$\widehat{y}_t = \underset{\widehat{y}\in[-B,B]}{\operatorname{argmin}} \max\left\{\ell(\widehat{y}, B) + \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, B)\right), \ell(\widehat{y}, -B) + \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, -B)\right)\right\}. \tag{22}$$

6.1 Recipe for designing online regression algorithms for general loss functions

We now provide a schema for deriving forecasters for general online non-parametric regression:

1. Find relaxation $\mathbf{Rel}_n$ s.t. $\mathfrak{R}_n\left(x_{1:t}, y_{1:t}\right) \le \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)$, and $\ell(\widehat{y}, y_t) + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t-1}, y_t\right)$ is convex in y_t.
2. Check the condition
$$\sup_{x_t \in \mathscr{X}, p_t \in \Delta([-B,B])} \left\{ \inf_{\widehat{y}_t} \mathbb{E}_{y_t \sim p_t}\left[\ell\left(\widehat{y}_t, y_t\right)\right] + \mathbb{E}_{y_t \sim p_t}\left[\mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)\right] \right\} \le \mathbf{Rel}_n\left(x_{1:t-1}, y_{1:t-1}\right)$$
3. Given x_t, the prediction $\widehat{y}_t$ is given by
$$\widehat{y}_t = \underset{\widehat{y} \in [-B,B]}{\operatorname{argmin}} \max\left\{\ell(\widehat{y}, B) + \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, B)\right), \ell(\widehat{y}, -B) + \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, -B)\right)\right\}$$

Proposition 16. *For any algorithm derived from the above schema,* $\mathbf{Reg}_n \le \mathbf{Rel}_n\left(\emptyset\right)$.

The proof of this Proposition follows very closely the proof in [19] (see also [20]), and we omit it.

6.1.1 Square Loss

To provide concrete examples of how the recipe can be used to derive algorithms, we now consider the square loss setting, $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$. In this case, we observe that in (22), the first term in the maximum decreases as $\widehat{y}$ increases to B and likewise the second term monotonically decreases as $\widehat{y}$ decreases to $-B$. Hence, the solution to (22) is given when both terms are equal (if this does not happen within the range $[-B, B]$ then we clip the prediction to this range). In other words, for the case of square loss, if we have an admissible relaxation, then the prediction based on this relaxation is simply given by:

$$\widehat{y}_t = \operatorname{Clip}\left(\frac{\mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, B)\right) - \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, -B)\right)}{4B}\right)$$

where $\operatorname{Clip}(z) = B\mathbf{1}\{z > B\} + (-B)\mathbf{1}\{z < -B\} + z\mathbf{1}\{z \in [-B, B]\}$. Hence, for any admissible relaxation such that $(\widehat{y} - y_t)^2 + \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, y_t)\right)$ is a convex function of y_t, the above prediction based on the relaxation enjoys the bound on regret $\mathbf{Rel}_n$. Based on the above observations and the recipe outline, we now provide two examples of $\mathscr{F}$ for which algorithms are derived.

Example : Finite class of experts
As an example of estimator derived from the schema for the square loss learning setting, we first consider the simple case $|\mathscr{F}| < \infty$.

Corollary 17. *The following is an admissible relaxation:*

$$\mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right) = B^2 \log\left(\sum_{f \in \mathscr{F}} \exp\left(-B^{-2} \sum_{j=1}^{t} (f(x_j) - y_j)^2\right)\right).$$

It leads to the algorithm

$$\widehat{y}_t = \operatorname{Clip}\left(\frac{B}{4}\log\left(\frac{\sum_{f\in\mathscr{F}} \exp\left(-B^{-2}\sum_{j=1}^{t-1}(f(x_j)-y_j)^2 - B^{-2}(f(x_t)-B)^2\right)}{\sum_{f\in\mathscr{F}} \exp\left(-B^{-2}\sum_{j=1}^{t-1}(f(x_j)-y_j)^2 - B^{-2}(f(x_t)+B)^2\right)}\right)\right)$$

which enjoys a regret bound $\mathbf{Reg}_n \le B^2 \log|\mathscr{F}|$.

Example : Linear regression
Next, consider the problem of online linear regression in $\mathbb{R}^d$. Here $\mathscr{F}$ is the class of linear functions. For this problem we consider a slightly modified notion of regret,

$$\sum_{t=1}^{n} (\widehat{y}_t - y_t)^2 - \inf_{f \in \mathscr{F}} \left\{ \sum_{t=1}^{n} (f^\top x_t - y_t)^2 + \lambda \left\| f \right\|_2^2 \right\}.$$

This regret can be seen alternatively as regret if we assume that on rounds $-d+1$ to 0 Nature plays $(\lambda e_1, 0), \ldots,$ $(\lambda e_d, 0)$, where $\{e_i\}$ are the standard basis vectors, and that on these rounds the learner (knowing this) predicts 0, thus incurring zero loss over these initial rounds. We can readily apply the schema for designing an algorithm for this problem.

Corollary 18. *For any $\lambda > 0$, the following is an admissible relaxation:*

$$\mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right) = \left\| \sum_{j=1}^{t} y_j z_j \right\|^2_{\left(\sum_{j=1}^{t} z_j z_j^\top + \lambda I\right)^{-1}} + 4B^2 \log \left(\frac{\left(\frac{n}{d}\right)^d}{\Delta\left(\sum_{j=1}^{t} z_j z_j^\top + \lambda I\right)} \right) - \sum_{j=1}^{t} y_j^2 \,.$$

It leads to the Vovk-Azoury-Warmuth forecaster [29, 4]

$$\widehat{y}_t = \mathrm{Clip}\left(x_t^\top \left(\sum_{j=1}^{t} x_j x_j^\top + \lambda I \right)^{-1} \left(\sum_{j=1}^{t-1} y_j x_j \right) \right)$$

and enjoys the regret bound

$$\frac{1}{n} \sum_{t=1}^{n} (\widehat{y}_t - y_t)^2 \leq \frac{1}{n} \sum_{t=1}^{n} (f^\top x_t - y_t)^2 + \frac{\lambda}{2n} \left\| f \right\|_2^2 + \frac{4dB^2 \log\left(\frac{n}{\lambda d}\right)}{n}.$$

The proofs of Corollaries 17 and 18 already appeared in [23], and we omit them here.

7 Discussion and Related Work

In the past twenty years, progress in online regression for arbitrary sequences, starting with the paper of [10], has been almost exclusively on finite-dimensional *linear* regression (an incomplete list includes [29, 15, 17, 30, 6, 3, 4, 16, 11]). This is to be contrasted with Statistics, where regression has been studied for rich (nonparametric) classes of functions. Important exceptions to this limitation in the online regression framework – and works that partly motivated the present findings – are the papers of [33, 31, 32]. Vovk considers regression with large classes, such as subsets of a Besov or Sobolev space, and remarks that there appears to be two distinct approaches to obtaining the upper bounds in online competitive regression. The first approach, which Vovk terms Defensive Forecasting, exploits uniform convexity of the space, while the second – an aggregating technique (such as the Exponential Weights Algorithm) – is based on the metric entropy of the space. Interestingly, the two seemingly different approaches yield distinct upper bounds, based on the respective properties of the space. In particular, Vovk asks whether there is a unified view of these techniques. The present paper addresses these questions and establishes optimal performance for online regression.

