Online Learning: Random Averages, Combinatorial Parameters, and Learnability

Alexander Rakhlin
Department of Statistics
University of Pennsylvania

Karthik Sridharan
Toyota Technological Institute
at Chicago

Ambuj Tewari
Computer Science Department
University of Texas at Austin

October 29, 2010

Abstract

We study learnability in the online learning model. We define several complexity measures which capture the difficulty of learning in a sequential manner. Among these measures are analogues of Rademacher complexity, covering numbers and fat shattering dimension from statistical learning theory. Relationship among these complexity measures, their connection to online learning, and tools for bounding them are provided. In the setting of supervised learning, finiteness of the introduced scale-sensitive parameters is shown to be equivalent to learnability. The complexities we define also ensure uniform convergence for non-i.i.d. data, extending the uniform Glivenko-Cantelli type results. We conclude by showing online learnability for an array of examples.

1 Introduction

In the online learning framework, the learner is faced with a sequence of data appearing at discrete time intervals. In contrast to the classical "batch" learning scenario where the learner is being evaluated after the sequence is completely revealed, in the online framework the learner is evaluated at every round. Furthermore, in the batch scenario the data source is typically assumed to be *i.i.d.* with an unknown distribution, while in the online framework we relax or eliminate any stochastic assumptions on the data source. As such, the online learning problem can be phrased as a repeated two-player game between the learner (player) and the adversary (Nature).

Let $\mathcal{F}$ be a class of functions and $\mathcal{X}$ some set. The `Online Learning Model` is defined as the following T-round interaction between the learner and the adversary: On round $t = 1, \ldots, T$, the learner chooses $f_t \in \mathcal{F}$, the adversary picks $x_t \in \mathcal{X}$, and the learner suffers loss $f_t(x_t)$. At the end of T rounds we define *regret*

$$\mathbf{R}(f_{1:T}, x_{1:T}) = \sum_{t=1}^{T} f_t(x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} f(x_t)$$

as the difference between the cumulative loss of the player as compared to the cumulative loss of the best fixed comparator. For the given pair $(\mathcal{F}, \mathcal{X})$, the problem is said to be *online learnable* if there exists an algorithm for the learner such that regret grows sublinearly. Learnability is closely related to *Hannan consistency* [14, 10].

There has been a lot of interest in a particular setting of the online learning model, called *online convex optimization*. In this setting, we write $x_t(f_t)$ as the loss incurred by the learner, and the assumption is made that the function x_t is convex in its argument. The particular convexity structure enables the development of optimization-based algorithms for learner's choices. Learnability and precise rates of growth of regret have been shown in a number of recent papers (e.g. [41, 32, 5, 1]).

The online learning model also subsumes the *prediction* setting. In the latter, the learner's choice of a $\mathcal{Y}$-valued function g_t leads to the loss of $\ell(g_t(z_t), y_t)$ according to a fixed loss function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$. It is evident that the choice of the learner is equivalently written as $f_t(x) = \ell(g_t(z), y)$, and $x_t = (z_t, y_t)$ is the choice of the adversary. In Section 6 we discuss the prediction setting in more detail.

In the "batch" learning scenario, data $\{(x_i, y_i)\}_{i=1}^T$ is presented as an i.i.d. draw from a fixed distribution over some product $\mathcal{X} \times \mathcal{Y}$. Learnability results have been extensively studied in the PAC framework [36] and its agnostic extensions [15, 19]. It is well-known that learnability in the binary case (that is, $\mathcal{Y} = \{-1, +1\}$) is completely characterized by finiteness of the Vapnik-Chervonenkis combinatorial dimension of the function class [40, 38]. In the real-valued case, a number of combinatorial quantities have been proposed: P-dimension [28], V-dimension [39], as well as the *scale-sensitive* versions P_γ-dimension [19, 6] and V_γ-dimension [3]. The last two dimensions were shown to be characterizing learnability [3] and uniform convergence of means to expectations for function classes.

