2-12-2014

Online Nonparametric Regression

Alexander Rakhlin
University of Pennsylvania

Karthik Sridharan
University of Pennsylvania

Online Nonparametric Regression

Abstract
We establish optimal rates for online regression for arbitrary classes of regression functions in terms of the sequential entropy introduced in [14]. The optimal rates are shown to exhibit a phase transition analogous to the i.i.d./statistical learning case, studied in [16]. In the frequently encountered situation when sequential entropy and i.i.d. empirical entropy match, our results point to the interesting phenomenon that the rates for statistical learning with squared loss and online nonparametric regression are the same. In addition to a non-algorithmic study of minimax regret, we exhibit a generic forecaster that enjoys the established optimal rates. We also provide a recipe for designing online regression algorithms that can be computationally efficient. We illustrate the techniques by deriving existing and new forecasters for the case of finite experts and for online linear regression.

Disciplines
Physical Sciences and Mathematics

Online Nonparametric Regression

Alexander Rakhlin
University of Pennsylvania

Karthik Sridharan
University of Pennsylvania

February 12, 2014

Abstract

We establish optimal rates for online regression for arbitrary classes of regression functions in terms of the sequential entropy introduced in [14]. The optimal rates are shown to exhibit a phase transition analogous to the i.i.d./statistical learning case, studied in [16]. In the frequently encountered situation when sequential entropy and i.i.d. empirical entropy match, our results point to the interesting phenomenon that the rates for statistical learning with squared loss and online nonparametric regression are the same.

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

1 Introduction

Within the online regression framework, data $(x_1, y_1), \ldots, (x_n, y_n), \ldots$ arrive in a stream, and we are tasked with sequentially predicting each next response y_t given the current x_t and the data $\{(x_i, y_i)\}_{i=1}^{t-1}$ observed thus far. Let $\hat{y}_t$ denote our prediction, and let the quality of this forecast be evaluated via square loss $(\hat{y}_t - y_t)^2$. Within the field of time series analysis, it is assumed that data are generated according to some model. The parameters of the model can then be estimated from data, leveraging the laws of probability. Alternatively, in the *competitive approach*, studied within the field of online learning, the aim is to develop a prediction method that does not assume a generative process of the data [7]. The problem is then formulated as that of minimizing regret

$$\sum_{t=1}^{n} (\hat{y}_t - y_t)^2 - \inf_{f \in \mathscr{F}} \sum_{t=1}^{n} (f(x_t) - y_t)^2 \tag{1}$$

with respect to some benchmark class of functions $\mathscr{F}$. This class encodes our prior belief about the family of regression functions that we expect to perform well on the sequence. Notably, an upper bound on regret is required to hold for all sequences.

In the past twenty years, progress in online regression for arbitrary sequences, starting with the paper of Foster [8], has been almost exclusively on finite-dimensional *linear* regression (an incomplete list includes [19, 11, 20, 4, 2, 3, 9]). 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 Vovk [23, 21, 22]. 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

mentioned above. 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 cover in the supremum norm on the underlying space of covariates (see Remark 3). The same difficulty arises when one uses PAC-Bayesian bounds [1] 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 [16] 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.

Sequential analogues of covering numbers, combinatorial parameters, and the Rademacher complexity have been recently introduced in [15]. These complexity measures were shown to both upper and lower bound minimax regret of online learning with absolute loss for arbitrary classes of functions. These rates, however, are not correct for the square loss case. Consider, for instance, finite-dimensional regression, where the behavior of minimax regret is known to be logarithmic in n; the Rademacher rate, however, cannot yield rates faster than $\sqrt{n}$. A hint as to how to modify the analysis for "curved" losses appears in the paper of [6] 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.[1]

Interestingly, our results imply the same phase transition as the one exhibited in [15] 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. [5]). 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 [16] 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 [13] provides a toolkit for developing improper learning algorithms in the i.i.d. scenario.

After characterizing minimax rates for online regression, we turn to the question of developing algorithms. We first show that an algorithm based on the Rademacher relaxation is admissible (see [13]) and yields the rates derived in a non-constructive manner in the first part of the paper. This algorithm is not generally computationally feasible, but, in particular, does achieve optimal rates, improving on those exhibited by Vovk [21] for Besov spaces. We show that further relaxations in finite dimensional space lead to the famous Vovk-Azoury-Warmuth forecaster. For illustration purposes, we also derive a prediction method for finite class $\mathscr{F}$.

2 Background

Let $\mathscr{X}$ be some set of covariates, and let $\mathscr{F}$ be a class of functions $\mathscr{X} \to [-1,1] = \mathscr{Y}$. We study the online regression scenario where on round $t \in \{1,\dots,n\}$, $x_t \in \mathscr{X}$ is revealed to the learner who subsequently makes a prediction $\hat{y}_t \in \mathbb{R}$; Nature then reveals[2] $y_t \in [-B,B]$. Instead of (1), we consider a slightly modified notion of regret

$$(1-\alpha)\sum_{t=1}^{n}(\hat{y}_t - y_t)^2 - \inf_{f\in\mathscr{F}}\sum_{t=1}^{n}(f(x_t)-y_t)^2 \tag{2}$$

[1] While we develop our results for square loss, similar statements hold for much more general losses, as will be shown in the full version of this paper.

[2] The assumption of bounded responses can be removed by standard truncation arguments (see e.g. [10]).

for some $\alpha \in [0,1)$. It is well-known that an upper bound on such a regret notion leads to the so-called *optimistic rates* which scale favorably with the cumulative loss $L^* = \inf_{f\in\mathscr{F}} \sum_{t=1}^n (f(x_t)-y_t)^2$ [2, 18]. More precisely, suppose we show an upper bound of $U_1/\alpha + U_2$ on regret in (2). Then regret in (1) is upper bounded by

$$4\sqrt{L^* U_1} + 12U_1 + 4U_2 \tag{3}$$

by considering the case $L^* \geq 4U_1$ and its converse.

Unlike most previous approaches to the study of online regression, we do not start from an algorithm, but instead directly work with minimax regret. We will be able to extract a (not necessarily efficient) algorithm after getting a handle 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 inside repeated over $t = 1 \dots n$ rounds. With this notation, the minimax regret of the online regression problem described earlier can be written as

$$V_n^\alpha = \left\langle\!\!\left\langle \sup_{x_t} \inf_{\hat{y}_t} \sup_{y_t} \right\rangle\!\!\right\rangle_{t=1}^n \left\{ (1-\alpha) \sum_{t=1}^n (\hat{y}_t - y_t)^2 - \inf_{f\in\mathscr{F}} \sum_{t=1}^n (f(x_t)-y_t)^2 \right\} \tag{4}$$

where each x_t ranges over $\mathscr{X}$ and $\hat{y}_t, y_t$ range over $[-B,B]$. The usual minimax regret notion is simply given when $\alpha = 0$ as V_n^0.

As mentioned above, in the i.i.d. scenario it is possible to employ a notion of a cover based on a sample, thanks to the symmetrization technique. In the online prediction scenario, symmetrization is more subtle, and involves the notion of a binary tree, the smallest entity that captures the sequential nature of the problem. To this end, let us state a few definitions. A $\mathscr{Z}$-valued tree $\mathbf{z}$ of depth n is a complete rooted binary tree with nodes labeled by elements of $\mathscr{Z}$. 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 $\epsilon = (\epsilon_1, \dots, \epsilon_n) \in \{\pm 1\}^n$, $\mathbf{z}_t(\epsilon) = \mathbf{z}_t(\epsilon_1, \dots, \epsilon_{t-1}) \in \mathscr{Z}$ is the label of the node on the t-th level of the tree obtained by following the path ϵ. For a function $g : \mathscr{Z} \to \mathbb{R}$, $g(\mathbf{z})$ 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 $\mathbf{z}$).

Next, let us define sequential covering numbers – one of the key complexity measures of $\mathscr{F}$.

Definition 1 ([15])**.** 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 $\mathscr{F} \subseteq \mathbb{R}^{\mathscr{X}}$ on a given $\mathscr{X}$-valued tree $\mathbf{x}$ of depth n if

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

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

We will refer to $\sup_{\mathbf{x}} \log \mathscr{N}_q(\beta, \mathscr{F}, \mathbf{x})$ as *sequential entropy* of $\mathscr{F}$. In particular, we will study the behavior of $V_n^\alpha(\mathscr{F})$ when sequential entropy grows polynomially[3] as the scale β decreases:

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

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

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

(e.g. linear regression in a bounded set in $\mathbb{R}^d$). We remark that the ℓ_∞ cover is necessarily n-dependent, so the form we assume there is

$$\mathscr{N}_\infty(\beta, \mathscr{F}, n) = (n/\beta)^{-d} . \tag{7}$$

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

3 Main Results

We now state the main results of this paper. They follow from the more general technical statements of Lemmas 4, 5, 6 and 7. We normalize V_n^α by n in order to make the rates comparable to those in statistical learning. Further, throughout the paper C, c refer to constants that may depend on B, p. Their values can be found in the proofs.

Theorem 1. *For a class $\mathscr{F}$ with sequential entropy growth $\log \mathscr{N}_2(\beta, \mathscr{F}, n) \le \beta^{-p}$,*

- *For $p > 2$, the minimax regret[4] is bounded as $\frac{1}{n} V_n^0 \le C n^{-1/p}$*
- *For $p \in (0,2)$, the minimax regret is bounded as $\frac{1}{n} V_n^0 \le C n^{-2/(2+p)}$*
- *For the parametric case (6), $\frac{1}{n} V_n^0 \le C d n^{-1} \log(n)$*
- *For finite set $\mathscr{F}$, $\frac{1}{n} V_n^0 \le C n^{-1} \log|\mathscr{F}|$*

Theorem 2. *The upper bounds of Theorem 1 are tight[5]:*

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

For the following theorem, we assume that L^* is known a priori. Adaptivity to L^* can be obtained through a doubling-type argument [17].

Theorem 3. *Additionally, the following optimistic rates hold for regret (1):*

- *For $p > 2$, regret is upper bounded by $C\sqrt{L^* n^{1-1/(p-1)} \log(n)} + C n^{1-1/(p-1)} \log(n)$*
- *For $p \in (0,2)$, regret is upper bounded by $C\sqrt{L^* \log(n)} + C \log(n)$. The bound gains an extra $\log(n)$ factor for $p = 2$*
- *For the parametric case (7), regret is upper bounded by $C\sqrt{L^* d \log(n)} + C d \log(n)$*

where $L^* = \inf_{f \in \mathscr{F}} \sum_{t=1}^n (f(x_t) - y_t)^2$.

Remark 1. *The optimistic rate for $p > 2$ appears to be slower than the hypothesized $\sqrt{L^* n^{1-2/p}} + n^{1-2/p}$ rate, and we leave the question of obtaining this rate as future work.*

Remark 2. *If we assume that y_t's are drawn from distributions with bounded mean and subgaussian tails, the same upper bounds can be shown with an extra $\log(n)$ factor.*

Next, we prove the three theorems stated above. The proofs are of the "plug-and-play style": the overarching idea is that the optimal rates can be derived simply by assuming an appropriate control of sequential entropy, be it a parametric or a nonparametric class.

Proof of Theorem 1. We appeal to Eq. (13) in Lemma 4 below. Fix $\mathbf{x}, \boldsymbol{\mu}$ and let $\mathbf{z}$ denote the $\mathscr{X} \times \mathbb{R}$-valued tree $(\mathbf{x}, \boldsymbol{\mu})$. Define the class $\mathscr{G} = \{g_f : g_f(\mathbf{z}) = f(\mathbf{x}) - \boldsymbol{\mu}, f \in \mathscr{F}\}$. Observe that the values of g_f outside of range of $\mathbf{z}$ are

[4] For $p = 2$, $\frac{1}{n} V_n^0 \le C \log(n) n^{-1/2}$.