Since most work in online learning is algorithmic, the boundaries of what can be proved are defined by the regret minimization algorithms one can find. One of the main algorithmic workhorses is the aggregating procedure (or, the Exponential Weights Algorithm). However, the difficulty in using an aggregating procedure beyond simple parametric classes (e.g. subsets of $\mathbb{R}^d$) lies in the need for a “pointwise” cover of the set of functions – that is, a data-independent cover in the supremum norm on the underlying space of covariates. The same difficulty arises when one uses PAC-Bayesian bounds [2] that, at the end of the day, require a volumetric argument. Notably, this difficulty has been overcome in statistical learning, where it has long been recognized (since the work of Vapnik and Chervonenkis) that it is sufficient to consider an *empirical* cover of the class – a potentially much smaller quantity. Such an empirical entropy is necessarily finite, and its growth with n is one of the key complexity measures for i.i.d. learning. In particular, the recent work of [27] shows that the behavior of empirical entropy characterizes the optimal rates for i.i.d. learning with square loss. To mimic this development, it appears that we need to understand empirical covering numbers in the sequential prediction framework.

A hint as to how to modify the analysis of [24] for "curved" losses appears in the paper of [8] where the authors derived rates for log-loss via a two-level procedure: the set of densities is first partitioned into small balls of a critical radius γ; a minimax algorithm is employed on each of these small balls; and an overarching aggregating procedure combines these algorithms. Regret within each small ball is upper bounded by classical Dudley entropy integral (with respect to a pointwise metric) defined up to the γ radius. The main technical difficulty in this paper is to prove a similar statement using "empirical" sequential covering numbers.

Interestingly, our results imply the same phase transition as the one exhibited in [26] for i.i.d. learning with square loss. More precisely, under the assumption of the $O(\beta^{-p})$ behavior of sequential entropy, the minimax regret normalized by time horizon n decays as $n^{-\frac{2}{2+p}}$ if $p \in (0,2]$, and as $n^{-1/p}$ for $p \geq 2$. We prove lower bounds that match up to a logarithmic factor, establishing that the phase transition is real. Even more surprisingly, it follows that, under a mild assumption that sequential Rademacher complexity of $\mathscr{F}$ behaves similarly to its i.i.d. cousin, *the rates of minimax regret in online regression with arbitrary sequences match, up to a logarithmic factor, those in the i.i.d. setting of Statistical Learning.* This phenomenon has been noticed for some parametric classes by various authors (e.g. [7]). The phenomenon is even more striking given the simple fact that one may convert the regret statement, that holds for all sequences, into an i.i.d. guarantee. Thus, in particular, we recover the result of [27] through completely different techniques. Since in many situations, one obtains optimal rates for i.i.d. learning from a regret statement, the relaxation framework of [19] provides a toolkit for developing improper learning algorithms in the i.i.d. scenario.

A Proofs

Proof of Lemma 3. Denoting the set of distributions on $\mathscr{Y}$ by $\mathscr{P}$, minimax regret can be written as

$$\begin{aligned}
V_n &= \left\langle\!\!\left\langle \sup_{x_t} \inf_{q_t} \sup_{p_t \in \mathscr{P}} \mathop{\mathbb{E}}_{\substack{\widehat{y}_t \sim q_t \\ y_t \sim p_t}} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{y}_t, y_t) - \inf_{f \in \mathscr{F}} \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \right\} && (23)\\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \mathscr{P}} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sum_{t=1}^{n} \inf_{\widehat{y}_t} \mathbb{E}_{y_t}\left[\boldsymbol{\ell}(\widehat{y}_t, y_t)\right] - \inf_{f \in \mathscr{F}} \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \right\}\\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \mathscr{P}} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}} \left\{ \sum_{t=1}^{n} \inf_{\widehat{y}_t} \{\mathbb{E}_{y_t}\left[\boldsymbol{\ell}(\widehat{y}_t, y_t)\right]\} - \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \right\} \right]
\end{aligned}$$

where the first equality is by definition, the second follows from an argument of [1, 24], and the third is a simple rearrangement. Taking $\widehat{y}_t^* = \operatorname*{argmin}_{\widehat{y} \in \widehat{\mathscr{Y}}} \mathbb{E}_{y \sim p_t}\left[\boldsymbol{\ell}(\widehat{y}, y)\right]$, we write the above as

$$\begin{aligned}
&\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \mathscr{P}} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}} \left\{ \sum_{t=1}^{n} \mathbb{E}_{y_t}\left[\boldsymbol{\ell}(\widehat{y}_t^*, y_t)\right] - \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \right\} \right] && (24)\\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \mathscr{P}} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}} \left\{ \sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{y}_t^*, y_t) - \boldsymbol{\ell}(f(x_t), y_t) \right\} \right] && (25)
\end{aligned}$$

The last step holds true by observing that the terms $\boldsymbol{\ell}(\widehat{y}_t^*, y_t)$ do not depend on f and can therefore be moved outside the supremum over $f \in \mathscr{F}$. The equality then follows by the linearity of expectation. By definition of $\underline{\boldsymbol{\Delta}}$ in (2),

$$V_n \leq \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \mathscr{P}} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathscr{F}} \left\{ \sum_{t=1}^{n} \partial\boldsymbol{\ell}(\widehat{y}_t^*, y_t) \cdot \left(\widehat{y}_t^* - f(x_t)\right) - \underline{\boldsymbol{\Delta}}(\widehat{y}_t^* - f(x_t)) \right\} \right] \qquad (26)$$

By definition of $\widehat{y}_t^*$, we have, $\mathbb{E}_{y_t\sim p_t}\left[\partial\boldsymbol{\ell}(\widehat{y}_t^*,y_t)\right] = \partial\mathbb{E}_{y_t\sim p_t}\left[\boldsymbol{\ell}(\widehat{y}_t^*,y_t)\right] = 0$ by the assumption that the minimum is attained in $\widehat{\mathscr{Y}}$ (see Section 2.1). Thus we can view $\left(\partial\boldsymbol{\ell}(\widehat{y}_t^*,y_t)\right)_{t=1}^T$ as a martingale difference sequence. Hence,

$$V_n \le \left\langle\!\!\left\langle \sup_{x_t}\sup_{p_t\in\mathscr{P}}\mathbb{E}_{y_t}\right\rangle\!\!\right\rangle_{t=1}^n \left[\sup_{f\in\mathscr{F}}\left\{\sum_{t=1}^n \left(\partial\boldsymbol{\ell}(\widehat{y}_t^*,y_t)-\mathbb{E}_{y_t'\sim p_t}\left[\partial\boldsymbol{\ell}(\widehat{y}_t^*,y_t')\right]\right)\cdot\left(\widehat{y}_t^*-f(x_t)\right)-\underline{\boldsymbol{\Delta}}(\widehat{y}_t^*-f(x_t))\right\}\right].$$