In contrast to the classical learning setting, there has been surprisingly little work on characterizing learnability for the online learning framework. Littlestone [24] has shown that, in the setting of prediction of binary outcomes, a certain combinatorial property of the binary-valued function class characterizes learnability in the realizable case (that is, when the outcomes presented by the adversary are given according to some function in the class $\mathcal{F}$). The result has been extended to the non-realizable case by Shai Ben-David, Dávid Pál and Shai Shalev-Shwartz [8] who named this combinatorial quantity the *Littlestone's dimension*. Coincident with [8], *minimax* analysis of online convex optimization yielded new insights into the value of the game, its minimax dual representation, as well as algorithm-independent upper and lower bounds [1, 34]. In this paper, we build upon these results and the findings of [8] to develop a theory of online learning.

We show that in the online learning model, a notion which we call *Sequential Rademacher complexity* allows us to easily prove learnability for a vast array of problems. The role of this complexity is similar to the role of the Rademacher complexity in statistical learning theory. Next, we extend Littlestone's dimension to the real-valued case. We show that finiteness of this scale-sensitive version, which we call the *fat-shattering dimension*, is necessary and sufficient for learnability in the prediction setting. Extending the binary-valued result of [8], we introduce a generic algorithm which plays the role similar to that of empirical risk minimization for i.i.d. data: if the problem is learnable in the supervised setting, then it is learnable by this algorithm. Along the way we develop analogues of Massart's finite class lemma, the Dudley integral upper bound on the Sequential Rademacher complexity, appropriately defined packing and covering numbers, and even an analogue of the Sauer-Shelah combinatorial lemma. We also introduce a generalization of the uniform law of large numbers for non-i.i.d. distributions and show that finiteness of the fat-shattering dimension implies this convergence.

Many of the results come with more work than their counterparts in statistical learning theory. In particular, instead of training sets we have to work with trees, making the results somewhat involved. While the spirit of the online theory is that it provides a "temporal" generalization of the "batch" learning problem, not all the results from statistical learning theory transfer to our setting. For instance, two distinct notions of a packing set exist for trees, and these notions can be seen to coincide in "batch" learning. The fact that many notions of statistical learning theory can be extended to the online learning model is indeed remarkable.

2 Preliminaries

By phrasing the online learning model as a repeated game and considering its minimax value, we naturally arrive at an important object in combinatorial game theory: trees. Unless specified, all trees considered in this paper are *rooted binary* trees with equal-depth paths from the root to the leaves. While it is useful to have the tree picture in mind when reading the paper, it is also necessary to precisely define trees as mathematical objects. We opt for the following definition.

Definition 1 (Trees)**.** Given some set $\mathcal{Z}$, a *$\mathcal{Z}$-valued tree of depth T* is a sequence $(\mathbf{z}_1, \ldots, \mathbf{z}_T)$ of T mappings $\mathbf{z}_i : \{\pm 1\}^{i-1} \mapsto \mathcal{Z}$. The *root* of the tree $\mathbf{z}$ is the constant function $\mathbf{z}_1 \in \mathcal{Z}$.

Armed with this definition, we can talk about various operations on trees. For a function $f : \mathcal{Z} \mapsto \mathcal{U}$, $f(\mathbf{x})$ denotes the U-valued tree defined by the mappings $(f \circ \mathbf{x}_1, \ldots, f \circ \mathbf{x}_T)$. Analogously, for $f : \mathcal{Z} \times \mathcal{Z} \mapsto \mathcal{U}$, the $\mathcal{U}$-valued tree $f(\mathbf{x}, \mathbf{x}')$ is defined as mappings $(f(\mathbf{x}_1, \mathbf{x}'_1), \ldots, f(\mathbf{x}_T, \mathbf{x}'_T))$. In particular, this defines the usual binary arithmetic operations on real-valued trees. Furthermore, for a class of functions $\mathcal{F}$ and a tree $\mathbf{x}$, the projection of $\mathcal{F}$ onto $\mathbf{x}$ is $\mathcal{F}(\mathbf{x}) = \{f(\mathbf{x}) : f \in \mathcal{F}\}$.

Definition 2 (Path)**.** A *path* of length T is a sequence $\epsilon = (\epsilon_1, \ldots, \epsilon_{T-1}) \in \{\pm 1\}^{T-1}$.

We shall abuse notation by referring to $\mathbf{x}_i(\epsilon_1, \ldots, \epsilon_{i-1})$ by $\mathbf{x}_i(\epsilon)$. Clearly $\mathbf{x}_i$ only depends on the first $i-1$ elements of ϵ. We will also refer to $\epsilon = (\epsilon_1, \ldots, \epsilon_T) \in \{\pm 1\}^T$ as a path in a tree of depth T even though the value of ϵ_T is inconsequential. Next we define the notion of subtrees.