[5] The $\tilde{\Omega}(\cdot)$ notation suppresses logarithmic factors

immaterial. Also note that the covering number of $\mathscr{G}$ on $\mathbf{z}$ coincides with the covering number of $\mathscr{F}$ on $\mathbf{x}$. Now, Lemma 5 applied to this class $\mathscr{G}$, together with $\boldsymbol{\eta} \equiv B$, yields

$$V_n^0 \le 32B^2 \log \mathscr{N}_2(\gamma, \mathscr{F}, n) + B \inf_{\rho \in (0,\gamma)} \left\{ 4\rho n + 12\sqrt{n} \int_\rho^\gamma \sqrt{\log \mathscr{N}_2(\delta, \mathscr{F}, n)}\, d\delta \right\} \tag{8}$$

We now evaluate the above upper bound for the β^{-p} growth of sequential entropy at scale β. In particular, for the case $p > 2$, we may choose $\gamma = 1$ (maximum of the function) and $\rho = n^{-1/p}$. Then $\mathscr{N}_2(B, \mathscr{F}, n) = 1$ and the first term disappears. We are left with

$$B^{-1} V_n^0 \le 4n^{1-p} + 12\sqrt{n} \left[\left(\frac{2}{2-p} \right) \delta^{(2-p)/2} \right]_{n^{-1/p}}^{B} \le 4n^{1-\frac{1}{p}} + \frac{24}{p-2} n^{-\frac{2-p}{2p} + \frac{1}{2}} = \left(4 + \frac{24}{p-2} \right) n^{1-1/p}$$

For the case $p \in (0,2)$, Eq. (8) gives an upper bound

$$32B^2 \gamma^{-p} + B \inf_{\rho \in (0,\gamma)} \left\{ 4\rho n + 12\sqrt{n} \int_\rho^\gamma \delta^{-p/2} d\delta \right\} \tag{9}$$

We choose $\gamma = n^{-1/(p+2)}$ and $\rho = n^{-1}$:

$$32B^2 n^{\frac{p}{p+2}} + 4B + 12\sqrt{n} \left[\left(\frac{2}{2-p} \right) \delta^{\frac{2-p}{2}} \right]_{n^{-1}}^{n^{-\frac{1}{p+2}}} \le 4B + \left(32B^2 + 12B \left(\frac{2}{2-p} \right) \right) n^{\frac{p}{p+2}}$$

For the case $p = 2$, we gain an extra factor of $\log(n)$ since the integral of δ^{-1} is the logarithm. For the parametric case (6), we choose $\gamma = n^{-1/2}$ and $\rho = n^{-1}$. Then Eq. (8) yields (for $n > 8$),

$$V_n^0 \le 16B^2 d \log n + 4B + 12\sqrt{n} \int_{n^{-1}}^{n^{-1/2}} \sqrt{d \log(1/\delta)}\, d\delta \le 16B^2 d \log n + 4B + 12\sqrt{d \log(n)}\,.$$

In the finite case, $\log \mathscr{N}_2(\gamma, \mathscr{F}, n) \le \log |\mathscr{F}|$ for any γ. We then have take $\gamma = 0$ (one can see that this value is allowed for the particular case of a finite class; or, use a small enough value). Then,

$$V_n^0 \le 32B^2 \log |\mathscr{F}|\,.$$

Normalizing by n yields the desired rates in the statement of the theorem. □

Proof of Theorem 2. The first two lower bounds are proved in Lemma 9 and 10. The lower bound for the parametric case follows from the i.i.d. lower bound in [16]. □

Proof of Theorem 3. For optimistic rates, we start with the upper bound in (12) and define $\mathscr{G}$ as above. We then appeal to Lemma 6 and obtain

$$V_n^\alpha \le \alpha^{-1} 16 \log \mathscr{N}_\infty(\gamma, \mathscr{F}, \mathbf{z}) + \alpha^{-1} \inf_{\rho \in (0,\gamma)} \left\{ 4\rho n + 16 \log(\gamma/\rho) \int_\rho^\gamma \delta \log \mathscr{N}_\infty(\delta, \mathscr{F}, \mathbf{z})\, d\delta \right\}. \tag{10}$$

For $\log \mathscr{N}_\infty(\beta, \mathscr{F}, n) \le \beta^{-p}$ decay of entropy for $p < 2$, we take $\rho = (nB)^{-1}$, $\gamma = 1$. The first term in (10) can be taken to be zero, as we may take one function at scale $\gamma = 1$. The infimum in (10) evaluates to

$$4 + 16 \log(nB) \int_{1/(nB)}^1 \delta^{1-p} d\delta \le 4 + 16 \log(nB) \left[\frac{1}{2-p} \delta^{2-p} \right]_{1/(nB)}^1 \le 4 + 16 \log(nB) \frac{1}{2-p}\,.$$

For $p = 2$, we gain an extra $\log(n)$ factor: $4 + 16(\log(nB))^2$.

For $p > 2$, we take $\rho = n^{-\frac{1}{p-1}}$ and $\gamma = 1$. Then infimum in (10) evaluates to

$$4n \cdot n^{-\frac{1}{p-1}} + 16 p^{-1} \log(n) \left[\frac{1}{2-p} \delta^{2-p} \right]_{n^{-\frac{1}{p-1}}}^1 \le 4n^{\frac{p-2}{p-1}} + 16 p^{-1} \log(n) \frac{1}{2-p} n^{\frac{p-2}{p-1}}\,.$$

For the parametric case (7), we take $\gamma = 1$ and $\rho = (nB)^{-1}$. Then (10) is upper bounded by

$$4 + 16 \log(nB) \int_{1/(nB)}^1 d\delta \log(1/\delta) d\delta \le 4 + 4d \log(nB)\,.$$

The final optimistic rates are obtained by following the bound in (3). □

3.1 Offset Rademacher Complexity and the Chaining Technique

Let us recall the definition of sequential Rademacher complexity of a class $\mathscr{F}$

$$\sup_{\mathbf{x}} \mathbb{E} \sup_{f \in \mathscr{F}} \left[\sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right] \tag{11}$$

introduced in [14], where the expectation is over a sequence of independent Rademacher random variables $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ and the supremum is over all $\mathscr{X}$-valued trees of depth n. While this complexity both upper- and lower-bounds minimax regret for absolute loss, it fails to capture the possibly faster rates one can obtain for regression. We show below that modified, or *offset*, versions of this complexity do in fact give optimal rates. These complexities have an extra quadratic term being subtracted off. Intuitively, this variance term "extinguishes" the $\sqrt{n}$-type fluctuations above a certain scale. Below this scale, complexity is given by the Dudley-type integral. The optimal balance of the scale gives the correct rates. As can be seen from the proof of Theorem 1, the critical scale γ is trivial (zero) for a finite case, then $n^{-1/2}$ for a parametric class, $n^{-1/(p+2)}$ for $p \in (0,2]$, and then becomes irrelevant (e.g. constant) at $p > 2$. Indeed, for $p > 2$, the rate is given purely by sequential Rademacher complexity, as curvature of the loss does not help. In particular, can achieve these rates for $p > 2$ by simply linearizing the square loss. The same phenomenon occurs in statistical learning with i.i.d. data [16].

We remark that [12] studies bounds for estimation with squared loss for the empirical risk minimization procedure and observes that it is enough to only consider one-sided estimates rather than concentration statements. The offset sequential Rademacher complexities are of this one-sided nature.

In Lemma 4 below, we provide a bound on minimax regret via offset sequential Rademacher complexities.

Lemma 4. *The minimax value V_n^α of online regression with responses y_t in a bounded interval $[-B,B]$ is upper bounded by*

$$V_n^\alpha \le \sup_{\mathbf{x},\boldsymbol{\mu},\boldsymbol{\eta}} \mathbb{E}_\epsilon \sup_{f \in \mathscr{F}} \left[\sum_{t=1}^{n} 4\epsilon_t \boldsymbol{\eta}_t(\epsilon)(f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon))^2 - \alpha \boldsymbol{\eta}_t(\epsilon)^2 \right] \tag{12}$$

and

$$V_n^0 \le \sup_{\mathbf{x},\boldsymbol{\mu}} \mathbb{E}_\epsilon \sup_{f \in \mathscr{F}} \left[\sum_{t=1}^{n} 4B\epsilon_t (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon))^2 \right] \tag{13}$$

where $\mathbf{x}$ ranges over all $\mathscr{X}$-valued trees, $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ over all $[-B,B]$-valued trees of depth n. Furthermore,

$$V_n^0 \ge \sup_{\mathbf{x},\boldsymbol{\mu}} \mathbb{E} \sup_{f \in \mathscr{F}} \left[\sum_{t=1}^{n} B\epsilon_t (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon))^2 \right] \tag{14}$$

where $\boldsymbol{\mu}$ ranges over $[-B/2, B/2]$-valued trees.

We now show that offset Rademacher complexities can be upper bounded by sequential entropies via the chaining technique. Lemma 5 below is an analogue of the Dudley-type integral bound

$$\sup_{\mathbf{x}} \mathbb{E} \sup_{g \in \mathscr{G}} \left[\sum_{t=1}^{n} \epsilon_t g(\mathbf{x}_t(\epsilon)) \right] \le \inf_{\rho \in (0,1]} \left\{ 4\rho n + 12\sqrt{n} \int_\rho^1 \sqrt{\log \mathscr{N}_2(\delta, \mathscr{G}, \mathbf{z})}\, d\delta \right\} \tag{15}$$

for sequential Rademacher proved in [15]. Crucially, the upper bound of Lemma 5 allows us to choose a critical scale γ.

Lemma 5. *Let $\boldsymbol{\eta}$ be a $[-B,B]$-valued tree of depth n. For any $\mathscr{Z}$-valued tree $\mathbf{z}$ and a class $\mathscr{G}$ of functions $\mathscr{Z} \to [-A,A]$ and any $\gamma \in (0,A]$,*

$$\mathbb{E} \sup_{g \in \mathscr{G}} \left[\sum_{t=1}^{n} 4\epsilon_t \boldsymbol{\eta}_t(\epsilon) g(\mathbf{z}_t(\epsilon)) - g(\mathbf{z}_t(\epsilon))^2 \right] \le 32B^2 \log \mathscr{N}_2(\gamma, \mathscr{G}, \mathbf{z}) + B \inf_{\rho \in (0,\gamma)} \left\{ 4\rho n + 12\sqrt{n} \int_\rho^\gamma \sqrt{\log \mathscr{N}_2(\delta, \mathscr{G}, \mathbf{z})}\, d\delta \right\}$$

For optimistic rates, we can take advantage of an additional offset. This offset arises from the quadratic term due to the α multiple of the loss of the algorithm.

Lemma 6. *Let $\boldsymbol{\eta}$ be a $[-B,B]$-valued tree of depth n. For any $\mathcal{Z}$-valued tree $\mathbf{z}$ and a class $\mathcal{G}$ of functions $\mathcal{Z}\to[-A,A]$, for any $\gamma\in(0,A]$,*

$$\mathbb{E}\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^{n} 4\epsilon_t\boldsymbol{\eta}_t(\epsilon)g(\mathbf{z}_t(\epsilon)) - g(\mathbf{z}_t(\epsilon))^2 - \alpha\boldsymbol{\eta}_t(\epsilon)^2\right] \le \alpha^{-1}16A^2\log\mathcal{N}_\infty(\gamma,\mathcal{G},\mathbf{z}) \tag{16}$$
$$+\alpha^{-1}\inf_{\rho\in(0,\gamma)}\left\{4\rho n + 16\log(\gamma/\rho)\int_\rho^\gamma \delta\log\mathcal{N}_\infty(\delta,\mathcal{G},\mathbf{z})d\delta\right\}$$

The chaining arguments of Lemmas 5 and 6 are based on the following key finite-class lemma:

Lemma 7. *Let $\boldsymbol{\eta}$ be a $[-B,B]$-valued tree of depth n. For a finite set W of $[-A,A]$-valued trees of depth n, it holds that*

$$\mathbb{E}\max_{\mathbf{w}\in W}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\mathbf{w}_t(\epsilon) - C\mathbf{w}_t(\epsilon)^2 - \alpha\boldsymbol{\eta}_t(\epsilon)^2\right] \le \min\left\{B^2(2C)^{-1}, A^2(2\alpha)^{-1}\right\}\log|W| \tag{17}$$

for any $C\ge 0$, $\alpha\ge 0$. It also holds that

$$\mathbb{E}\max_{\mathbf{w}\in W}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\mathbf{w}_t(\epsilon)\right] \le B\sqrt{2\log|W|\cdot\max_{\mathbf{w}\in W,\epsilon_{1:n}}\sum_{t=1}^{n}\mathbf{w}_n(\epsilon)^2}\,. \tag{18}$$

Remark 3. *Let us compare the upper bound of Lemma 5 to the bound we may obtain via a metric entropy approach, as in the work of Vovk [21]. 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)*

$$n\epsilon + \mathcal{H}(\epsilon,\mathcal{F}) \tag{19}$$

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. While the balance (19) can yield correct rates for small classes, it 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. Let us also mention that sequential covering number of $\mathcal{F}$ is an "empirical" quantity and is finite even if we cannot upper bound metric entropy.

4 Further Examples

For the sake of illustration we show bounds on minimax rates for a couple of examples.

Example 1 (Sparse linear predictors)**.** *Let $\mathcal{G}=\{g_1,\dots,g_M\}$ be a set of M functions such that each $g_i:\mathcal{X}\mapsto[-1,1]$. Define $\mathcal{F}$ to be the convex combination of at most s out of these M functions. That is*

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

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 further the ℓ_∞ 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, the upper bound is

$$\tfrac{1}{n} V_n^0 \le O\left(\frac{s \log(M/s)}{n}\right)$$

Example 2 (Besov Spaces)**.** *Let $\mathscr{X}$ be a compact subset of $\mathbb{R}^d$. Let $\mathscr{F}$ be a ball in Besov space $B^s_{p,q}(\mathscr{X})$. When $s > d/p$, pointwise metric entropy bounds at scale β scale as $\Omega(\beta^{-d/s})$ [21, p. 20]. On the other hand, when $s \in (d/p, \infty)$, one can show that the space is a Banach space that is p-uniformly convex. From [15], it can be shown that sequential Rademacher can be upper bounded by $O(n^{1-1/p})$, yielding an bound on sequential entropy at scale β as $O(\beta^{-p})$. 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 $O\left(n^{\frac{2s}{2s+d}}\right)$.*
- *If $s < d/2$, the minimax rate depends on the interaction of p and d, s:*
  - *if $p > 1 + \frac{d}{2s}$, the minimax rate is $O\left(n^{\frac{2s}{2s+d}}\right)$, as above.*
  - *otherwise, the minimax rate is $O\left(n^{1-\frac{1}{p}}\right)$*

5 Lower Bounds

The lower bounds will involve a notion of a "dimension" of $\mathscr{F}$ called the sequential fat-shattering dimension. Let us introduce this notion.