By Jensen's inequality the above is upper bounded by

$$\left\langle\!\!\left\langle \sup_{x_t}\sup_{p_t\in\mathscr{P}}\mathbb{E}_{y_t,y_t'\sim p_t}\mathbb{E}_{\epsilon_t}\right\rangle\!\!\right\rangle_{t=1}^n \left[\sup_{f\in\mathscr{F}}\left\{\sum_{t=1}^n \epsilon_t\left(\partial\boldsymbol{\ell}(\widehat{y}_t^*,y_t)-\partial\boldsymbol{\ell}(\widehat{y}_t^*,y_t')\right)\cdot\left(\widehat{y}_t^*-f(x_t)\right)-\underline{\boldsymbol{\Delta}}(\widehat{y}_t^*-f(x_t))\right\}\right]$$

where we introduced Rademacher random variables. The next step involves splitting the upper bound into two equal terms, one for the y_t sequence and the other for the y_t' sequence:

$$V_n \le \left\langle\!\!\left\langle \sup_{x_t}\sup_{p_t\in\mathscr{P}}\mathbb{E}_{y_t\sim p_t}\mathbb{E}_{\epsilon_t}\right\rangle\!\!\right\rangle_{t=1}^n \left[\sup_{f\in\mathscr{F}}\left\{\sum_{t=1}^n 2\epsilon_t\partial\boldsymbol{\ell}(\widehat{y}_t^*,y_t)\cdot\left(\widehat{y}_t^*-f(x_t)\right)-\underline{\boldsymbol{\Delta}}(\widehat{y}_t^*-f(x_t))\right\}\right]$$

Using Jensen's inequality once again leads to an upper bound of

$$\left\langle\!\!\left\langle \sup_{x_t}\sup_{p_t\in\mathscr{P}}\mathbb{E}_{y_t\sim p_t}\mathbb{E}_{\epsilon_t}\right\rangle\!\!\right\rangle_{t=1}^n \left[\sup_{f\in\mathscr{F}}\left\{\sum_{t=1}^n 2\epsilon_t\partial\boldsymbol{\ell}(\widehat{y}_t^*,y_t)\cdot\left(\widehat{y}_t^*-f(x_t)\right)-\underline{\boldsymbol{\Delta}}(\widehat{y}_t^*-f(x_t))\right\}\right]$$

Now, observe that y_t^* is a function of p_t and $\partial\boldsymbol{\ell}(\widehat{y}_t^*,y_t)$ is a function of y_t and p_t. Hence, we may pass to a further upper bound by inserting $\sup_{\eta_t,\mu_t}$, and by replacing each subgradient with the respective η_t and each $\widehat{y}_t^*$ with μ_t:

$$\begin{aligned}
&\left\langle\!\!\left\langle \sup_{x_t}\sup_{p_t\in\mathscr{P}}\mathbb{E}_{y_t\sim p_t}\sup_{\eta_t\in[-G,G]}\sup_{\mu_t}\mathbb{E}_{\epsilon_t}\right\rangle\!\!\right\rangle_{t=1}^n \left[\sup_{f\in\mathscr{F}}\left\{\sum_{t=1}^n 2\epsilon_t\eta_t\cdot\left(\mu_t-f(x_t)\right)-\underline{\boldsymbol{\Delta}}(\mu_t-f(x_t))\right\}\right]\\
&=\left\langle\!\!\left\langle \sup_{x_t\in\mathscr{X}}\sup_{\eta_t\in[-G,G]}\sup_{\mu_t}\mathbb{E}_{\epsilon_t}\right\rangle\!\!\right\rangle_{t=1}^n \left[\sup_{f\in\mathscr{F}}\left\{\sum_{t=1}^n 2\epsilon_t\eta_t\cdot\left(\mu_t-f(x_t)\right)-\underline{\boldsymbol{\Delta}}(\mu_t-f(x_t))\right\}\right]
\end{aligned}$$

Since each η_t range over $[-G,G]$, we can represent it as G times the expectation of a random variable $u_t\in\{-1,1\}$. Denoting this distribution by q_t, by Jensen's inequality

$$\begin{aligned}
V_n &\le \left\langle\!\!\left\langle \sup_{x_t,\mu_t,q_t}\mathbb{E}_{\epsilon_t}\right\rangle\!\!\right\rangle_{t=1}^n \left\{\sup_{f\in\mathscr{F}}\left[\sum_{t=1}^n 2G\epsilon_t\mathbb{E}(u_t)(\mu_t-f(x_t))-\underline{\boldsymbol{\Delta}}(\mu_t-f(x_t))\right]\right\}\\
&\le \left\langle\!\!\left\langle \sup_{x_t,\mu_t,q_t}\mathbb{E}_{u_t}\mathbb{E}_{\epsilon_t}\right\rangle\!\!\right\rangle_{t=1}^n \left\{\sup_{f\in\mathscr{F}}\left[\sum_{t=1}^n 2G\epsilon_t u_t(f(x_t)-\mu_t)-\underline{\boldsymbol{\Delta}}(\mu_t-f(x_t))\right]\right\}\\
&= \left\langle\!\!\left\langle \sup_{x_t,\mu_t}\max_{u_t\in\{-1,1\}}\mathbb{E}_{\epsilon_t}\right\rangle\!\!\right\rangle_{t=1}^n \left\{\sup_{f\in\mathscr{F}}\left[\sum_{t=1}^n 2G\epsilon_t u_t(f(x_t)-\mu_t)-\underline{\boldsymbol{\Delta}}(\mu_t-f(x_t))\right]\right\}
\end{aligned}$$

Since for any fixed $u_t\in\{+1,-1\}$, the distribution of $\epsilon_t u_t$ is the same as that of ϵ_t, the above expression is simply

$$\begin{aligned}
&\left\langle\!\!\left\langle \sup_{x_t,\mu_t}\mathbb{E}_{\epsilon_t}\right\rangle\!\!\right\rangle_{t=1}^n \left\{\sup_{f\in\mathscr{F}}\left[\sum_{t=1}^n 2G\epsilon_t(f(x_t)-\mu_t)-\underline{\boldsymbol{\Delta}}(\mu_t-f(x_t))\right]\right\}\\
&=\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_{\epsilon}\left[\sup_{f\in\mathscr{F}}\left\{\sum_{t=1}^n 2G\epsilon_t\left(f(\mathbf{x}_t(\epsilon)-\boldsymbol{\mu}_t(\epsilon))\right)-\underline{\boldsymbol{\Delta}}\left(\boldsymbol{\mu}_t(\epsilon)-f(\mathbf{x}_t(\epsilon))\right)\right\}\right]
\end{aligned}$$

which is the same as the desired upper bound in (8), in the tree notation. □

Proof of Lemma 4. It holds that

$$\begin{aligned}
&\mathbb{E}_\epsilon\left[\max_{\mathbf{w}\in W}\left\{\sum_{t=1}^n 2C\epsilon_t\mathbf{w}_t(\epsilon)-\Delta\left(\mathbf{w}_t(\epsilon)\right)\right\}\right]\\
&=\mathbb{E}_\epsilon\left[\inf_{\lambda>0}\frac{1}{\lambda}\log\left(\sum_{\mathbf{w}\in W}\exp\left(\lambda\left(\sum_{t=1}^n 2C\epsilon_t\mathbf{w}_t(\epsilon)-\Delta\left(\mathbf{w}_t(\epsilon)\right)\right)\right)\right)\right]
\end{aligned}$$

which, by Jensen's inequality, is upper bounded by

$$\begin{aligned}
&\inf_{\lambda>0}\left\{\frac{1}{\lambda}\log\left(\sum_{\mathbf{w}\in W}\mathbb{E}_\epsilon\left[\exp\left(\lambda\left(\sum_{t=1}^n 2C\epsilon_t\mathbf{w}_t(\epsilon)-\Delta\left(\mathbf{w}_t(\epsilon)\right)\right)\right)\right]\right)\right\}\\
&=\inf_{\lambda>0}\left\{\frac{1}{\lambda}\log\left(\sum_{\mathbf{w}\in W}\mathbb{E}_\epsilon\left[\prod_{t=1}^n \exp\left(\lambda\left(2C\epsilon_t\mathbf{w}_t(\epsilon)-\Delta\left(\mathbf{w}_t(\epsilon)\right)\right)\right)\right]\right)\right\}
\end{aligned}$$