Definition 3 (Subtrees)**.** The *left subtree* $\mathbf{z}^\ell$ of $\mathbf{z}$ at the root is defined as $T-1$ mappings $(\mathbf{z}^\ell_1, \ldots, \mathbf{z}^\ell_{T-1})$ with $\mathbf{z}^\ell_i(\epsilon) = \mathbf{z}_{i+1}(\{-1\} \times \epsilon)$ for $\epsilon \in \{\pm 1\}^{T-1}$. The right subtree $\mathbf{z}^r$ is defined analogously by conditioning on the first coordinate of $\mathbf{z}_{i+1}$ to be $+1$.

Given two subtrees $\mathbf{z}$, $\mathbf{v}$ of the same depth $T-1$ and a constant mapping $\mathbf{w}_1$, we can *join* the two subtrees to obtain a new set of mappings $(\mathbf{w}_1, \ldots, \mathbf{w}_T)$ as follows. The root is the constant mapping $\mathbf{w}_1$. For $i \in \{2, \ldots, T\}$ and $\epsilon \in \{\pm 1\}^T$, $\mathbf{w}_i(\epsilon) = \mathbf{z}_{i-1}(\epsilon)$ if $\epsilon_1 = -1$ and $\mathbf{w}_i(\epsilon) = \mathbf{v}_{i-1}(\epsilon)$ if $\epsilon_1 = +1$.

In the sequel, we will need to talk about the values given by the tree $\mathbf{x}$ over all the paths. Formally, let $\mathrm{Img}(\mathbf{x}) = \mathbf{x}\left(\{\pm 1\}^T\right) = \{\mathbf{x}_t(\epsilon) : t \in [T], \epsilon \in \{\pm 1\}^T\}$ be the image of the mappings of $\mathbf{x}$.

Let us also introduce some notation not related to trees. We denote a sequence of the form $(y_a, \ldots, y_b)$, where $a \leq b$, by simply writing $y_{a:b}$. The set of all functions from $\mathcal{X}$ to $\mathcal{Y}$ is denoted by $\mathcal{Y}^{\mathcal{X}}$, and the t-fold product $\mathcal{X} \times \ldots \times \mathcal{X}$ is denoted by $\mathcal{X}^t$. For any $T \in \mathbb{N}$, $[T]$ denotes the set $\{1, \ldots, T\}$. A conditional distribution is written as $\mathbb{E}_t\left[A\right] = \mathbb{E}\left[A | \mathcal{G}_t\right]$, where $\mathcal{G}_t$ is an appropriate filtration which will be specified. Whenever a supremum (infimum) is written in the form $\sup_a$ without a being quantified, it is assumed that a ranges over the set of all possible values which will be understood from the context.

For the sake of readability, almost all the proofs are deferred to the appendix.

3 Value of the Game

Fix the sets $\mathcal{F}$ and $\mathcal{X}$ and consider the online learning model stated in the introduction. We assume that $\mathcal{F}$ is a subset of a separable metric space. Let $\mathcal{Q}$ be the set of probability distributions on $\mathcal{F}$. Assume that $\mathcal{Q}$ is weakly compact. We consider randomized learners who predict a distribution $q_t \in \mathcal{Q}$ on every round. Formally, define a learner's strategy π as a sequence of mappings $\pi_t : \mathcal{X}^{t-1} \times \mathcal{F}^{t-1} \mapsto \mathcal{Q}$ for each $t \in [T]$. We define the value of the game as

$$\mathcal{V}_T(\mathcal{F}, \mathcal{X}) = \inf_{q_1 \in \mathcal{Q}} \sup_{x_1 \in \mathcal{X}} \mathbb{E}_{f_1 \sim q_1} \cdots \inf_{q_T \in \mathcal{Q}} \sup_{x_T \in \mathcal{X}} \mathbb{E}_{f_T \sim q_T} \left[\sum_{t=1}^T f_t(x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^T f(x_t) \right] \tag{1}$$

where f_t has distribution q_t. We consider here the *adaptive* adversary who gets to choose each x_t based on the history of moves $f_{1:t-1}$ and $x_{1:t-1}$.