Definition 2. An $\mathscr{X}$-valued tree of depth d is said to be β-shattered by $\mathscr{F}$ if there exists an $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d such that

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

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

The sequential fat-shattering dimension is related to sequential covering numbers as follows:

Theorem 8 ([15])**.** *Let $\mathscr{F}$ be a class of functions $\mathscr{X} \to [-1,1]$. For any $\beta > 0$,*

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

Therefore, if $\log \mathscr{N}_2(\beta, \mathscr{F}, n) \ge (c/\beta)^p$, then

$$\mathrm{fat}_\beta(\mathscr{F}) \ge (c/\beta)^p / (\log(2en/\beta)) .$$

The lower bounds will now be obtained assuming $\mathrm{fat}_\beta(\mathscr{F}) \ge C/\beta^p$ behavior of the fat-shattering dimension, and the resulting statement of Theorem 2 in terms of the sequential entropy growth will involve extra logarithmic factors, hidden in the $\tilde{\Omega}(\cdot)$ notation.

Lemma 9. *Consider the problem of online regression with responses bounded by $B = 4$. For any class $\mathscr{F}$ of functions $\mathscr{X} \to [-1,1]$ and any $\beta > 0$ and $n = \mathrm{fat}_\beta(\mathscr{F})$,*

$$\frac{1}{n} V_n^0 \ge \beta$$

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

$$\frac{1}{n} V_n^0 \ge C n^{-1/p} .$$

Lemma 10. *For any class $\mathscr{F}'$ and $\beta > 0$, there exists a modified class $\mathscr{F}$ such that* $\mathrm{fat}_\beta(\mathscr{F}) \le 2\mathrm{fat}_\beta(\mathscr{F}') + 4$ *and for* $n > \mathrm{fat}_\beta(\mathscr{F})$,

$$\frac{1}{n} V_n^0 \ge C\left(2\sqrt{2}\beta\sqrt{\frac{\mathrm{fat}_\beta(\mathscr{F})}{n}} - \beta^2\right).$$

In particular, when $p \in (0,2]$ and $\mathrm{fat}_\beta(\mathscr{F}) = C/\beta^p$,

$$\frac{1}{n} V_n^0 \ge C n^{-\frac{2}{p+2}}.$$

6 Relaxations and Algorithms

To design generic forecasters for the problem of online non-parametric regression we follow the recipe provided in [13]. 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 initial and admissibility conditions then one can build estimators based on such relaxations. Specifically, we look for relaxations that satisfy the following initial condition

$$\mathbf{Rel}_n\left(x_{1:n}, y_{1:n}\right) \ge -\inf_{f\in\mathscr{F}} \sum_{t=1}^n (f(x_t) - y_t)^2$$

and the recursive admissibility condition that for any $t \in [n]$ and any $x_t \in \mathscr{X}$

$$\inf_{\hat{y}_t\in[-B,B]} \sup_{y_t\in[-B,B]} \left\{(\hat{y}_t - y_t)^2 + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)\right\} \le \mathbf{Rel}_n\left(x_{1:t-1}, y_{1:t-1}\right) \tag{20}$$

If a relaxation $\mathbf{Rel}_n$ satisfies these two conditions then one can define an algorithm via

$$\hat{y}_t = \operatorname*{argmin}_{\hat{y}\in[-B,B]} \sup_{y_t\in[-B,B]} \left\{(\hat{y} - y_t)^2 + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)\right\}$$

and for this forecast the associated bound on regret is automatically bounded as (see [13] for details) :

$$\mathbf{Reg}_n \le \mathbf{Rel}_n(\cdot)$$

Now further note that if $(\hat{y} - y_t)^2 + \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, y_t)\right)$ is a convex function of y_t then the prediction takes a very simple form, as the supremum over y_t is attained either at B or $-B$. The prediction can be written as

$$\hat{y}_t = \operatorname*{argmin}_{\hat{y}\in[-B,B]} \max\left\{(\hat{y} - B)^2 + \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, B)\right), (\hat{y} + B)^2 + \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, -B)\right)\right\}$$

Observe that the first term decreases as $\hat{y}$ increases to B and likewise the second term monotonically decreases as $\hat{y}$ decreases to $-B$. Hence the solution to the above is given when both terms are equal (if this doesn't happen within the range $[-B, B]$ then we clip). In other words,

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

Hence, for any admissible relaxation such that $(\hat{y} - y_t)^2 + \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, y_t)\right)$ is a convex function of y_t, the above prediction based on the relaxation enjoys the bound on regret $\frac{1}{n}\mathbf{Rel}_n$.

We now claim that the following conditional version of Equation (13) gives an admissible relaxation and leads to a method that enjoys the regret bounds shown in the first part of this paper.

Lemma 11. *The following relaxation is admissible :*

$$\mathfrak{R}_n(x_{1:t}, y_{1:t}) = \sup_{\mathbf{x},\boldsymbol{\mu}} \mathbb{E}_\epsilon \sup_{f\in\mathscr{F}} \left[\sum_{j=t+1}^n 4B\epsilon_j(f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon)) - (f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon))^2 - \sum_{j=1}^t (f(x_j) - y_j)^2\right]$$

The forecast corresponding to this relaxation is given by

$$\hat{y}_t = \frac{\mathfrak{R}_n(x_{1:t},(y_{1:t-1},B)) - \mathfrak{R}_n(x_{1:t},(y_{1:t-1},-B))}{4B}$$

The above algorithm enjoys the regret bound of an offset Rademacher complexity:

$$\mathbf{Reg}_n \leq \sup_{\mathbf{x},\boldsymbol{\mu}} \mathbb{E}_\epsilon \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 4B\epsilon_t (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon))^2 \right]$$

Notice that since the regret bound for the above prediction based on the sequential Rademacher relaxation is exactly the one given in Equation (13), the upper bounds provided for V_n^0 in Theorem 1 also hold for the above algorithm.

6.1 Recipe for designing online regression algorithms

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

1. Find relaxation $\mathbf{Rel}_n$ such that
$$\mathfrak{R}_n\left(x_{1:t},y_{1:t}\right) \leq \mathbf{Rel}_n\left(x_{1:t},y_{1:t}\right)$$
and s.t. $(\hat{y}-y_t)^2 + \mathfrak{R}_n\left(x_{1:t},(y_{1:t-1},y_t)\right)$ is a convex function of y_t
2. Check the condition
$$\sup_{x_t\in\mathscr{X},p_t\in\Delta([-B,B])} \left\{ \mathbb{E}_{y_t\sim p_t}\left[\left(\mathbb{E}_{y_t\sim p_t}\left[y_t\right] - y_t \right)^2 \right] + \mathbb{E}_{y_t\sim p_t}\left[\mathbf{Rel}_n\left(x_{1:t},y_{1:t}\right)\right] \right\} \leq \mathbf{Rel}_n\left(x_{1:t-1},y_{1:t-1}\right)$$
3. Given x_t on round t, the prediction $\hat{y}_t$ is given by
$$\hat{y}_t = \mathrm{Clip}\left(\frac{\mathbf{Rel}_n\left(x_{1:t},(y_{1:t-1},B)\right) - \mathbf{Rel}_n\left(x_{1:t},(y_{1:t-1},-B)\right)}{4B} \right)$$

Proposition 12. *Any algorithm derived from the above schema using relaxation* $\mathbf{Rel}_n$ *enjoys a bound*

$$\mathbf{Reg}_n \leq \tfrac{1}{n}\mathbf{Rel}_n(\cdot)$$

on regret.

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

Corollary 13. *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 following algorithm

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

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

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 :

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

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. Hence we can readily apply the schema for designing an algorithm for this problem.

Corollary 14. *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 [19, 3]:

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

and enjoys the following upper bound on regret:

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

References

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

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

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

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

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

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

[12] S. Mendelson. Learning without Concentration. *ArXiv e-prints*, January 2014.

[13] 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.

[14] 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.

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

[16] A. Rakhlin, K. Sridharan, and A. Tsybakov. Entropy, minimax regret and minimax risk. In submission, 2013.

[17] S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, Hebrew University, 2007.

[18] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010.

[19] 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.

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

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

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

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

A Proofs

Proof of Lemma 4. Let us now study the value (4). We will do so "from inside out" by considering the last step $t = n$, then working our way back to $t = 1$. Given a value x_n, by the minimax theorem,

$$\inf_{q_n}\sup_{p_n}\mathbb{E}_{\hat{y}_n\sim q_n, y_n\sim p_n}\left\{(1-\alpha)(\hat{y}_n-y_n)^2+\sup_{f\in\mathscr{F}}\sum_{t=1}^{n}-(f(x_t)-y_t)^2\right\} \tag{21}$$

$$=\sup_{p_n}\left\{(1-\alpha)\inf_{\hat{y}_n}\mathbb{E}_{y_n}(\hat{y}_n-y_n)^2+\mathbb{E}_{y_n}\sup_{f\in\mathscr{F}}\sum_{t=1}^{n}-(f(x_t)-y_t)^2\right\}$$

$$=\sup_{p_n}\mathbb{E}_{y_n}\left\{(1-\alpha)(\mu_n-y_n)^2+\sup_{f\in\mathscr{F}}\sum_{t=1}^{n}-(f(x_t)-y_t)^2\right\} \tag{22}$$

where $\mu_n = \mathbb{E}[y_n]$ under the distribution p_n with support on $[-B,B]$. Observe that

$$(\mu_n-y_n)^2-(f(x_n)-y_n)^2 = 2(y_n-\mu_n)(f(x_n)-\mu_n)-(f(x_n)-\mu_n)^2 \tag{23}$$

and hence the expression in (21) can be written as

$$\sup_{p_n}\mathbb{E}_{y_n}\sup_{f\in\mathscr{F}}\left[\sum_{t=1}^{n-1}-(f(x_t)-y_t)^2+\left\{2(y_n-\mu_n)(f(x_n)-\mu_n)-(f(x_n)-\mu_n)^2-\alpha(\mu_n-y_n)^2\right\}\right]$$

Continuing in this fashion back to $t = 1$, the minimax value is equal to

$$V_n^\alpha = \left\langle\!\!\left\langle \sup_{x_t}\sup_{p_t}\mathbb{E}_{y_t}\right\rangle\!\!\right\rangle_{t=1}^{n}\left\{\sup_{f\in\mathscr{F}}\left[\sum_{t=1}^{n}2(y_t-\mu_t)(f(x_t)-\mu_t)-(f(x_t)-\mu_t)^2-\alpha(\mu_t-y_t)^2\right]\right\}. \tag{24}$$

The supremum over p_t can now be upper bounded by the supremum over the mean $\mu_t \in [-B,B]$ and a zero-mean distribution p'_t with support on $[-B,B]$. Denoting by η_t a random variable with this distribution p'_t, the variable $\mu_t+\eta_t$ is then in $[-2B,2B]$. We upper bound (24) by

$$V_n^\alpha \le \left\langle\!\!\left\langle \sup_{x_t}\sup_{p'_t,\mu_t}\mathbb{E}_{\eta_t}\right\rangle\!\!\right\rangle_{t=1}^{n}\left\{\sup_{f\in\mathscr{F}}\left[\sum_{t=1}^{n}2\eta_t(f(x_t)-\mu_t)-(f(x_t)-\mu_t)^2-\alpha\eta_t^2\right]\right\}. \tag{25}$$

Since the $-\alpha\eta^2$ term does not depend on f, we use linearity of expectation to write

$$V_n^\alpha = \left\langle\!\!\left\langle \sup_{x_t}\sup_{p'_t,\mu_t}\mathbb{E}_{\eta_t}\right\rangle\!\!\right\rangle_{t=1}^{n}\left\{\sup_{f\in\mathscr{F}}\left[\sum_{t=1}^{n}2\eta_t(f(x_t)-\mu_t)-(f(x_t)-\mu_t)^2-\mathscr{D}(p'_1,\dots,p'_n)\right]\right\} \tag{26}$$

where

$$\mathscr{D}(p'_1,\dots,p'_n) = \frac{1}{n}\sum_{t=1}^{n}\alpha\mathbb{E}\eta_t^2.$$

We now symmetrize the linear term. Let (η'_t) be a sequence tangent to (η_t) (that is, η_t and η'_t are i.i.d. conditionally on $\eta_{1:t-1}$). We write $\mu_t = \mathbb{E}[\eta'_t]$ and use convexity of the supremum to arrive at an upper bound

$$V_n^\alpha \le \left\langle\!\!\left\langle \sup_{x_t}\sup_{p'_t,\mu_t}\mathbb{E}_{\eta_t}\right\rangle\!\!\right\rangle_{t=1}^{n}\left\{\sup_{f\in\mathscr{F}}\left[\sum_{t=1}^{n}2(\eta_t-\eta'_t)(f(x_t)-\mu_t)-(f(x_t)-\mu_t)^2-\mathscr{D}(p'_1,\dots,p'_n)\right]\right\} \tag{27}$$

$$= \left\langle\!\!\left\langle \sup_{x_t}\sup_{p'_t,\mu_t}\mathbb{E}_{\eta_t,\eta'_t}\mathbb{E}_{\epsilon_t}\right\rangle\!\!\right\rangle_{t=1}^{n}\left\{\sup_{f\in\mathscr{F}}\left[\sum_{t=1}^{n}2\epsilon_t(\eta_t-\eta'_t)(f(x_t)-\mu_t)-(f(x_t)-\mu_t)^2-\mathscr{D}(p'_1,\dots,p'_n)\right]\right\} \tag{28}$$

where in the second equality holds because η'_t and η_t are i.i.d. from p'_t, conditionally on the past observations. We now split the above supremum over f into two parts, thus passing to the upper bound

$$\begin{aligned}
&\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t,\mu_t} \mathbb{E}_{\eta_t,\eta'_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 2\epsilon_t\eta_t(f(x_t)-\mu_t) - \frac{1}{2}(f(x_t)-\mu_t)^2 - \frac{1}{2}\mathscr{D}(p'_1,\dots,p'_n) \right] \right\} \\
&+ \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t,\mu_t} \mathbb{E}_{\eta_t,\eta'_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} -2\epsilon_t\eta'_t(f(x_t)-\mu_t) - \frac{1}{2}(f(x_t)-\mu_t)^2 - \frac{1}{2}\mathscr{D}(p'_1,\dots,p'_n) \right] \right\} \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p'_t,\mu_t} \mathbb{E}_{\eta_t\sim p_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 4\epsilon_t\eta_t(f(x_t)-\mu_t) - (f(x_t)-\mu_t)^2 - \mathscr{D}(p'_1,\dots,p'_n) \right] \right\} \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p'_t,\mu_t} \mathbb{E}_{\eta_t\sim p_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 4\epsilon_t\eta_t(f(x_t)-\mu_t) - (f(x_t)-\mu_t)^2 - \alpha\eta_t^2 \right] \right\} \\
&\le \left\langle\!\!\left\langle \sup_{x_t} \sup_{\mu_t,\eta_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 4\epsilon_t\eta_t(f(x_t)-\mu_t) - (f(x_t)-\mu_t)^2 - \alpha\eta_t^2 \right] \right\} \\
&= \sup_{\mathbf{x},\boldsymbol{\mu},\boldsymbol{\eta}} \mathbb{E}_\epsilon \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 4\epsilon_t\boldsymbol{\eta}_t(\epsilon)(f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon))^2 - \alpha\boldsymbol{\eta}_t(\epsilon)^2 \right]
\end{aligned}$$