Since $e^x+e^{-x}\le 2e^{x^2/2}$, we have that

$$\begin{aligned}
\mathbb{E}_{\epsilon_n}\left[\exp\left(\lambda\left(2C\epsilon_n\mathbf{w}_n(\epsilon)-\Delta\left(\mathbf{w}_n(\epsilon)\right)\right)\right)\right]&\le\exp\left(2C^2\lambda^2\mathbf{w}_n(\epsilon)^2-\lambda\Delta\left(\mathbf{w}_n(\epsilon)\right)\right)\\
&=\exp\left(2C^2\lambda^2\mathbf{w}_n(\epsilon)^2-\lambda\Gamma\left(\mathbf{w}_n(\epsilon)^2\right)\right)
\end{aligned}$$

By definition of conjugacy, we pass to a further upper bound of

$$\exp\left(\lambda\Gamma^*\left(2C^2\lambda\right)\right)\le\exp\left(\lambda\Gamma^*\left(2C^2\lambda\right)\right)$$

where the last step is because Γ^* is non-decreasing. Hence we have that

$$\begin{aligned}
&\frac{1}{\lambda}\log\left(\mathbb{E}_{\epsilon_{1:n}}\left[\prod_{t=1}^n\exp\left(\lambda\left(2C\epsilon_t\mathbf{w}_t(\epsilon)-\Delta\left(\mathbf{w}_t(\epsilon)\right)\right)\right)\right]\right)\\
&\qquad\le\frac{1}{\lambda}\log\left(\mathbb{E}_{\epsilon_{1:n-1}}\left[\prod_{t=1}^{n-1}\exp\left(\lambda\left(2C\epsilon_t\mathbf{w}_t(\epsilon)-\Delta\left(\mathbf{w}_t(\epsilon)\right)\right)\right)\right]\right)+\Gamma^*\left(2C^2\lambda\right)
\end{aligned}$$

Proceeding in similar fashion from $n-1$ down to 1, we arrive at an upper bound of

$$\frac{1}{\lambda}\log|W|+n\Gamma^*\left(2C^2\lambda\right)$$

This proves the first claim. The second statement (which already appears in [24]) is proved similarly, except the tuning value λ is chosen at the end, and we need to account for the worst-case ℓ_2 norm along any paths. We provide the proof here for completeness. For any tree $\mathbf{w}\in W$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\sum_{t=1}^n\lambda\epsilon_t\mathbf{w}_t(\epsilon)\right\}\,\middle|\,\epsilon_{1:n-1}\right]&\le\exp\left\{\sum_{t=1}^{n-1}\lambda\epsilon_t\mathbf{w}_t(\epsilon)\right\}\exp\left\{\lambda^2\mathbf{w}_n(\epsilon)^2/2\right\}\\
&\le\exp\left\{\sum_{t=1}^{n-1}\lambda\epsilon_t\mathbf{w}_t(\epsilon)\right\}\max_{\epsilon_n}\exp\left\{\lambda^2\mathbf{w}_n(\epsilon)^2/2\right\}
\end{aligned}$$

Continuing in this fashion backwards to $t=1$, for any $\mathbf{w}\in W$

$$\mathbb{E}\left[\exp\left\{\sum_{t=1}^n\lambda\epsilon_t\mathbf{w}_t(\epsilon)\right\}\right]\le\max_{\epsilon_1,\dots,\epsilon_n}\exp\left\{(\lambda^2/2)\sum_{t=1}^n\mathbf{w}_n(\epsilon)^2\right\}$$

and thus

$$\mathbb{E}\left[\sum_{\mathbf{w}\in W}\exp\left\{\sum_{t=1}^n\lambda\epsilon_t\mathbf{w}_t(\epsilon)\right\}\right]\le|W|\max_{\epsilon_1,\dots,\epsilon_n}\max_{\mathbf{w}\in W}\exp\left\{(\lambda^2/2)\sum_{t=1}^n\mathbf{w}_n(\epsilon)^2\right\}.$$

Choosing

$$\lambda = \sqrt{\frac{2\log|W|}{\max_{\epsilon_{1:n},\mathbf{w}\in W}\sum_{t=1}^n \mathbf{w}_n(\epsilon)^2}}$$

we obtain

$$\begin{aligned}\mathbb{E}\max_{\mathbf{w}\in W}\left[\sum_{t=1}^n \epsilon_t \mathbf{w}_t(\epsilon)\right] &\le \frac{1}{\lambda}\log\mathbb{E}\left[\sum_{\mathbf{w}\in W}\exp\left\{\sum_{t=1}^n \lambda\epsilon_t\mathbf{w}_t(\epsilon)\right\}\right]\\ &\le \sqrt{2\log|W|\cdot \max_{\mathbf{w}\in W,\epsilon_{1:n}}\sum_{t=1}^n \mathbf{w}_n(\epsilon)^2}\end{aligned}$$

□

Proof of Lemma 5. Fix $\gamma > 0$. Let V' be a sequential $\gamma/2$-cover of $\mathcal{G}$ on $\mathbf{z}$ in the ℓ_∞ sense, i.e.

$$\forall\epsilon,\ \ \forall g\in\mathcal{G},\ \ \exists\mathbf{v}'\in V'\ \ \text{s.t.}\ \ \max_{t\in[n]}\left|g(\mathbf{z}_t(\epsilon)) - \mathbf{v}'_t(\epsilon)\right| \le \gamma/2$$

We now modify V' to construct a γ-cover of $\mathcal{G}$ on $\mathbf{z}$, which we shall denote by V. The γ-cover is built as follows. For every $\mathbf{v}'\in V'$ we include $\mathbf{v}$ in V as defined by a soft-thresholding operation:

$$\forall\epsilon\in\{\pm1\}^n, \forall t\in[n],\quad \mathbf{v}_t(\epsilon) = \begin{cases} 0 & \text{if } |\mathbf{v}'_t(\epsilon)| \le \gamma/2\\ \text{sign}(\mathbf{v}'_t(\epsilon))\left(|\mathbf{v}'_t(\epsilon)| - \gamma/2\right) & \text{otherwise}\end{cases}$$

Since we change each $\mathbf{v}'\in V'$ only by $\gamma/2$ on each coordinate, V is indeed a γ-cover in the ℓ_∞ sense. Also note that by the way we constructed V from V', we also have that for every ϵ and any $g\in\mathcal{G}$, there exists a $\mathbf{v}\in V$ that is γ-close in the ℓ_∞ sense and for this $\mathbf{v}$, $|g(\mathbf{z}_t(\epsilon))| \ge |\mathbf{v}_t(\epsilon)|$ for every t. Hence, $\Delta(g(\mathbf{z}_t(\epsilon))) \ge \Delta(\mathbf{v}_t(\epsilon))$ by the assumption that Δ is nondecreasing on $\mathbb{R}_{\ge 0}$ and non-increasing on $\mathbb{R}_{\le 0}$. Denote such a γ-close tree $\mathbf{v}$ by $\mathbf{v}[\epsilon, g]$ to make the dependence on g,ϵ explicit. Since, for all ϵ and all t, $\Delta(g(\mathbf{z}_t(\epsilon))) \ge \Delta(\mathbf{v}[\epsilon,g]_t(\epsilon))$ we have,

$$\begin{aligned}&\mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^n 2C\epsilon_t g(\mathbf{z}_t(\epsilon)) - \Delta\left(g(\mathbf{z}_t(\epsilon))\right)\right] \qquad (27)\\ &= \mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^n 2C\epsilon_t\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t(\epsilon)\Big) + 2C\epsilon_t\mathbf{v}[\epsilon,g]_t(\epsilon) - \Delta(g(\mathbf{z}_t(\epsilon)))\right]\\ &\le \mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^n 2C\epsilon_t\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t(\epsilon)\Big) + 2C\epsilon_t\mathbf{v}[\epsilon,g]_t(\epsilon) - \Delta(\mathbf{v}[\epsilon,g]_t(\epsilon))\right]\end{aligned}$$

Since $\mathbf{v}[\epsilon, g]$ ranges over V, the last expression is upper bounded by

$$\mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^n 2C\epsilon_t\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t(\epsilon)\Big)\right] + \mathbb{E}\max_{\mathbf{v}\in V}\left[\sum_{t=1}^n 2C\epsilon_t\mathbf{v}_t(\epsilon) - \Delta(\mathbf{v}_t(\epsilon))\right] \qquad (28)$$

Now let $\mathbf{v}[\epsilon, g]$ be denoted by $\mathbf{v}^0[\epsilon, g]$ and V be denoted by V^0. Let $\beta_j = 2^{-j}\gamma$ and let V^j denote a sequential β_j-cover of $\mathcal{G}$ on the tree $\mathbf{z}$, for $j = 1,\dots,N$, $N \ge 1$ to be specified later. We can now write the first term in (28) as the constant C times

$$\begin{aligned}&\mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^n 2\epsilon_t\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}^N[\epsilon,g]_t(\epsilon)\Big) + \sum_{t=1}^n\sum_{j=1}^N 2\epsilon_t\Big(\mathbf{v}^j[\epsilon,g]_t(\epsilon) - \mathbf{v}^{j-1}[\epsilon,g]_t(\epsilon)\Big)\right]\\ &\le \mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^n 2\epsilon_t\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}^N[\epsilon,g]_t(\epsilon)\Big)\right] + \sum_{j=1}^N \mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^n 2\epsilon_t\Big(\mathbf{v}^j[\epsilon,g]_t(\epsilon) - \mathbf{v}^{j-1}[\epsilon,g]_t(\epsilon)\Big)\right]\end{aligned}$$

Observe that $|g(\mathbf{z}_t(\epsilon)) - \mathbf{v}^N[\epsilon,g]_t(\epsilon)| \le 2\beta_N$, and hence the first term above is upper-bounded by $4\beta_N n$.

We now upper bound the second term. Fix $\rho \in (0,\gamma)$ and choose $N = \max\{j : \beta_j > 2\rho\}$. Then $\beta_{N+1} \le 2\rho$ and $\beta_N \le 4\rho$. Further, $\beta_{N+1} > \alpha$. Then second term is upper bounded via Lemma 4 by

$$\sum_{j=1}^{N} 3\beta_j \sqrt{2n\,\log(|V^j|\,|V^{j-1}|)} \le 12\sqrt{n}\int_{\rho}^{\gamma}\sqrt{\log\mathcal{N}_\infty(\delta,\mathcal{G},\mathbf{z})}\,d\delta$$

and, finally, the second term in (28) is upper bounded via Lemma 4 by

$$\inf_{\lambda>0}\left\{\frac{1}{\lambda}\log\mathcal{N}_\infty\left(\tfrac{\gamma}{2},\mathcal{G},\mathbf{z}\right) + n\,\Gamma^*\left(2C^2\lambda\right)\right\}$$

Combining the results,

$$\begin{aligned}
&\mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^{n} 2C\epsilon_t g(\mathbf{z}_t(\epsilon)) - \Delta\left(g(\mathbf{z}_t(\epsilon))\right)\right] \\
&\le C\inf_{\rho\in(0,\gamma)}\left\{4\rho n + 12\sqrt{n}\int_{\rho}^{\gamma}\sqrt{\log\mathcal{N}_\infty(\delta,\mathcal{G},\mathbf{z})}\,d\delta\right\} + \inf_{\lambda>0}\left\{\frac{1}{\lambda}\log\mathcal{N}_\infty\left(\tfrac{\gamma}{2},\mathcal{G},\mathbf{z}\right) + n\,\Gamma^*\left(2C^2\lambda\right)\right\}
\end{aligned}$$

Since γ was chosen arbitrarily, the result follows.

□

Proof of Lemma 6. Recall that by definition,

$$\boldsymbol{\ell}(y^*,y_t) - \boldsymbol{\ell}(f(x_t),y_t) = \partial\boldsymbol{\ell}(y^*,y_t)\cdot\left(y^* - f(x_t)\right) - \overline{\boldsymbol{\Delta}}_{y^*,f(x_t)}^{y_t}.$$

From Eq. (24) in the proof of Lemma 3,

$$\begin{aligned}
V_n &= \left\langle\!\!\left\langle \sup_{x_t}\ \sup_{p_t\in\mathcal{P}} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\boldsymbol{\ell}(\widehat{y}_t^*,y_t) - \boldsymbol{\ell}(f(x_t),y_t)\right\}\right] \\
&= \left\langle\!\!\left\langle \sup_{x_t}\ \sup_{p_t\in\mathcal{P}} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\partial\boldsymbol{\ell}(\widehat{y}_t^*,y_t)\cdot\left(\widehat{y}_t^* - f(x_t)\right) - \overline{\boldsymbol{\Delta}}_{y_t^*,f(x_t)}^{y_t}\right\}\right]
\end{aligned}$$

The above inequality holds true if $\widehat{y}_t^*$ ensures $\mathbb{E}_{y\sim p_t}\left[\partial\boldsymbol{\ell}(\widehat{y}_t^*,y)\right] = 0$. Let us now pass to a lower bound by restricting the set of possible optima $\widehat{y}_t^*$ to be in S and the set of associated distributions to be two-point uniform distributions on the corresponding $y_1(\widehat{y}_t^*), y_2(\widehat{y}_t^*)$. Recall that by definition

$$\overline{\boldsymbol{\Delta}}_S(f(x_t) - s) \ge \max\left\{\overline{\boldsymbol{\Delta}}_{s,f(x_t)}^{y_1}, \overline{\boldsymbol{\Delta}}_{s,f(x_t)}^{y_2}\right\}$$

The lower bound is then

$$\begin{aligned}
V_n &\ge R\left\langle\!\!\left\langle \sup_{x_t}\sup_{s_t\in S}\mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\epsilon_t\cdot\left(s_t - f(x_t)\right) - \frac{1}{R}\overline{\boldsymbol{\Delta}}_S(f(x_t) - s_t)\right\}\right] \\
&\ge R\sup_{\mathbf{x}}\mathbb{E}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\epsilon_t\cdot\left(f(\mathbf{x}_t(\epsilon) - \boldsymbol{\mu}_t(\epsilon))\right) - \frac{1}{R}\overline{\boldsymbol{\Delta}}_{\boldsymbol{\mu}_t(\epsilon)}(f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon))\right\}\right]
\end{aligned}$$

for any S-valued tree $\boldsymbol{\mu}$.