This proves the first statement. For the case $\alpha = 0$, we have

$$\begin{aligned}
V_n^0 &\le \sup_{\mathbf{x},\boldsymbol{\mu},\boldsymbol{\eta}} \mathbb{E}_\epsilon \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 4\epsilon_t\boldsymbol{\eta}_t(\epsilon)(f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon))^2 \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t,\mu_t,\eta_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 4\epsilon_t\eta_t(f(x_t)-\mu_t) - (f(x_t)-\mu_t(\epsilon))^2 \right] \right\}
\end{aligned}$$

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

$$\begin{aligned}
V_n^0 &\le \left\langle\!\!\left\langle \sup_{x_t,\mu_t,q_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 4\epsilon_t\mathbb{E}(u_t)B(f(x_t)-\mu_t) - (f(x_t)-\mu_t(\epsilon))^2 \right] \right\} \\
&\le \left\langle\!\!\left\langle \sup_{x_t,\mu_t,q_t} \mathbb{E}_{u_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 4\epsilon_t u_t B(f(x_t)-\mu_t) - (f(x_t)-\mu_t(\epsilon))^2 \right] \right\} \\
&= \left\langle\!\!\left\langle \sup_{x_t,\mu_t,u_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 4\epsilon_t u_t B(f(x_t)-\mu_t) - (f(x_t)-\mu_t(\epsilon))^2 \right] \right\} \\
&= \left\langle\!\!\left\langle \sup_{x_t,\mu_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 4\epsilon_t B(f(x_t)-\mu_t) - (f(x_t)-\mu_t(\epsilon))^2 \right] \right\}
\end{aligned}$$

which is the same as the desired upper bound in (13), in the tree notation.

As for the lower bound, Recall from Eq. (24) that the value with $\alpha = 0$ is equal to

$$V_n^0 = \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t} \mathbb{E}_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 2(y_t-\mu_t)(f(x_t)-\mu_t) - (f(x_t)-\mu_t)^2 \right] \right\}. \tag{29}$$

For the purposes of a lower bound, let us pick particular distributions p_t as follows. Let $\epsilon_1,\dots,\epsilon_n$ be independent Rademacher random variables. Fix a $[-B/2,B/2]$-valued tree $\boldsymbol{\mu}$. Let $y_t = \boldsymbol{\mu}_t(\epsilon_{1:t-1}) + (B/2)\epsilon_t$. Hence, $y_t \in [-B,B]$ as

required. We can then lower bound the above expression as

$$V_n^0 \geq \sup_{\boldsymbol{\mu}} \left\langle\!\!\left\langle \sup_{x_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} 2\epsilon_t (f(x_t) - \boldsymbol{\mu}_t(\epsilon)) - (f(x_t) - \boldsymbol{\mu}_t(\epsilon))^2 \right] \right\}$$
$$= \sup_{\mathbf{x},\boldsymbol{\mu}} \mathbb{E} \sup_{f\in\mathscr{F}} \left[\sum_{t=1}^{n} B\epsilon_t (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon))^2 \right]$$

□

Proof of Lemma 7. For any $\lambda > 0$,

$$\mathbb{E} \max_{\mathbf{w}\in W} \left[\sum_{t=1}^{n} \epsilon_t \boldsymbol{\eta}_t(\epsilon) \mathbf{w}_t(\epsilon) - C\mathbf{w}_t(\epsilon)^2 - \alpha \boldsymbol{\eta}_t(\epsilon)^2 \right] \leq \frac{1}{\lambda} \log \mathbb{E} \sum_{\mathbf{w}\in W} \exp\left\{ \sum_{t=1}^{n} \lambda\epsilon_t \boldsymbol{\eta}_t(\epsilon) \mathbf{w}_t(\epsilon) - \lambda C\mathbf{w}_t(\epsilon)^2 - \lambda\alpha \boldsymbol{\eta}_t(\epsilon)^2 \right\}$$

Conditioning on $\epsilon_{1:n-1}$, we analyze

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\mathbf{w}\in W} \exp\left\{ \sum_{t=1}^{n} \lambda\epsilon_t \boldsymbol{\eta}_t(\epsilon) \mathbf{w}_t(\epsilon) - \lambda C\mathbf{w}_t(\epsilon)^2 - \lambda\alpha \boldsymbol{\eta}_t(\epsilon)^2 \right\} \;\middle|\; \epsilon_{1:n-1} \right] \\
&= \sum_{\mathbf{w}\in W} \exp\left\{ \sum_{t=1}^{n-1} \lambda\epsilon_t \boldsymbol{\eta}_t(\epsilon) \mathbf{w}_t(\epsilon)) - \sum_{t=1}^{n} \lambda C\mathbf{w}_t(\epsilon)^2 - \sum_{t=1}^{n} \lambda\alpha \boldsymbol{\eta}_t(\epsilon)^2 \right\} \mathbb{E}\left[\exp\left\{ \lambda\epsilon_n \boldsymbol{\eta}_n(\epsilon) \mathbf{w}_n(\epsilon) \right\} \;\middle|\; \epsilon_{1:n-1} \right] \\
&\leq \sum_{\mathbf{w}\in W} \exp\left\{ \sum_{t=1}^{n-1} \lambda\epsilon_t \eta_t \mathbf{w}_t(\epsilon) - \sum_{t=1}^{n-1} \lambda C\mathbf{w}_t(\epsilon)^2 - \sum_{t=1}^{n-1} \lambda\alpha \boldsymbol{\eta}_t(\epsilon)^2 \right\} \exp\left\{ \lambda^2 \boldsymbol{\eta}_n(\epsilon)^2 \mathbf{w}_n(\epsilon)^2/2 - \lambda C\mathbf{w}_n(\epsilon)^2 - \lambda\alpha \boldsymbol{\eta}_n(\epsilon)^2 \right\} \qquad (30)
\end{aligned}$$

The choice $\lambda = 2C/B^2$ ensures

$$\lambda^2 \boldsymbol{\eta}_n(\epsilon)^2 \mathbf{w}_n(\epsilon)^2/2 - \lambda C\mathbf{w}_n(\epsilon)^2 \leq 0$$

Alternatively, the choice $\lambda = 2\alpha/A^2$ ensures

$$\lambda^2 \boldsymbol{\eta}_n(\epsilon)^2 \mathbf{w}_n(\epsilon)^2/2 - \lambda\alpha \boldsymbol{\eta}_n(\epsilon)^2 \leq 0$$

In both cases, the exponential factor peeled off in (30) is no greater than 1. We proceed all the way to $t = 1$ to arrive at an upper bound of

$$\frac{1}{\lambda} \log \sum_{\mathbf{w}\in W} \exp\{0\} = \min\left\{ B^2(2C)^{-1}, A^2(2\alpha)^{-1} \right\} \log|W| \,.$$

The second statement (which already appears in [14]) 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. For any tree $\mathbf{w} \in W$,

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

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

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

and thus

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

Choosing

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

we obtain

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

□

Proof of Lemma 5. Let V' be a sequential γ-cover of $\mathscr{G}$ on $\mathbf{z}$ in the ℓ_2 sense, i.e.

$$\forall\epsilon,\ \ \forall g\in\mathscr{G},\ \ \exists\mathbf{v}\in V'\ \ \text{s.t.}\ \ \frac{1}{n}\sum_{t=1}^n (g(\mathbf{z}_t(\epsilon))-\mathbf{v}_t(\epsilon))^2 \le \gamma^2$$

Let us augment V' to include the all-zero tree, and denote the resulting set by $V = V'\cup\{\mathbf{0}\}$. Denote by $\mathbf{v}[\epsilon,g]$ a γ-close tree promised above, but we leave the choice for later. Then for any $c\in[0,1]$

$$\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n 4\epsilon_t\boldsymbol{\eta}_t(\epsilon)g(\mathbf{z}_t(\epsilon)) - g(\mathbf{z}_t(\epsilon))^2\right] \tag{31}$$

$$= \mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n 4\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t(\epsilon)\Big) - \Big(g(\mathbf{z}_t(\epsilon))^2 - c^2\mathbf{v}[\epsilon,g]_t(\epsilon)^2\Big)\right. \tag{32}$$

$$\left. + \Big(4\epsilon_t\boldsymbol{\eta}_t(\epsilon)\mathbf{v}[\epsilon,g]_t(\epsilon) - c^2\mathbf{v}[\epsilon,g]_t(\epsilon)^2\Big)\right] \tag{33}$$

$$\le \mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n 4\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t(\epsilon)\Big) - \sum_{t=1}^n\Big(g(\mathbf{z}_t(\epsilon))^2 - c^2\mathbf{v}[\epsilon,g]_t(\epsilon)^2\Big)\right] \tag{34}$$

$$+\mathbb{E}\max_{\mathbf{v}\in V'}\left[\sum_{t=1}^n 4\epsilon_t\boldsymbol{\eta}_t(\epsilon)\mathbf{v}_t(\epsilon) - c^2\mathbf{v}_t(\epsilon)^2\right] \tag{35}$$

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

$$\sum_{t=1}^n g(\mathbf{z}_t(\epsilon))^2 \ge c^2\sum_{t=1}^n \mathbf{v}[\epsilon,g]_t(\epsilon)^2 \tag{36}$$

and so we can drop the corresponding negative term in the supremum over $\mathscr{G}$. To prove this claim, first consider the easy case $\frac{1}{n}\sum_{t=1}^n g(\mathbf{z}_t(\epsilon))^2 \le C^2\gamma^2$, where $C = \frac{c}{1-c}$. Then we may choose $\mathbf{0}\in V$ as a tree that provides a sequential $C\gamma$-cover in the ℓ_2 sense. Clearly, (36) is then satisfied with this choice of $\mathbf{v}[\epsilon,g]=\mathbf{0}$. Now, assume $\frac{1}{n}\sum_{t=1}^n g(\mathbf{z}_t(\epsilon))^2 > C^2\gamma^2$. Fix any tree $\mathbf{v}[\epsilon,g]\in V$ that is γ-close in the ℓ_2 sense to g on the path ϵ. Denote $u = (\mathbf{v}[\epsilon,g]_1(\epsilon),\dots,\mathbf{v}[\epsilon,g]_n(\epsilon))$ and $h = (g(\mathbf{z}_1(\epsilon)),\dots,g(\mathbf{z}_n(\epsilon)))$. Thus, we have that $\|u-h\|\le\gamma$ and $\|h\|\ge C\gamma$ for the norm $\|h\|^2 = \frac{1}{n}\sum_{t=1}^n h_t^2$. Then

$$\|u\| \le \|u-h\| + \|h\| \le \gamma + \|h\| \le (C^{-1}+1)\|h\|$$

and thus $\|h\|\ge c\|u\|$ as desired. By choosing $c=1/2$, we have $C=1$ and thus the zero tree also provides a γ-cover.

We conclude that

$$\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n 4\epsilon_t\boldsymbol{\eta}_t(\epsilon)g(\mathbf{z}_t(\epsilon)) - g(\mathbf{z}_t(\epsilon))^2\right] \le 4\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n \epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t(\epsilon)\Big)\right] \tag{37}$$

$$+\mathbb{E}\max_{\mathbf{v}\in V'}\left[\sum_{t=1}^n 4\epsilon_t\boldsymbol{\eta}_t(\epsilon)\mathbf{v}_t(\epsilon) - (1/4)\mathbf{v}_t(\epsilon)^2\right] \tag{38}$$

where $\mathbf{v}[\epsilon,g]$ is defined to be the all-zero tree if $\frac{1}{n}\sum_{t=1}^n g(\mathbf{z}_t(\epsilon))^2\le\gamma^2$ and otherwise as an element of the cover V' that is γ-close to g on the path ϵ.