□

Proof of Lemma 8. Fix a $\beta > 0$, and set $n = \mathrm{fat}_\beta(\mathcal{F})$. Suppose $\mathbf{x}$ is an $\mathcal{X}$-valued tree of depth n that is β-shattered by $\mathcal{F}$:

$$\forall\epsilon, \exists f^\epsilon \in \mathcal{F} \quad \text{s.t.} \quad \epsilon_t(f^\epsilon(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) \ge \beta/2$$

where $\boldsymbol{\mu}$ is the witness to shattering. Then from (14) with the particular choices of $\mathbf{x}$ and $\boldsymbol{\mu}$ described above,

$$V_n \ge \mathbb{E} \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} R\epsilon_t (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - \overline{\boldsymbol{\Delta}}_{\boldsymbol{\mu}_t(\epsilon)} (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) \right] \tag{29}$$

$$\ge \mathbb{E} \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} R\epsilon_t (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - \frac{R}{2} |f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)| \right] \tag{30}$$

$$\ge \mathbb{E} \left[\sum_{t=1}^{n} R\epsilon_t (f^{\epsilon}(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - \frac{R}{2} |f^{\epsilon}(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)| \right] \tag{31}$$

Using the definition of shattering, we can further lower bound the above quantity by

$$\mathbb{E} \left[\sum_{t=1}^{n} \frac{R}{2} |f^{\epsilon}(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)| \right] \ge \frac{Rn\beta}{2}.$$

Now, suppose $\mathrm{fat}_\beta(\mathscr{F}) = 1/\beta^p$, $p > 0$. Then $n = \mathrm{fat}_\beta(\mathscr{F})$ implies $\beta = n^{-1/p}$. The result follows. □

Proof of Lemma 9. Assume that $d = \mathrm{fat}_\beta(\mathscr{F}') \le n$. Let $\mathbf{z}$ be an $\mathscr{X}$-valued tree of depth d that is β-shattered by $\mathscr{F}'$ with a witness tree $\mathbf{s}$. Observe that the functions f^ϵ that guarantee

$$\forall\, t \in [n],\ \epsilon_t (f^{\epsilon}(\mathbf{z}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) \ge \beta/2 \tag{32}$$

do not necessarily take on values close to the $\mathbf{s}_t(\epsilon) \pm \beta/2$ interval. We augment $\mathscr{F}'$ with 2^d functions g^ϵ that take on the same values as f^ϵ, except on points on the $\mathbf{z}$ tree where, for some choice κ, we have,

$$g^{\epsilon}(\mathbf{z}_t(\epsilon)) = \epsilon_t \beta/2 + \kappa\ .$$

Let $\mathscr{F}$ be the resulting class of functions, and $\mathscr{G} = \mathscr{F} \setminus \mathscr{F}'$. We now argue that $\mathrm{fat}_\beta(\mathscr{F})$ cannot be more than $2d+4$, as we have only added at most 2^d functions to $\mathscr{F}'$. Suppose for the sake of contradiction that there exists a tree $\mathbf{z}$ of depth at least $2d+5$ shattered by $\mathscr{F}$. There must exist 2^{2d+5} functions that shatter $\mathbf{z}$ and only at most 2^d of them can be from $\mathscr{G}$. Let us label the leaves of $\mathbf{z}$ with the functions that shatter the corresponding path from the root; these functions are clearly distinct. Order the leaves of the tree in any way, and observe that there must exist a pair of functions from $\mathscr{G}$ with indices differing by at least 2^{d+4}. It is easy to see that such two leaves can only have a common parent at $d+3$ levels from the leaves, and this yields a complete binary subtree of size $d+1$ that is shattered by functions in $\mathscr{F}'$, a contradiction.

We will now use the function class $\mathscr{F}$ to prove a lower bound. Recall that $\mathbf{z}$ is an $\mathscr{X}$-valued tree of depth fat_β that is β-shattered by $\mathscr{G} \subseteq \mathscr{F}$, with a witness tree having the constant κ at every node. We will now show a construction of particular trees of depth

$$n' = \left\lceil \frac{n}{\mathrm{fat}_\beta} \right\rceil \mathrm{fat}_\beta \tag{33}$$

using the tree $\mathbf{z}$. Define $k = \lceil \frac{n}{\mathrm{fat}_\beta} \rceil = \frac{n'}{\mathrm{fat}_\beta} \ge 1$ and consider the $\mathscr{X}$-valued tree $\mathbf{x}$ and the $\mathbb{R}$-valued tree $\boldsymbol{\mu}$ of depth n' constructed as follows. For any path $\epsilon \in \{\pm 1\}^{n'}$ and any $t \in [n']$, set

$$\mathbf{x}_t(\epsilon) = \mathbf{z}_{\lceil \frac{t}{k} \rceil}(\tilde{\epsilon})\,, \quad \boldsymbol{\mu}_t(\epsilon) = \kappa$$

where $\tilde{\epsilon} \in \{\pm 1\}^{\mathrm{fat}_\beta}$ is the sequence of signs specified as

$$\tilde{\epsilon} = \left(\mathrm{sign}\left(\sum_{j=1}^{k} \epsilon_j \right), \mathrm{sign}\left(\sum_{j=k+1}^{2k} \epsilon_j \right), \ldots, \mathrm{sign}\left(\sum_{j=k(\mathrm{fat}_\beta - 1)}^{k\,\mathrm{fat}_\beta} \epsilon_j \right) \right).$$

We now lower bound (14) by choosing the particular $\mathbf{x}, \boldsymbol{\mu}$ defined above:

$$\begin{aligned} V_{n'} &\geq R\,\mathbb{E} \sup_{f \in \mathscr{F}} \left[\sum_{t=1}^{n'} \epsilon_t (f(\mathbf{x}_t(\epsilon)) - \kappa) - \frac{1}{R} \boldsymbol{\Delta}_\kappa (f(\mathbf{x}_t(\epsilon)) - \kappa) \right] \\ &= R\,\mathbb{E} \sup_{f \in \mathscr{F}} \left[\sum_{t=1}^{n'} \epsilon_t (f(\mathbf{z}_{\lceil \frac{t}{k} \rceil}(\tilde{\epsilon})) - \kappa) - \frac{1}{R} \boldsymbol{\Delta}_\kappa (f(\mathbf{z}_{\lceil \frac{t}{k} \rceil}(\tilde{\epsilon})) - \kappa) \right] . \end{aligned}$$

Splitting the sum over t into fat_β blocks, the above expression is equal to

$$\begin{aligned} & R\,\mathbb{E} \sup_{f \in \mathscr{F}} \left[\sum_{i=1}^{\mathrm{fat}_\beta} \sum_{j=(i-1)k+1}^{i \cdot k} \epsilon_j (f(\mathbf{z}_i(\tilde{\epsilon})) - \kappa) - \frac{1}{R} \boldsymbol{\Delta}_\kappa (f(\mathbf{z}_i(\tilde{\epsilon})) - \kappa) \right] \\ &= R\,\mathbb{E} \sup_{f \in \mathscr{F}} \left[\sum_{i=1}^{\mathrm{fat}_\beta} (f(\mathbf{z}_i(\tilde{\epsilon})) - \kappa) \left(\sum_{j=(i-1)k+1}^{i \cdot k} \epsilon_j \right) - \frac{k}{R}\, \boldsymbol{\Delta}_\kappa (f(\mathbf{z}_i(\tilde{\epsilon})) - \kappa) \right] \\ &= R\,\mathbb{E} \sup_{f \in \mathscr{F}} \left[\sum_{i=1}^{\mathrm{fat}_\beta} \tilde{\epsilon}_i (f(\mathbf{z}_i(\tilde{\epsilon})) - \kappa) \left| \sum_{j=(i-1)k+1}^{i \cdot k} \epsilon_j \right| - \frac{k}{R}\, \boldsymbol{\Delta}_\kappa (f(\mathbf{z}_i(\tilde{\epsilon})) - \kappa) \right] \end{aligned}$$

where the last step follows by the definition of $\tilde{\epsilon}$. Recall that $\mathbf{z}$ is shattered by the subset $\mathscr{G}$ and that the functions in $\mathscr{G}$ stay close to the witness tree $\mathbf{s}$. We obtain a lower bound

$$\begin{aligned} R\,\mathbb{E} \sup_{g \in \mathscr{G}} \left[\sum_{i=1}^{\mathrm{fat}_\beta} \tilde{\epsilon}_i (g(\mathbf{z}_i(\tilde{\epsilon})) - \kappa) \left| \sum_{j=(i-1)k+1}^{i \cdot k} \epsilon_j \right| - \frac{k}{R}\, \boldsymbol{\Delta}_\kappa (g(\mathbf{z}_i(\tilde{\epsilon})) - \kappa) \right] &\geq R\,\mathbb{E} \sum_{i=1}^{\mathrm{fat}_\beta} \left(\frac{\beta}{2} \left| \sum_{j=(i-1)k+1}^{i \cdot k} \epsilon_j \right| - \frac{k}{R}\, \boldsymbol{\Delta}_\kappa \left(\frac{\beta}{2} \right) \right) \\ &\geq R\, \mathrm{fat}_\beta(\mathscr{F}) \left(\frac{\beta}{2} \sqrt{\frac{k}{2}} - \frac{k}{R}\, \boldsymbol{\Delta}_\kappa \left(\frac{\beta}{2} \right) \right) \end{aligned}$$

where we used Khinchine's inequality in the last step. By the definition of k,

$$\mathrm{fat}_\beta(\mathscr{F}) \frac{\beta}{2} \sqrt{\frac{k}{2}} = \mathrm{fat}_\beta(\mathscr{F}) \frac{\beta}{2} \sqrt{\frac{n'}{2\,\mathrm{fat}_\beta(\mathscr{F})}} = \frac{1}{2\sqrt{2}} \beta \sqrt{n' \mathrm{fat}_\beta(\mathscr{F})}$$

and $\mathrm{fat}_\beta(\mathscr{F}) k = n'$ and so we conclude that,

$$V_{n'} \geq \frac{R\,\beta}{2} \sqrt{\frac{n' \mathrm{fat}_\beta(\mathscr{F})}{2}} - n'\, \boldsymbol{\Delta}_\kappa \left(\frac{\beta}{2} \right)$$

Since we are free to choose κ and β,

$$V_{n'} \geq \sup_{\beta, \kappa} \left\{ \frac{R\,\beta}{2} \sqrt{\frac{n' \mathrm{fat}_\beta(\mathscr{F})}{2}} - n'\, \boldsymbol{\Delta}_\kappa \left(\frac{\beta}{2} \right) \right\} \tag{34}$$

Examining (23), we see that V_n is nondecreasing with n. To see this, let $n' > n$. For $t \in \{n+1, \dots, n'\}$, we may choose p_t in (23) as a delta distribution on $f^*(x_t)$, for any sequence of x_t, where f^* is an optimal function over steps $\{1, \dots, n\}$. Clearly, $V_{n'} \geq V_n$. In view of (33) and the above discussion, $V_{n'} \leq V_{2n-1}$, and thus

$$V_{2n} \geq V_{2n-1} \geq V_{n'}.$$

□

Proof of Theorem 10. We have $\Delta(t) = Kt^r$ for $r \geq 2$. It will suffice to take the conjugate of $t \mapsto Kt^{r/2}$ over all of $\mathbb{R}$. A straightforward calculation shows that

$$\Gamma^*(s) \leq \frac{K}{2}(r-2)\left(\frac{2s}{Kr}\right)^{\frac{r}{r-2}} \leq \frac{r-2}{2e}\frac{s^{\frac{r}{r-2}}}{K^{\frac{2}{r-2}}}$$

Combining Lemma 3 and Lemma 5, the minimax value V_n is upper bounded by

$$\inf_{\gamma>0}\left\{G\inf_{\rho\in(0,\gamma)}\left\{4\rho n + 12\sqrt{n}\int_\rho^\gamma\sqrt{\log\mathcal{N}_\infty(\delta,\mathcal{F},n)}\,d\delta\right\} + \inf_{\lambda>0}\left\{\frac{1}{\lambda}\log\mathcal{N}_\infty\left(\tfrac{\gamma}{2},\mathcal{F},n\right) + n\,\Gamma^*\left(2\lambda G^2\right)\right\}\right\} \tag{35}$$

Consider the case $p \in (0,2)$. By the assumption on the growth of the covering numbers, and taking $\rho = 1/n$,

$$\int_\rho^\gamma\sqrt{\log\mathcal{N}_\infty(\delta,\mathcal{F},n)}\,d\delta \leq \int_\rho^\gamma\delta^{-p/2}\sqrt{\log(n/\delta)}\,d\delta \leq \frac{c\sqrt{\log n}}{2-p}\gamma^{1-p/2}$$

Then (35) is upper bounded by

$$\inf_{\gamma>0}\left\{4G + G\sqrt{n}\frac{c\sqrt{\log n}}{2-p}\gamma^{1-p/2} + \inf_{\lambda>0}\left\{\frac{1}{\lambda}\gamma^{-p}\log(n/\gamma) + n\,\Gamma^*\left(2\lambda G^2\right)\right\}\right\}$$

We take $\gamma \geq n^{-c}$ and divide through by $\log n$:

$$\frac{V_n}{n\log n} \leq \frac{4G}{n} + \inf_{\gamma\geq n^{-c}}\left\{c_pGn^{-1/2}\gamma^{1-p/2} + \inf_{\lambda>0}\left\{\frac{1}{n\lambda}\gamma^{-p} + \Gamma^*\left(2\lambda G^2\right)\right\}\right\}$$

where c_p is a constant that depends on p. Balancing the terms in the inner infimum,

$$\inf_{\lambda>0}\left\{\frac{1}{n\lambda}\gamma^{-p} + \Gamma^*\left(2\lambda G^2\right)\right\} = \inf_{\lambda>0}\left\{\frac{1}{n\lambda}\gamma^{-p} + c_r\frac{(\lambda G^2)^{\frac{r}{r-2}}}{K^{\frac{2}{r-2}}}\right\} \leq c_r n^{-\frac{r}{2(r-1)}}\gamma^{-\frac{rp}{2(r-1)}}G^{\frac{r}{r-1}}K^{-\frac{1}{r-1}}$$

where c_r is a constant that depends on r and may change from one expression to next. The value of γ that balances

$$c_pGn^{-1/2}\gamma^{1-p/2} = c_r n^{-\frac{r}{2(r-1)}}\gamma^{-\frac{rp}{2(r-1)}}G^{\frac{r}{r-1}}$$

is

$$\gamma = c_{r,p}n^{-\frac{1}{2(r-1)+p}}G^{\frac{2}{2(r-1)+p}}K^{-\frac{2}{2(r-1)+p}}$$

and it gives

$$\begin{aligned}\frac{V_n}{n\log n} &\leq c_{r,p}\left(n^{-\frac{1}{2(r-1)+p}}G^{\frac{2}{2(r-1)+p}}K^{-\frac{2}{2(r-1)+p}}\right)^{1-p/2}Gn^{-1/2}\\ &= c_{r,p}\,n^{-\frac{r}{2(r-1)+p}}G^{\frac{2r}{2(r-1)+p}}K^{-\frac{2-p}{2(r-1)+p}}\end{aligned}$$