By Lemma 7, the term (38) is upper bounded as

$$\mathbb{E}_\epsilon \max_{\mathbf{v}\in V'}\left[\sum_{t=1}^n 4\epsilon_t \boldsymbol{\eta}_t(\epsilon)\mathbf{v}_t(\epsilon) - (1/4)\mathbf{v}_t(\epsilon)^2\right] \le 32B^2 \log \mathcal{N}_2(\gamma,\mathscr{G},\mathbf{z})$$

We now turn to the analysis of the first term on the right-hand side of (37). Let $\mathbf{v}[\epsilon,g]$ be denoted by $\mathbf{v}[\epsilon,g]^0$ and V be denoted by V^0. Let V^j denote a sequential $(2^{-j}\gamma)$-cover of $\mathscr{G}$ on the tree $\mathbf{z}$, for $j=1,\dots,N$, $N\ge 1$ to be specified later. We can now write

$$\begin{aligned}
&\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n \epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t^0(\epsilon)\Big)\right]\\
&= \mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n \epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t^N(\epsilon)\Big) + \sum_{t=1}^n\sum_{j=1}^N \epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(\mathbf{v}[\epsilon,g]_t^j(\epsilon) - \mathbf{v}[\epsilon,g]_t^{j-1}(\epsilon)\Big)\right]\\
&\le \mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n \epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t^N(\epsilon)\Big)\right] + \sum_{j=1}^N \mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n \epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(\mathbf{v}[\epsilon,g]_t^j(\epsilon) - \mathbf{v}[\epsilon,g]_t^{j-1}(\epsilon)\Big)\right]
\end{aligned}$$

From here, the analysis is very similar to the one in [15], except for the additional random variables $\boldsymbol{\eta}_t(\epsilon)$ multiplying the differences, and also for the minor fact that $\mathbf{v}[\epsilon,g]^0$ is defined as $\mathbf{0}$ for some (g,ϵ) pairs. This latter fact, however, does not affect the proof since $\mathbf{0}$ does provide a valid γ-cover when it is used.

First, by Cauchy-Schwartz inequality,

$$\begin{aligned}
\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n \epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t^N(\epsilon)\Big)\right] &\le n\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n \left(\frac{\epsilon_t\boldsymbol{\eta}_t(\epsilon)}{\sqrt{n}}\right)\left(\frac{1}{\sqrt{n}}\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t^N(\epsilon)\Big)\right)\right]\\
&\le n\mathbb{E}\left(\frac{1}{n}\sum_{t=1}^n \boldsymbol{\eta}_t(\epsilon)^2\right)^{-1/2}\beta_N\\
&\le B\beta_N n
\end{aligned}$$

where $\beta_j = 2^{-j}\gamma$. For the second term, fix a particular j and consider all pairs $(\mathbf{v}^s,\mathbf{v}^r)$ with $\mathbf{v}^s\in V^j$ and $\mathbf{v}^r\in V^{j-1}$. For each such pair, define a tree $\mathbf{w}^{(s,r)}$ by

$$\mathbf{w}_t^{(s,r)}(\epsilon) = \begin{cases}\mathbf{v}_t^s(\epsilon) - \mathbf{v}_t^r(\epsilon) & \text{if there exists } g\in\mathscr{G} \text{ s.t. } \mathbf{v}^s = \mathbf{v}[g,\epsilon]^j, \mathbf{v}^r = \mathbf{v}[g,\epsilon]^{j-1}\\ 0 & \text{otherwise.}\end{cases}$$

for all $t\in[n]$ and $\epsilon\in\{\pm1\}^n$. One can check that the tree is well-defined, and we set

$$W_j = \left\{w^{(s,r)} : 1\le s\le |V_j|, 1\le r\le |V_{j-1}|\right\}.$$

Then for any $j\in[N]$ and ϵ,

$$\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n \epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(\mathbf{v}[\epsilon,g]_t^j(\epsilon) - \mathbf{v}[\epsilon,g]_t^{j-1}(\epsilon)\Big)\right] \le \max_{\mathbf{w}\in W}\left[\sum_{t=1}^n \epsilon_t\boldsymbol{\eta}_t(\epsilon)\mathbf{w}_t(\epsilon)\right]$$

By the argument outlined in [15], for any $\mathbf{w}\in W^j$ and any path ϵ,

$$\sqrt{\sum_{t=1}^n \mathbf{w}_t(\epsilon)^2} \le 3\sqrt{n}\beta_j\,.$$

Putting everything together, and using Lemma 7,

$$\mathbb{E}_\epsilon\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^n \epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t^0(\epsilon)\Big)\right] \le B\beta_N n + B\sqrt{n}\sum_{j=1}^N 3\beta_j\sqrt{2\log(|V^j||V^{j-1}|)}$$

and the last term is upper bounded by

$$6B\sqrt{n}\sum_{j=1}^{N}\beta_j\sqrt{\log(|V^j|)} \le 12B\sqrt{n}\sum_{j=1}^{N}(\beta_j-\beta_{j+1})\sqrt{\log(|V^j|)} \le 12B\sqrt{n}\int_{\beta_{N+1}}^{\beta_0}\sqrt{\log \mathscr{N}_2(\delta\mathscr{G},\mathbf{z})}d\delta$$

Given any $\rho\in(0,\gamma)$, we let $N=\max\{j:\beta_j>2\rho\}$. Then $\beta_N<4\rho$ and $\beta_{N+1}>\rho$, and thus

$$\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon))-\mathbf{v}[\epsilon,g]_t^0(\epsilon)\Big)\right] \le B\inf_{\rho\in(0,\gamma)}\left\{4\rho n+12\sqrt{n}\int_{\rho}^{\gamma}\sqrt{\log \mathscr{N}_2(\delta,\mathscr{G},\mathbf{z})}d\delta\right\}$$

This concludes the proof. □

Proof of Lemma 6. The proof closely follows that of Lemma 5, except for the way we use Lemma 7 to take advantage of the subtracted quadratic term. We also employ an ℓ_∞ notion of sequential cover, rather than ℓ_2. To this end, let V' be a sequential γ-cover of $\mathscr{G}$ on $\mathbf{z}$ in the ℓ_∞ sense, i.e.

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

As before, let $V=V'\cup\{\mathbf{0}\}$ and denote by $\mathbf{v}[\epsilon,g]$ a γ-close tree promised by the definition. As in (31), for any $c\in[0,1]$,

$$\begin{aligned}
&\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}4\epsilon_t\boldsymbol{\eta}_t(\epsilon)g(\mathbf{z}_t(\epsilon))-g(\mathbf{z}_t(\epsilon))^2-\alpha\boldsymbol{\eta}_t(\epsilon)^2\right]\\
&\le\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}\left\{4\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon))-\mathbf{v}[\epsilon,g]_t(\epsilon)\Big)-\frac{\alpha}{2}\boldsymbol{\eta}_t(\epsilon)^2\right\}-\sum_{t=1}^{n}\Big(g(\mathbf{z}_t(\epsilon))^2-c^2\mathbf{v}[\epsilon,g]_t(\epsilon)^2\Big)\right]\\
&\quad+\mathbb{E}\max_{\mathbf{v}\in V'}\left[\sum_{t=1}^{n}4\epsilon_t\boldsymbol{\eta}_t(\epsilon)\mathbf{v}_t(\epsilon)-c^2\mathbf{v}_t(\epsilon)^2-\frac{\alpha}{2}\boldsymbol{\eta}_t(\epsilon)^2\right]
\end{aligned}$$

Following the proof of Lemma 5, we claim that for any ϵ,g there exists an element $\mathbf{v}[\epsilon,g]\in V$ such that for any $t\in[n]$,

$$|g(\mathbf{z}_t(\epsilon))|\ge c|\mathbf{v}[\epsilon,g]_t(\epsilon)| \tag{39}$$

First consider the easy case $\|g(\mathbf{z}_t(\epsilon))\|_\infty\le C\gamma$, where $C=\frac{c}{1-c}$. Then $\mathbf{0}\in V$ provides a sequential $C\gamma$-cover in the ℓ_∞ sense. If, on the other hand, $\|g(\mathbf{z}_t(\epsilon))\|_\infty>C\gamma$, we fix any tree $\mathbf{v}[\epsilon,g]\in V$ that is γ-close in the ℓ_∞ sense to g on the path ϵ. With the same argument as in the proof of Lemma 5, we conclude (39).

Hence,

$$\begin{aligned}
&\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}4\epsilon_t\boldsymbol{\eta}_t(\epsilon)g(\mathbf{z}_t(\epsilon))-g(\mathbf{z}_t(\epsilon))^2-\alpha\boldsymbol{\eta}_t(\epsilon)^2\right]\\
&\le 4\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon))-\mathbf{v}[\epsilon,g]_t(\epsilon)\Big)-\frac{\alpha}{2}\boldsymbol{\eta}_t(\epsilon)^2\right] &(40)\\
&\quad+\mathbb{E}\max_{\mathbf{v}\in V'}\left[\sum_{t=1}^{n}4\epsilon_t\boldsymbol{\eta}_t(\epsilon)\mathbf{v}_t(\epsilon)-(1/4)\mathbf{v}_t(\epsilon)^2-\frac{\alpha}{2}\boldsymbol{\eta}_t(\epsilon)^2\right] &(41)
\end{aligned}$$

By Lemma 7, the term (41) is upper bounded as

$$\mathbb{E}_\epsilon\max_{\mathbf{v}\in V'}\left[\sum_{t=1}^{n}4\epsilon_t\boldsymbol{\eta}_t(\epsilon)\mathbf{v}_t(\epsilon)-(1/4)\mathbf{v}_t(\epsilon)^2-\frac{\alpha}{2}\boldsymbol{\eta}_t(\epsilon)^2\right]\le\alpha^{-1}16A^2\log\mathscr{N}_\infty(\gamma,\mathscr{G},\mathbf{z}) \tag{42}$$

As for the term in (40), we write

$$\begin{aligned}
&\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon))-\mathbf{v}[\epsilon,g]_t^0(\epsilon)\Big)-\frac{\alpha}{2}\boldsymbol{\eta}_t(\epsilon)^2\right]\\
&=\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon))-\mathbf{v}[\epsilon,g]_t^N(\epsilon)\Big)-\frac{\alpha}{4}\boldsymbol{\eta}_t(\epsilon)^2\right.\\
&\qquad\left.+\sum_{t=1}^{n}\sum_{j=1}^{N}\left\{\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(\mathbf{v}[\epsilon,g]_t^j(\epsilon)-\mathbf{v}[\epsilon,g]_t^{j-1}(\epsilon)\Big)-\frac{\alpha}{4N}\boldsymbol{\eta}_t(\epsilon)^2\right\}\right]\\
&\le\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon))-\mathbf{v}[\epsilon,g]_t^N(\epsilon)\Big)-\frac{\alpha}{4}\boldsymbol{\eta}_t(\epsilon)^2\right]\\
&\qquad+\sum_{j=1}^{N}\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(\mathbf{v}[\epsilon,g]_t^j(\epsilon)-\mathbf{v}[\epsilon,g]_t^{j-1}(\epsilon)\Big)-\frac{\alpha}{4N}\boldsymbol{\eta}_t(\epsilon)^2\right]
\end{aligned}$$

Using Cauchy-Schwartz inequality along with $ab\le(1/2)(a^2+b^2)$,

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon))-\mathbf{v}[\epsilon,g]_t^N(\epsilon)\Big)&\le\sum_{t=1}^{n}\left(\frac{\sqrt{\alpha}\epsilon_t\boldsymbol{\eta}_t(\epsilon)}{\sqrt{2n}}\right)\left(\sqrt{\frac{2}{n\alpha}}\Big(g(\mathbf{z}_t(\epsilon))-\mathbf{v}[\epsilon,g]_t^N(\epsilon)\Big)\right)\\
&\le\frac{1}{4n}\sum_{t=1}^{n}\alpha\boldsymbol{\eta}_t(\epsilon)^2+\sum_{t=1}^{n}\frac{1}{n\alpha}\Big(g(\mathbf{z}_t(\epsilon))-\mathbf{v}[\epsilon,g]_t^N(\epsilon)\Big)^2
\end{aligned}$$

and thus

$$\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon))-\mathbf{v}[\epsilon,g]_t^N(\epsilon)\Big)-\frac{\alpha}{4}\boldsymbol{\eta}_t(\epsilon)^2\right]\le\alpha^{-1}\beta_N n$$

where $\beta_j=2^{-j}\gamma$. For the j-th link in the chain, recall that we can define

$$\mathbf{w}_t^{(s,r)}(\epsilon)=\begin{cases}\mathbf{v}_t^s(\epsilon)-\mathbf{v}_t^r(\epsilon) & \text{if there exists } g\in\mathscr{G} \text{ s.t. } \mathbf{v}^s=\mathbf{v}[g,\epsilon]^j,\mathbf{v}^r=\mathbf{v}[g,\epsilon]^{j-1}\\ 0 & \text{otherwise.}\end{cases}$$

for all $t\in[n]$ and $\epsilon\in\{\pm1\}^n$. Then for any $j\in[N]$ and ϵ,

$$\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(\mathbf{v}[\epsilon,g]_t^j(\epsilon)-\mathbf{v}[\epsilon,g]_t^{j-1}(\epsilon)\Big)-\frac{\alpha}{4N}\boldsymbol{\eta}_t(\epsilon)^2\right]\le\max_{\mathbf{w}\in W}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\mathbf{w}_t(\epsilon)-\frac{\alpha}{4N}\boldsymbol{\eta}_t(\epsilon)^2\right]$$

and it must hold by the definition of the cover that

$$|\mathbf{w}_t(\epsilon)|\le2\beta_j$$

for any $\mathbf{w}\in W^j$ and any path ϵ and any t. Putting everything together, and using Lemma 7,

$$\mathbb{E}_\epsilon\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon))-\mathbf{v}[\epsilon,g]_t^0(\epsilon)\Big)-\frac{\alpha}{2}\boldsymbol{\eta}_t(\epsilon)^2\right]\le\frac{n\beta_N}{\alpha}+\sum_{j=1}^{N}\frac{8N\beta_j^2}{\alpha}\log(|V^j||V^{j-1}|)$$