On the other hand, using [25, Theorem 8], we have

$$V_n \leq cG\mathfrak{R}_n(\mathcal{F})$$

In turn, sequential Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$ is upper bounded via (9) by either $\mathcal{O}(n^{1-1/p})$ or $\mathcal{O}(n^{1/2})$ for $p > 2$ and $p \in (0,2)$, respectively. More precisely, for the $p > 2$ regime, taking $\rho = n^{-1/p}$ we obtain

$$\frac{1}{n}\mathfrak{R}_n(\mathcal{F}) \leq 4n^{-1/p} + 12n^{-1/2}\int_{n^{-1/p}}^\infty\sqrt{\delta^{-p}\log(n/\delta)}\,d\delta \tag{36}$$

$$= cn^{-1/p} + c\sqrt{\log(n)/n}\left[\left(\frac{2}{2-p}\right)\delta^{(2-p)/2}\right]_{n^{-1/p}}^\infty \tag{37}$$

$$\leq c_pn^{-1/p}\sqrt{\log(n)}. \tag{38}$$

The same calculation for $p \in (0,2)$ gives $\frac{1}{n}\mathfrak{R}_n(\mathcal{F}) \leq c_{\mathcal{F}}n^{-1/2}\sqrt{\log(n)}$. □

Proof of Theorem 11. From Lemma 9,

$$V_n \geq \sup_{\beta \leq 1} \left\{ \frac{R\beta}{2} \sqrt{\frac{n \mathrm{fat}_\beta(\mathscr{F})}{2}} - n\, \boldsymbol{\Delta}_\kappa \left(\frac{\beta}{2}\right) \right\} \geq \sup_{\beta \leq 1} \left\{ \frac{R\sqrt{n}\beta^{1-p/2}}{2\sqrt{2}} - nK\beta^r \right\}$$

Using $\beta = \min\left\{1, \left(\frac{R^2}{c_{p,r}K^2 n}\right)^{\frac{1}{2r+p-2}}\right\}$, we find that

$$\frac{1}{n} V_n \geq c_{p,r} \min\left\{ \frac{R}{\sqrt{n}} \,,\;\; K^{-\frac{2-p}{2(r-1)+p}} R^{\frac{2r}{2(r-1)+p}} n^{-\frac{r}{2(r-1)+p}} \right\}$$

for some constant $c_{p,r}$.

□

Proof of Corollary 14. The second derivative of this loss with respect to the first argument is given by $q(q-1)|y - \widehat{y}|^{q-2}$, and it is lower bounded by $\frac{q(q-1)}{2}$ because $q \in (1,2)$ and $y, \widehat{y} \in [-1,1]$. This means that the loss is $q(q-1)/2$ strongly convex and so

$$\underline{\boldsymbol{\Delta}}(x) \geq \frac{q(q-1)}{2} x^2.$$

We now turn to upper bounding $\overline{\boldsymbol{\Delta}}$. Choose $S = \{0\}$ and take $y_1 = 1, y_2 = -1$. By symmetry of the loss function, the optimal $\widehat{y}^* = 0$, verifying property (12). Then

$$\overline{\boldsymbol{\Delta}}_0(x) = \sup_{x \in \mathscr{Y}} \max\left\{\Delta^1_{0,x}, \Delta^{-1}_{0,x}\right\}, \tag{39}$$

with domain $\widehat{\mathscr{Y}} - \{0\} = [-1,1]$. For any $y \in \mathscr{Y}$, the generalized binomial theorem gives an expansion of $\ell(\cdot, y)$ at the point $a \neq y$ as

$$\Delta^y_{a,b} = \ell(b,y) - [\ell(a,y) + \partial_a \ell(a,y) \cdot (b-a)] = \sum_{j=2}^{\infty} \frac{\prod_{k=0}^{j-1}(q-k)}{j!} (a-y)^{q-j} \cdot (b-a)^j$$

Then, taking $b = x$ and $a = 0$,

$$\overline{\boldsymbol{\Delta}}_0(x) \leq \sum_{j=2}^{\infty} \frac{\prod_{k=0}^{j-1}|q-k|}{j!} |x|^j = \left(\sum_{j=2}^{\infty} \frac{\prod_{k=2}^{j-1}(k-q)}{j!} |x|^{j-2} \right) q(q-1)x^2$$

Since $q > 1$ we can bound the above by

$$\left(\sum_{j=2}^{\infty} \frac{(j-2)!}{j!} |x|^{j-2} \right) q(q-1)x^2 \leq \left(\sum_{j=2}^{\infty} \frac{1}{j(j-1)} \right) q(q-1)x^2 \leq 2q(q-1)x^2.$$

The result follow from Theorems 10 and 11.

□

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.Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.

[3] P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

[4] K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, June 2001.

[5] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[6] N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. *Journal of Computer and System Sciences*, 59(3):392–411, 1999.

[7] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[8] 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.

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

[10] D. P. Foster. Prediction in the worst case. *Annals of Statistics*, 19(2):1084–1090, 1991.

[11] S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14:729–769, 2013.

[12] S. Gerchinovitz and J. Yu. Adaptive and optimal online linear regression on ℓ_1-balls. *Theoretical Computer Science*, 2013.

[13] E. Giné and J. Zinn. Some limit theorems for empirical processes. *Annals of Probability*, 12(4):929–989, 1984.

[14] W. Han, A. Rakhlin, and K. Sridharan. Competing with strategies. In *Conference on Learning Theory*, 2013.

[15] D. Haussler, J. Kivinen, and M. Warmuth. Sequential prediction of individual sequences under general loss functions. *Information Theory, IEEE Transactions on*, 44(5):1906–1925, 1998.

[16] E. Hazan and N. Megiddo. Online learning with prior knowledge. In *Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pages 499–513. 2007.

[17] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inf. Comput.*, 132(1):1–63, 1997.

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

[19] 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.

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

[21] A. Rakhlin and K. Sridharan. On semi-probabilistic universal prediction. In *Information Theory Workshop (ITW), 2013 IEEE*, pages 1–5. IEEE, 2013.

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

[23] A. Rakhlin and K. Sridharan. Online nonparametric regression. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, page 1232âĂŞ1264, 2014.

[24] 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.

[25] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research*, 2014. To appear.

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

[27] A. Rakhlin, K. Sridharan, and A. Tsybakov. Entropy, minimax regret and minimax risk. In submission, 2014.

[28] V. Vovk. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 51–60. ACM, 1995.

[29] V. Vovk. Competitive on-line linear regression. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 364–370, Cambridge, MA, USA, 1998. MIT Press.

[30] V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

[31] V. Vovk. Metric entropy in competitive on-line prediction. *CoRR*, abs/cs/0609045, 2006.

[32] V. Vovk. On-line regression competitive with reproducing kernel hilbert spaces. In *Theory and Applications of Models of Computation*, pages 452–463. Springer, 2006.

[33] V. Vovk. Competing with wild prediction rules. *Machine Learning*, 69(2):193–212, 12 2007.