Simplifying and using $\beta_j=\beta_{j-1}-\beta_j$, we obtain an upper bound of

$$\begin{aligned}
\frac{n\beta_N}{\alpha}+\frac{16N}{\alpha}\sum_{j=1}^{N}\beta_j^2\log(|V^j|)&=\frac{n\beta_N}{\alpha}+\frac{16N}{\alpha}\sum_{j=1}^{N}(\beta_{j-1}-\beta_j)\beta_j\log(|V^j|)\\
&\le\frac{n\beta_N}{\alpha}+\frac{16N}{\alpha}\int_{\beta_{N+1}}^{\beta_0}\delta\log\mathscr{N}_\infty(\delta,\mathscr{G},\mathbf{z})d\delta
\end{aligned}$$

Given any $\rho \in (0,\gamma)$, we let $N = \max\{j : \beta_j > 2\rho\}$. Then $\beta_N < 4\rho$ and $\beta_{N+1} > \rho$. Further, $N \le \log(\gamma/\rho)$. Thus

$$\mathbb{E}\sup_{g\in\mathscr{G}}\left[\sum_{t=1}^{n}\epsilon_t\boldsymbol{\eta}_t(\epsilon)\Big(g(\mathbf{z}_t(\epsilon)) - \mathbf{v}[\epsilon,g]_t^0(\epsilon)\Big) - \frac{\alpha}{2}\boldsymbol{\eta}_t(\epsilon)^2\right]$$
$$\le \alpha^{-1}\inf_{\rho\in(0,\gamma)}\left\{4\rho n + 16\log(\gamma/\rho)\int_\rho^\gamma \delta\log\mathscr{N}_\infty(\delta,\mathscr{G},\mathbf{z})d\delta\right\}$$

Together with (42) this concludes the proof. □

Proof of Lemma 9. Fix a $\beta > 0$, and set $n = \mathrm{fat}_\beta(\mathscr{F})$. Suppose $\mathbf{x}$ is an $\mathscr{X}$-valued tree of depth n that is β-shattered by $\mathscr{F}$:

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

where $\boldsymbol{\mu}$ is the witness to shattering. Since functions in $\mathscr{F}$ take values in $[-1,1]$, it is also the case that $\boldsymbol{\mu}$ is $[-1,1]$-valued, and thus $|f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)| \le 2$ for all $f \in \mathscr{F}$. Then from (14) with the particular choices of $\mathbf{x}$ and $\boldsymbol{\mu}$ described above,

$$V_n^0 \ge \mathbb{E}\sup_{f\in\mathscr{F}}\left[\sum_{t=1}^{n} 4\epsilon_t(f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon))^2\right] \tag{43}$$
$$\ge \mathbb{E}\sup_{f\in\mathscr{F}}\left[\sum_{t=1}^{n} 4\epsilon_t(f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - 2|f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)|\right] \tag{44}$$
$$\ge \mathbb{E}\left[\sum_{t=1}^{n} 4\epsilon_t(f^\epsilon(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - 2|f^\epsilon(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)|\right] \tag{45}$$

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

$$\mathbb{E}\left[\sum_{t=1}^{n} 4|f^\epsilon(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)| - 2|f^\epsilon(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)|\right] \ge \mathbb{E}\left[\sum_{t=1}^{n}\beta\right] = n\beta$$

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

Proof of Lemma 10. Assume $d = \mathrm{fat}_\beta(\mathscr{F}') \le n$. 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

$$\forall\, t \in [n],\ \epsilon_t(f^\epsilon(\mathbf{z}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) \ge \beta/2 \tag{46}$$

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 (46) is satisfied with equality: $\epsilon_t(g^\epsilon(\mathbf{z}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) = \beta/2$. 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 $\mathscr{G} \subseteq \mathscr{F}$. Let $\mathbf{s}$ be the witness tree for the shattering. We will now show a construction of particular trees of depth

$$n' = \left\lceil \frac{n}{\mathrm{fat}_\beta} \right\rceil \mathrm{fat}_\beta \tag{47}$$

using the pair $\mathbf{z},\mathbf{s}$. Define $k = \lceil \frac{n}{\mathrm{fat}_\beta} \rceil = \frac{n'}{\mathrm{fat}_\beta} \geq 1$ 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

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

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:

$$\begin{aligned} V^0_{n'} &\geq \mathbb{E} \sup_{f \in \mathscr{F}} \left[\sum_{t=1}^{n'} 2\epsilon_t (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon)) - (f(\mathbf{x}_t(\epsilon)) - \boldsymbol{\mu}_t(\epsilon))^2 \right] \\ &= \mathbb{E} \sup_{f \in \mathscr{F}} \left[\sum_{t=1}^{n'} 2\epsilon_t (f(\mathbf{z}_{\lceil \frac{t}{k} \rceil}(\tilde{\epsilon})) - \mathbf{s}_{\lceil \frac{t}{k} \rceil}(\tilde{\epsilon})) - (f(\mathbf{z}_{\lceil \frac{t}{k} \rceil}(\tilde{\epsilon})) - \mathbf{s}_{\lceil \frac{t}{k} \rceil}(\tilde{\epsilon}))^2 \right]. \end{aligned}$$

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

$$\begin{aligned} &\mathbb{E} \sup_{f \in \mathscr{F}} \left[\sum_{i=1}^{\mathrm{fat}_\beta} \sum_{j=(i-1)k+1}^{i \cdot k} 2\epsilon_j (f(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon})) - (f(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon}))^2 \right] \\ &= \mathbb{E} \sup_{f \in \mathscr{F}} \left[\sum_{i=1}^{\mathrm{fat}_\beta} 2(f(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon})) \left(\sum_{j=(i-1)k+1}^{i \cdot k} \epsilon_j \right) - k(f(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon}))^2 \right] \\ &= \mathbb{E} \sup_{f \in \mathscr{F}} \left[\sum_{i=1}^{\mathrm{fat}_\beta} 2\tilde{\epsilon}_i (f(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon})) \left| \sum_{j=(i-1)k+1}^{i \cdot k} \epsilon_j \right| - k(f(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon}))^2 \right] \end{aligned}$$

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

$$\begin{aligned} \mathbb{E} \sup_{g \in \mathscr{G}} \left[\sum_{i=1}^{\mathrm{fat}_\beta} 2\tilde{\epsilon}_i (f(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon})) \left| \sum_{j=(i-1)k+1}^{i \cdot k} \epsilon_j \right| - k(f(\mathbf{z}_i(\tilde{\epsilon})) - \mathbf{s}_i(\tilde{\epsilon}))^2 \right] &\geq \mathbb{E} \sum_{i=1}^{\mathrm{fat}_\beta} \left(\beta \left| \sum_{j=(i-1)k+1}^{i \cdot k} \epsilon_j \right| - \frac{k\beta^2}{4} \right) \\ &\geq \mathrm{fat}_\beta(\mathscr{F}) \left(\beta \sqrt{\frac{k}{2}} - \frac{k\beta^2}{4} \right) \end{aligned}$$

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

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

and

$$\mathrm{fat}_\beta(\mathscr{F}) \frac{k\beta^2}{4} = \frac{1}{4} n' \beta^2$$

We conclude that

$$V^0_{n'} \geq \frac{1}{4} \left(2\sqrt{2} \beta \sqrt{n' \mathrm{fat}_\beta(\mathscr{F})} - n' \beta^2 \right) \tag{48}$$

Now suppose $\mathrm{fat}_\beta(\mathscr{F}) = c/\beta^p$ for some $c > 0$. First, we need to ensure that $\mathrm{fat}_\beta(\mathscr{F}) = c/\beta^p \le n'$, as required by our construction. This means that $\beta \ge (cn')^{-1/p}$. Plugging in the rate of $\mathrm{fat}_\beta(\mathscr{F})$ into (48),

$$2\sqrt{2}\beta\sqrt{n'\mathrm{fat}_\beta(\mathscr{F})} - n'\beta^2 = 2\sqrt{2}c^{1/2}\beta^{1-p/2}\sqrt{n'} - n'\beta^2$$

Observe that the setting of $\beta = (32c)^{1/(2+p)}(n')^{-1/(p+2)}$ yields a lower bound of

$$c_p \cdot (n')^{\frac{p}{p+2}}$$

where c_p denotes a constant that may depend on p, and whose value may change from line to line.

Examining (29), we see that V_n^0 is nondecreasing with n. To illustrate this, let $n' > n$. For $t \in \{n+1, \dots, n'\}$, we may choose p_t in (29) 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, $V_{n'}^0 \ge V_n^0$. In view of (47) and the above discussion, $V_{n'}^0 \le V_{2n-1}^0$, and thus

$$V_{2n}^0 \ge V_{2n-1}^0 \ge V_{n'}^0 \ge c_p n^{\frac{p}{p+2}} .$$

□

Proof of Lemma 11. First note that when $t = n$ the initial condition is trivially satisfied as

$$\mathfrak{R}_n(x_{1:n}, y_{1:n}) = \sup_{f\in\mathscr{F}}\left\{-\sum_{j=1}^n (f(x_j) - y_j)^2\right\} = -\inf_{f\in\mathscr{F}}\sum_{j=1}^n (f(x_j) - y_j)^2 .$$

Let us denote

$$\widehat{L}_t(f) = \sum_{j=1}^t (f(x_j) - y_j)^2$$

and

$$A_{t+1}(f) = \sum_{j=t+1}^n B\epsilon_j(f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon)) - (f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon))^2$$

To check admissibility note that we need to check the inequality in Equation (49). To do so note that for any $x_t \in \mathscr{X}$, $p_t \in \Delta([-B, B])$,

$$\mathbb{E}_{y_t\sim p_t}\left[\left(\mathbb{E}_{y_t\sim p_t}[y_t] - y_t\right)^2\right] + \mathbb{E}_{y_t\sim p_t}\left[\mathfrak{R}_n(x_{1:t}, y_{1:t})\right] = \mathbb{E}_{y_t\sim p_t}\left[\left(\mathbb{E}_{y_t\sim p_t}[y_t] - y_t\right)^2 + \sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon \sup_{f\in\mathscr{F}}\left\{A_{t+1}(f) - \widehat{L}_t(f)\right\}\right]$$

Expanding the square in the first term and then the loss of f at time t, we obtain

$$\mathbb{E}_{y_t\sim p_t}\left[\left(\mathbb{E}_{y_t\sim p_t}[y_t]\right)^2 - 2y_t\mathbb{E}_{y_t\sim p_t}[y_t] + y_t^2 + \sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon \sup_{f\in\mathscr{F}}\left\{A_{t+1}(f) - \widehat{L}_t(f)\right\}\right]$$

$$= \mathbb{E}_{y_t\sim p_t}\left[\left(\mathbb{E}_{y_t\sim p_t}[y_t]\right)^2 - 2y_t\mathbb{E}_{y_t\sim p_t}[y_t] + \sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon \sup_{f\in\mathscr{F}}\left\{A_{t+1}(f) - f^2(x_t) + 2f(x_t)y_t - \widehat{L}_{t-1}(f)\right\}\right]$$

Rearranging, the above is equal to

$$\mathbb{E}_{y_t\sim p_t}\left[\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon \sup_{f\in\mathscr{F}}\left\{A_{t+1}(f) - \widehat{L}_{t-1}(f) + 2\left(\mathbb{E}_{y_t\sim p_t}[y_t]\right)^2 - f^2(x_t) - \left(\mathbb{E}_{y_t\sim p_t}[y_t]\right)^2 + 2(f(x_t) - \mathbb{E}_{y_t\sim p_t}[y_t])y_t\right\}\right]$$

$$= \mathbb{E}_{y_t\sim p_t}\left[\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon \sup_{f\in\mathscr{F}}\left\{A_{t+1}(f) - \widehat{L}_{t-1}(f) + 2\left(\mathbb{E}_{y_t\sim p_t}[y_t]\right)^2 - \left(f(x_t) - \mathbb{E}_{y_t\sim p_t}[y_t]\right)^2 - 2f(x_t)\mathbb{E}_{y_t\sim p_t}[y_t] + 2\left(f(x_t) - \mathbb{E}_{y_t\sim p_t}[y_t]\right)y_t\right\}\right]$$

which is

$$\mathbb{E}_{y_t\sim p_t}\left[\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon \sup_{f\in\mathscr{F}}\left\{A_{t+1}(f) - \widehat{L}_{t-1}(f) - 2\left(\mathbb{E}_{y_t\sim p_t}[y_t]\right)^2 - \left(f(x_t) - \mathbb{E}_{y_t\sim p_t}[y_t]\right)^2 - 2\left(f(x_t) - \mathbb{E}_{y_t\sim p_t}[y_t]\right)\mathbb{E}_{y_t\sim p_t}[y_t] + 2\left(f(x_t) - \mathbb{E}_{y_t\sim p_t}[y_t]\right)y_t\right\}\right]$$

$$\le \mathbb{E}_{y_t\sim p_t}\left[\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon \sup_{f\in\mathscr{F}}\left\{A_{t+1}(f) - \widehat{L}_{t-1}(f) - \left(f(x_t) - \mathbb{E}_{y_t\sim p_t}[y_t]\right)^2 + 2\left(f(x_t) - \mathbb{E}_{y_t\sim p_t}[y_t]\right)\left(y_t - \mathbb{E}_{y_t\sim p_t}[y_t]\right)\right\}\right]$$

By Jensen's inequality, the above can be upper bounded by

$$\mathbb{E}_{y_t,y_t'\sim p_t}\left[\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon\sup_{f\in\mathscr{F}}\left\{A_{t+1}(f)-\widehat{L}_{t-1}(f)-\left(f(x_t)-\mathbb{E}_{y_t\sim p_t}[y_t]\right)^2+2\left(f(x_t)-\mathbb{E}_{y_t\sim p_t}[y_t]\right)(y_t-y_t')\right\}\right]$$
$$=\mathbb{E}_{y_t,y_t'\sim p_t,\epsilon_t}\left[\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon\sup_{f\in\mathscr{F}}\left\{A_{t+1}(f)-\widehat{L}_{t-1}(f)-\left(f(x_t)-\mathbb{E}_{y_t\sim p_t}[y_t]\right)^2+2\epsilon_t\left(f(x_t)-\mathbb{E}_{y_t\sim p_t}[y_t]\right)(y_t-y_t')\right\}\right]$$

Since the inequalities above hold for any $x_t\in\mathscr{X}$, $p_t\in\Delta([-B,B])$, we have

$$\sup_{x_t\in\mathscr{X},p_t\in\Delta([-B,B])}\left[\mathbb{E}_{y_t\sim p_t}\left[\left(\mathbb{E}_{y_t\sim p_t}[y_t]-y_t\right)^2\right]+\mathbb{E}_{y_t\sim p_t}\left[\mathfrak{R}_n\left(x_{1:t},y_{1:t}\right)\right]\right]$$
$$\le\sup_{x_t\in\mathscr{X},p_t\in\Delta([-B,B])}\mathbb{E}_{y_t,y_t'\sim p_t,\epsilon_t}\left[\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon\sup_{f\in\mathscr{F}}\left\{A_{t+1}(f)-\widehat{L}_{t-1}(f)-\left(f(x_t)-\mathbb{E}_{y_t\sim p_t}[y_t]\right)^2+2\epsilon_t\left(f(x_t)-\mathbb{E}_{y_t\sim p_t}[y_t]\right)(y_t-y_t')\right\}\right]$$
$$\le\sup_{\substack{x_t\in\mathscr{X}\\y_t,y_t',\mu_t\in[-B,B]}}\mathbb{E}_{\epsilon_t}\left[\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon\sup_{f\in\mathscr{F}}\left\{A_{t+1}(f)-\widehat{L}_{t-1}(f)-\left(f(x_t)-\mu_t\right)^2+2\epsilon_t\left(f(x_t)-\mu_t\right)(y_t-y_t')\right\}\right]$$
$$\le\sup_{\substack{x_t\in\mathscr{X}\\y_t,\mu_t\in[-B,B]}}\mathbb{E}_{\epsilon_t}\left[\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon\sup_{f\in\mathscr{F}}\left\{A_{t+1}(f)-\widehat{L}_{t-1}(f)-\left(f(x_t)-\mu_t\right)^2+4\epsilon_t\left(f(x_t)-\mu_t\right)y_t\right\}\right]$$

Since the above is convex in y_t, we can replace the supremum over $[-B,B]$ to supremum over $\{-B,B\}$

$$\sup_{\substack{x_t\in\mathscr{X},\mu_t\in[-B,B]\\y_t\in\{-B,B\}}}\mathbb{E}_{\epsilon_t}\left[\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon\sup_{f\in\mathscr{F}}\left\{A_{t+1}(f)-\widehat{L}_{t-1}(f)-\left(f(x_t)-\mu_t\right)^2+4\epsilon_t\left(f(x_t)-\mu_t\right)y_t\right\}\right]$$
$$=\sup_{\substack{x_t\in\mathscr{X}\\\mu_t\in[-B,B]}}\mathbb{E}_{\epsilon_t}\left[\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon\sup_{f\in\mathscr{F}}\left\{A_{t+1}(f)-\widehat{L}_{t-1}(f)-\left(f(x_t)-\mu_t\right)^2+4B\epsilon_t\left(f(x_t)-\mu_t\right)\right\}\right]$$
$$=\sup_{\mathbf{x},\boldsymbol{\mu}}\mathbb{E}_\epsilon\left[\sup_{f\in\mathscr{F}}\left\{\sum_{j=t}^{n}B\epsilon_j(f(\mathbf{x}_j(\epsilon))-\boldsymbol{\mu}_j(\epsilon))-(f(\mathbf{x}_j(\epsilon))-\boldsymbol{\mu}_j(\epsilon))^2-\widehat{L}_{t-1}(f)\right\}\right]=\mathfrak{R}_n\left(x_{1:t-1},y_{1:t-1}\right)$$

Thus we have shown that $\mathfrak{R}_n$ is an admissible relaxation. Further, $(\hat{y}-y_t)^2+\mathfrak{R}_n\left(x_{1:t},(y_{1:t-1},y_t)\right)$ is a convex function of y_t and so for the estimator one can use

$$\hat{y}_t=\frac{\mathfrak{R}_n(x_{1:t},(y_{1:t-1},B))-\mathfrak{R}_n(x_{1:t},(y_{1:t-1},-B))}{4B}$$

(no clipping is needed above as $\hat{y}_t$ is always between $-B$ and B). For the above estimator one enjoys the regret bound

$$\mathbf{Reg}_n\le\mathfrak{R}_n(\cdot)$$

Note that this is exactly the bound in Eq. (13). □

Proof of Proposition 12. Notice that the above recipe closely follows the notion of relaxation provided in [13]. All we need to do is check that the relaxation derived satisfies admissibility and initial conditions. By Step 1 of the recipe, since the offset Rademacher relaxation is admissible to start with, the derived relaxation also satisfies initial condition. To show admissibility condition notice that the set $[-B,B]$ is compact and convex and $(\hat{y}_t-y_t)^2+\mathbf{Rel}_n\left(x_{1:t},y_{1:t}\right)$ is a convex function of $\hat{y}_t$. Hence applying minimax theorem, we see that,

$$\inf_{\hat{y}_t\in[-B,B]}\sup_{y_t\in[-B,B]}\left\{(\hat{y}_t-y_t)^2+\mathbf{Rel}_n\left(x_{1:t},y_{1:t}\right)\right\}$$
$$=\sup_{p_t\in\Delta([-B,B])}\inf_{\hat{y}_t}\left\{\mathbb{E}_{y_t\sim p_t}\left[(\hat{y}_t-y_t)^2+\mathbf{Rel}_n\left(x_{1:t},y_{1:t}\right)\right]\right\}$$
$$=\sup_{p_t\in\Delta([-B,B])}\left\{\inf_{\hat{y}_t}\mathbb{E}_{y_t\sim p_t}\left[(\hat{y}_t-y_t)^2\right]+\mathbb{E}_{y_t\sim p_t}\left[\mathbf{Rel}_n\left(x_{1:t},y_{1:t}\right)\right]\right\}$$
$$=\sup_{p_t\in\Delta([-B,B])}\left\{\mathbb{E}_{y_t\sim p_t}\left[\left(\mathbb{E}_{y_t\sim p_t}[y_t]-y_t\right)^2\right]+\mathbb{E}_{y_t\sim p_t}\left[\mathbf{Rel}_n\left(x_{1:t},y_{1:t}\right)\right]\right\}$$

Hence the admissibility condition can be rewritten as :

$$\forall x_t \in \mathscr{X}, \quad \sup_{p_t \in \Delta([-B,B])} \left\{ \mathbb{E}_{y_t \sim p_t} \left[\left(\mathbb{E}_{y_t \sim p_t} [y_t] - y_t \right)^2 \right] + \mathbb{E}_{y_t \sim p_t} \left[\mathbf{Rel}_n \left(x_{1:t}, y_{1:t} \right) \right] \right\} \leq \mathbf{Rel}_n \left(x_{1:t-1}, y_{1:t-1} \right) \tag{49}$$

□

Proof of Corollary 13. As done in [13] for the case of finite class of experts, in the Rademacher relaxation one can replace the $\max_{f \in \mathscr{F}}$ with a limit of soft-max as follows:

$$\begin{aligned}
\mathfrak{R}_n(x_{1:t}, y_{1:t}) &= \sup_{\mathbf{x}, \boldsymbol{\mu}} \mathbb{E}_\epsilon \max_{f \in \mathscr{F}} \left[\sum_{j=t+1}^{n} 4B\epsilon_j (f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon)) - (f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon))^2 - \sum_{j=1}^{t} (f(x_j) - y_j)^2 \right] \\
&= \sup_{\mathbf{x}, \boldsymbol{\mu}} \mathbb{E}_\epsilon \inf_{\lambda > 0} \lambda^{-1} \log \left(\sum_{f \in \mathscr{F}} \exp \left(\lambda \sum_{j=t+1}^{n} 4B\epsilon_j (f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon)) - (f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon))^2 - \lambda \sum_{j=1}^{t} (f(x_j) - y_j)^2 \right) \right) \\
&\leq \inf_{\lambda > 0} \left\{ \lambda^{-1} \log \left(\sum_{f \in \mathscr{F}} \exp \left(-\lambda \sum_{j=1}^{t} (f(x_j) - y_j)^2 \right) \right) \right. \\
&\qquad \left. + \sup_{\mathbf{x}, \boldsymbol{\mu}} \lambda^{-1} \log \left(\mathbb{E}_\epsilon \exp \left(\lambda \sum_{j=t+1}^{n} 4B\epsilon_j (f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon)) - (f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon))^2 \right) \right) \right\}
\end{aligned}$$

Not notice that if we set $\lambda = B^{-2}$, the proof of Lemma 7 exactly shows that

$$\sup_{\mathbf{x}, \boldsymbol{\mu}} \lambda^{-1} \log \left(\mathbb{E}_\epsilon \exp \left(\lambda \sum_{j=t+1}^{n} 4B\epsilon_j (f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon)) - (f(\mathbf{x}_j(\epsilon)) - \boldsymbol{\mu}_j(\epsilon))^2 \right) \right) \leq B^2 \log |\mathscr{F}|$$

Hence we arrive at our relaxation

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

Now to show admissibility, note that

$$\begin{aligned}
&\sup_{x_t, p_t} \mathbb{E}_{y_t \sim p_t} \left[(y_t - \mathbb{E}[y_t])^2 + \mathbf{Rel}_n \left(x_{1:t}, y_{1:t} \right) \right] \\
&\quad = \sup_{x_t, p_t} \mathbb{E}_{y_t \sim p_t} \left[y_t^2 - (\mathbb{E}[y_t])^2 + B^2 \log \left(\sum_{f \in \mathscr{F}} \exp \left(-B^{-2} \sum_{j=1}^{t} (f(x_j) - y_j)^2 \right) \right) \right] \\
&\quad = \sup_{x_t, p_t} \mathbb{E}_{y_t \sim p_t} \left[B^2 \log \left(\exp \left(B^{-2} y_t^2 - B^{-2} (\mathbb{E}[y_t])^2 \right) \right) + B^2 \log \left(\sum_{f \in \mathscr{F}} \exp \left(-B^{-2} \sum_{j=1}^{t} (f(x_j) - y_j)^2 \right) \right) \right] \\
&\quad = \sup_{x_t, p_t} \mathbb{E}_{y_t \sim p_t} \left[B^2 \log \left(\sum_{f \in \mathscr{F}} \exp \left(B^{-2} y_t^2 - B^{-2} (\mathbb{E}[y_t])^2 - B^{-2} \sum_{j=1}^{t} (f(x_j) - y_j)^2 \right) \right) \right] \\
&\quad = \sup_{x_t, p_t} \mathbb{E}_{y_t \sim p_t} \left[B^2 \log \left(\sum_{f \in \mathscr{F}} \exp \left(-B^{-2} (\mathbb{E}[y_t])^2 + 2B^{-2} f(x_t) y_t - B^{-2} f^2(x_t) - B^{-2} \sum_{j=1}^{t-1} (f(x_j) - y_j)^2 \right) \right) \right] \\
&\quad = \sup_{x_t, p_t} \mathbb{E}_{y_t \sim p_t} \left[B^2 \log \left(\sum_{f \in \mathscr{F}} \exp \left(-B^{-2} (\mathbb{E}[y_t] - f(x_t))^2 + 2B^{-2} f(x_t) (y_t - \mathbb{E}[y_t]) - B^{-2} \sum_{j=1}^{t-1} (f(x_j) - y_j)^2 \right) \right) \right]
\end{aligned}$$

Now by convexity (see [1]) we can take the expectation w.r.t. y_t inside and hence we get,

$$\begin{aligned}
&\sup_{x_t,p_t} \mathbb{E}_{y_t\sim p_t}\left[(y_t-\mathbb{E}\left[y_t\right])^2+\mathbf{Rel}_n\left(x_{1:t},y_{1:t}\right)\right]\\
&\quad\le \sup_{x_t,p_t}\left\{B^2\log\left(\sum_{f\in\mathscr{F}}\exp\left(-B^{-2}(\mathbb{E}\left[y_t\right]-f(x_t))^2-B^{-2}\sum_{j=1}^{t-1}(f(x_j)-y_j)^2\right)\right)\right\}\\
&\quad\le B^2\log\left(\sum_{f\in\mathscr{F}}\exp\left(-B^{-2}\sum_{j=1}^{t-1}(f(x_j)-y_j)^2\right)\right)\\
&\quad= \mathbf{Rel}_n\left(x_{1:t-1},y_{1:t-1}\right)
\end{aligned}$$

Again as we used above (see [1]) we have that the relaxation is such that $(\hat{y}-y_t)^2+\mathbf{Rel}_n\left(x_{1:t},(y_{1:t-1},y_t)\right)$ is a convex function of y_t and so the estimator is given by

$$\begin{aligned}
\hat{y}_t &= \mathrm{Clip}\left(\frac{\mathbf{Rel}_n\left(x_{1:t},(y_{1:t-1},B)\right)-\mathbf{Rel}_n\left(x_{1:t},(y_{1:t-1},-B)\right)}{4B}\right)\\
&= \mathrm{Clip}\left(\frac{B}{4}\log\left(\frac{\sum_{f\in\mathscr{F}}\exp\left(-B^{-2}\sum_{j=1}^{t-1}(f(x_j)-y_j)^2-B^{-2}(f(x_t)-B)^2\right)}{\sum_{f\in\mathscr{F}}\exp\left(-B^{-2}\sum_{j=1}^{t-1}(f(x_j)-y_j)^2-B^{-2}(f(x_t)+B)^2\right)}\right)\right)
\end{aligned}$$

Now the final regret bound we obtain is given by $\mathbf{Reg}_n\le\mathbf{Rel}_n(\cdot)$ and so we conclude that

$$\mathbf{Reg}_n\le B^2\log|\mathscr{F}|$$

□

Proof of Corollary 14. For simplicity, each input instance $x_t\in\mathbb{R}^d$ we define vector in $\mathbb{R}^{d+1}$ as $z_t=(0,x_t)$, the vector obtained by concatenating 0 before x_t. Further given trees $\mathbf{x}$ and μ, we write the $\mathbf{z}$ as the $[-B,B]\times\mathscr{X}$ valued tree corresponding to $\mathbf{x}$ and μ obtained by concatenating μ's before x's on every node. Also for every linear predictor $f\in\mathscr{F}$ define corresponding $w=(-1,f)$. The unnormalized regret over the rounds $-d$ to n can be written as

$$\sum_{t=1}^{n}(\hat{y}_t-y_t)^2-\inf_{w}\left\{\sum_{t=1}^{n}(\langle w,z_t\rangle-y_t)^2+\lambda\left\|w\right\|_2^2\right\}$$

Hence, we have,

$$\begin{aligned}
\mathfrak{R}_n(x_{1:t},y_{1:t}) &= \sup_{\mathbf{z}}\mathbb{E}_\epsilon\sup_{f\in\mathscr{F}}\left[\sum_{j=t+1}^{n-1}4B\epsilon_j\left\langle w,\mathbf{z}_j(\epsilon)\right\rangle-(\langle w,\mathbf{z}_j(\epsilon)\rangle)^2-\sum_{j=1}^{t}(\langle w,z_j\rangle-y_j)^2-\lambda\left\|w\right\|_2^2\right]\\
&= 2\sup_{\mathbf{z}}\mathbb{E}_\epsilon\sup_{w}\left[\left\langle w,\sum_{j=t+1}^{n}2B\epsilon_j\mathbf{z}_j(\epsilon)+\sum_{j=1}^{t}y_jz_j\right\rangle-\frac{1}{2}w^\top\left(\sum_{j=t+1}^{n}\mathbf{z}_j(\epsilon)\mathbf{z}_j(\epsilon)^\top+\sum_{j=1}^{t}z_jz_j^\top+\lambda I\right)w^\top\right]-\sum_{j=1}^{t}y_j^2
\end{aligned}$$

Let us denote $\mathbf{A}_{t+1:n}(\mathbf{z})=\sum_{j=t+1}^{n}\mathbf{z}_j(\epsilon)\mathbf{z}_j(\epsilon)^\top$ and $\mathbf{B}_t=\sum_{j=1}^{t}z_jz_j^\top$. Using Fenchel-Young inequality for

$$\frac{1}{2}w^\top\left(\mathbf{A}_{t+1:n}(\mathbf{z})+\mathbf{B}_t+\lambda I\right)w^\top$$

and its conjugate we get,

$$\mathfrak{R}_n(x_{1:t},y_{1:t})\le\sup_{\mathbf{z}}\mathbb{E}_\epsilon\left\|\sum_{j=t+1}^{n}2B\epsilon_j\mathbf{z}_j(\epsilon)+\sum_{j=1}^{t}y_jz_j\right\|^2_{(\mathbf{A}_{t+1:n}(\mathbf{z})+\mathbf{B}_t+\lambda I)^{-1}}-\sum_{j=1}^{t}y_j^2$$

The idea now is to obtain a further upper bound by removing the dependence on the tree $\mathbf{z}$. Opening the square with only the n-th term, the above expression is equal to

$$\sup_{\mathbf{z}} \mathbb{E}_{\epsilon}\left[\left\|\sum_{j=t+1}^{n-1} 2B\epsilon_j \mathbf{z}_j(\epsilon)+\sum_{j=1}^{t} y_j z_j\right\|^2_{(\mathbf{A}_{t+1:n}(\mathbf{z})+\mathbf{B}_t+\lambda I)^{-1}} - \sum_{j=1}^{t} y_j^2 + 4B^2 \mathbf{z}_n(\epsilon)^{\top}\left(\mathbf{A}_{t+1:n}(\mathbf{z})+\mathbf{B}_t+\lambda I\right)^{-1}\mathbf{z}_n(\epsilon)\right]$$

By the standard argument we may upper bound the quadratic terms by a ratio of determinants Δ:

$$\mathbf{z}_n(\epsilon)^{\top}\left(\mathbf{A}_{t+1:n}(\mathbf{z})+\mathbf{B}_t+\lambda I\right)^{-1}\mathbf{z}_n(\epsilon) \le \left(1-\frac{\Delta\left(\mathbf{A}_{t+1:n-1}(\mathbf{z})+\mathbf{B}_t+\lambda I\right)}{\Delta\left(\mathbf{A}_{t+1:n}(\mathbf{z})+\mathbf{B}_t+\lambda I\right)}\right)$$

Using the inequality $1-x \le -\log(x)$ for $x>0$, we obtain an upper bound

$$\sup_{\mathbf{z}}\left\{\mathbb{E}_{\epsilon}\left[\left\|\sum_{j=t+1}^{n-1} 2B\epsilon_j \mathbf{z}_j(\epsilon)+\sum_{j=1}^{t} y_j z_j\right\|^2_{(\mathbf{A}_{t+1:n-1}(\mathbf{z})+\mathbf{B}_t+\lambda I)^{-1}}\right] - \sum_{j=1}^{t} y_j^2 + 4B^2\,\mathbb{E}_{\epsilon}\left[\log\left(\frac{\Delta\left(\mathbf{A}_{t+1:n}(\mathbf{z})+\mathbf{B}_t+\lambda I\right)}{\Delta\left(\mathbf{A}_{t+1:n-1}(\mathbf{z})+\mathbf{B}_t+\lambda I\right)}\right)\right]\right\}$$

Proceeding in similar fashion by peeling off terms from the norm, we arrive at,

$$\begin{aligned}\mathfrak{R}_n(x_{1:t}, y_{1:t}) &\le \left\|\sum_{j=1}^{t} y_j z_j\right\|^2_{(\mathbf{B}_t+\lambda I)^{-1}} - \sum_{j=1}^{t} y_j^2 + 4B^2 \sup_{\mathbf{z}}\mathbb{E}_{\epsilon}\left[\log\left(\frac{\Delta\left(\mathbf{A}_{t+1:n}(\mathbf{z})+\mathbf{B}_t+\lambda I\right)}{\Delta\left(\mathbf{B}_t+\lambda I\right)}\right)\right] \\ &\le \left\|\sum_{j=1}^{t} y_j z_j\right\|^2_{(\mathbf{B}_t+\lambda I)^{-1}} + 4B^2\log\left(\frac{(n/d)^d}{\Delta\left(\mathbf{B}_t+\lambda I\right)}\right) - \sum_{j=1}^{t} y_j^2\end{aligned}$$

and we take this last expression as our relaxation $\mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)$. Now notice that since z_t's are 0 on the first coordinate, the relaxation can be rewritten as

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

where $\tilde{\mathbf{B}}_t = \sum_{j=1}^{t} x_j x_j^{\top}$. By conjugacy, the relaxation is equal to

$$\begin{aligned}&\sup_{f\in\mathscr{F}}\left\{2\sum_{j=1}^{t} y_j\langle f, x_j\rangle - f^{\top}\left(\tilde{\mathbf{B}}_t+\lambda I\right)f\right\} - \sum_{j=1}^{t} y_j^2 + 4B^2\log\left(\frac{(n/d)^d}{\Delta\left(\mathbf{B}_t+\lambda I\right)}\right) \\ &= -\inf_{f\in\mathscr{F}}\left\{\sum_{j=1}^{t}(f(x_j)-y_j)^2 + \lambda\|f\|_2^2\right\} + 4B^2\log\left(\frac{(n/d)^d}{\Delta\left(\mathbf{B}_t+\lambda I\right)}\right)\end{aligned}$$

We now prove admissibility of relaxation as follows:

$$\begin{aligned}&\sup_{p_t}\mathbb{E}_{y_t\sim p_t}\left[(y_t-\mathbb{E}[y_t])^2 + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)\right] \\ &\quad= \sup_{p_t}\mathbb{E}_{y_t\sim p_t}\left[(y_t-\mathbb{E}[y_t])^2 - \inf_{f\in\mathscr{F}}\left\{\sum_{j=1}^{t}\left(\langle f, x_j\rangle - y_j\right)^2 + \lambda\|f\|^2\right\}\right] + 4B^2\log\left(\frac{(n/d)^d}{\Delta\left(\mathbf{B}_t+\lambda I\right)}\right)\end{aligned}$$

The first term, in view of (23), is equal to

$$\begin{aligned}&\sup_{p_t}\mathbb{E}_{y_t\sim p_t}\left[\sup_{f\in\mathscr{F}}\left\{-\sum_{j=1}^{t-1}\left(\langle f, x_j\rangle - y_j\right)^2 - (\langle f, x_t\rangle - \mathbb{E}[y_t])^2 + 2(y_t-\mathbb{E}[y_t])(\langle f, x_t\rangle - \mathbb{E}[y_t]) + \lambda\|f\|^2\right\}\right] \\ &\le \sup_{\mu_t}\mathbb{E}_{\epsilon_t}\left[\sup_{f\in\mathscr{F}}\left\{-\sum_{j=1}^{t-1}\left(\langle f, x_j\rangle - y_j\right)^2 - (\langle f, x_t\rangle - \mu_t)^2 + 4B\epsilon_t\left(\langle f, x_t\rangle - \mu_t\right) + \lambda\|f\|^2\right\}\right]\end{aligned}$$

and the inequality arises from symmetrization exactly as in the proof of of Lemma 4. Once again, rewriting the above using conjugacy and converting to the z_t notation by appending a coordinate, the relaxation is upper bounded by

$$\begin{aligned}
&\sup_{z_t} \mathbb{E}_{\epsilon_t}\left[\left\|\sum_{j=1}^{t-1} y_j z_j + 2B\epsilon_t z_t\right\|^2_{\left(\mathbf{B}_{t-1}+z_t z_t^\top+\lambda I\right)^{-1}}\right] + 4B^2\log\left(\frac{(n/d)^d}{\Delta\left(\mathbf{B}_t+\lambda I\right)}\right) - \sum_{j=1}^{t-1} y_j^2 \\
&= \sup_{z_t}\left\|\sum_{j=1}^{t-1} y_j z_j\right\|^2_{\left(\mathbf{B}_{t-1}+z_t z_t^\top+\lambda I\right)^{-1}} + 4B^2 z_t^\top\left(\mathbf{B}_{t-1}+z_t z_t^\top+\lambda I\right)^{-1} z_t + 4B^2\log\left(\frac{(n/d)^d}{\Delta\left(\mathbf{B}_t+\lambda I\right)}\right) - \sum_{j=1}^{t-1} y_j^2
\end{aligned}$$

which is further upper bounded by

$$\begin{aligned}
&\sup_{z_t}\left\|\sum_{j=1}^{t-1} y_j z_j\right\|^2_{\left(\mathbf{B}_{t-1}+\lambda I\right)^{-1}} + 4B^2\log\left(\frac{\Delta\left(\mathbf{B}_t+\lambda I\right)}{\Delta\left(\mathbf{B}_{t-1}+\lambda I\right)}\right) + 4B^2\log\left(\frac{(n/d)^d}{\Delta\left(\mathbf{B}_t+\lambda I\right)}\right) - \sum_{j=1}^{t-1} y_j^2 \\
&= \left\|\sum_{j=1}^{t-1} y_j z_j\right\|^2_{\left(\mathbf{B}_{t-1}+\lambda I\right)^{-1}} + 4B^2\log\left(\frac{(n/d)^d}{\Delta\left(\mathbf{B}_{t-1}+\lambda I\right)}\right) - \sum_{j=1}^{t-1} y_j^2 \\
&= \mathbf{Rel}_n\left(x_{1:t-1}, y_{1:t-1}\right)
\end{aligned}$$

Thus we have shown admissibility and further this relaxation is such that $(\hat{y}-y_t)^2 + \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, y_t)\right)$ is a convex function of y_t and so the forecast associated with this relaxation is simply

$$\hat{y}_t = \mathrm{Clip}\left(\frac{\left\|\sum_{j=1}^{t-1} y_j x_j + B x_t\right\|^2_{\left(\mathbf{B}_t+\lambda I\right)^{-1}} - \left\|\sum_{j=1}^{t-1} y_j x_j - B x_t\right\|^2_{\left(\mathbf{B}_t+\lambda I\right)^{-1}}}{4B}\right)$$

Expanding out the two norm square terms we conclude that

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

Notice that this is exactly the clipped version of the Vovk-Azoury-Warmuth forecaster. The final regret bound we obtain is given by $\mathbf{Reg} \le \mathbf{Rel}_n(\cdot)$ and so we conclude that for any $f \in \mathscr{F}$, regret against this linear predictor is bounded as :

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

□