Note that our assumption that $\mathcal{F}$ is a subset of a separable metric space implies that $\mathcal{Q}$ is tight and Prokhorov's theorem states that compactness of $\mathcal{Q}$ under weak topology is equivalent to tightness [35]. Compactness under weak topology allows us to appeal to Theorem 1 stated below, which is adapted for our needs from [1].

Theorem 1. *Let $\mathcal{F}$ and $\mathcal{X}$ be the sets of moves for the two players, satisfying the necessary conditions for the minimax theorem to hold. Denote by $\mathcal{Q}$ and $\mathcal{P}$ the sets of probability distributions (mixed strategies) on $\mathcal{F}$ and $\mathcal{X}$, respectively. Then*

$$\begin{aligned}\mathcal{V}_T(\mathcal{F},\mathcal{X}) &= \inf_{q_1\in\mathcal{Q}}\sup_{x_1\in\mathcal{X}}\mathbb{E}_{f_1\sim q_1}\dots\inf_{q_T\in\mathcal{Q}}\sup_{x_T\in\mathcal{X}}\mathbb{E}_{f_T\sim q_T}\left[\sum_{t=1}^T f_t(x_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^T f(x_t)\right]\\ &= \sup_{p_1}\mathbb{E}_{x_1\sim p_1}\dots\sup_{p_1}\mathbb{E}_{x_1\sim p_1}\left[\sum_{t=1}^T\inf_{f_t\in\mathcal{F}}\mathbb{E}_{x_t\sim p_t}\left[f_t(x_t)\right]-\inf_{f\in\mathcal{F}}\sum_{t=1}^T f(x_t)\right]. \qquad (2)\end{aligned}$$

The question of learnability in the online learning model is now reduced to the study of $\mathcal{V}_T(\mathcal{F},\mathcal{X})$, taking Eq. (2) as the starting point. In particular, under our definition, showing that the value grows sublinearly with T is equivalent to showing learnability.

Definition 4. A class $\mathcal{F}$ is said to be *online learnable* with respect to the given $\mathcal{X}$ if

$$\limsup_{T\to\infty}\frac{\mathcal{V}_T(\mathcal{F},\mathcal{X})}{T}=0\ .$$

The rest of the paper is aimed at understanding the value of the game $\mathcal{V}_T(\mathcal{F},\mathcal{X})$ for various function classes $\mathcal{F}$. Since complexity of $\mathcal{F}$ is the focus of the paper, we shall often write $\mathcal{V}_T(\mathcal{F})$, and the dependence on $\mathcal{X}$ will be implicit.

One of the key notions introduced in this paper is the complexity which we term *Sequential Rademacher complexity.* A natural generalization of Rademacher complexity [21, 7, 26], the sequential analogue possesses many of the nice properties of its classical cousin. The properties are proved in Section 7 and then used to show learnability for many of the examples in Section 8. The first step, however, is to show that Sequential Rademacher complexity upper bounds the value of the game. This is the subject of the next section.

4 Random Averages

We propose the following definition. The key difference from the classical notion is the dependence of the sequence of data on the sequence of signs (Rademacher random variables). As shown in the sequel, this dependence captures the sequential nature of the problem.

Definition 5. The *Sequential Rademacher Complexity* of a function class $\mathcal{F}\subseteq\mathbb{R}^{\mathcal{X}}$ is defined as

$$\mathfrak{R}_T(\mathcal{F})=\sup_{\mathbf{x}}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^T\epsilon_t f(\mathbf{x}_t(\epsilon))\right]$$

where the outer supremum is taken over all $\mathcal{X}$-valued trees of depth T and $\epsilon=(\epsilon_1,\dots,\epsilon_T)$ is a sequence of i.i.d. Rademacher random variables.

In statistical learning, Rademacher complexity is shown to control uniform deviations of means and expectations, and this control is key for learnability in the "batch" setting. We now show that Sequential Rademacher complexity upper-bounds the value of the game, suggesting its importance for online learning (see Section 6 for a lower bound).

Theorem 2. *The minimax value of a randomized game is bounded as*

$$\mathcal{V}_T(\mathcal{F}) \le 2\mathfrak{R}_T(\mathcal{F})$$

Proof. From Eq. (2),

$$\begin{aligned}
\mathcal{V}_T(\mathcal{F}) &= \sup_{p_1} \mathbb{E}_{x_1\sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T\sim p_T} \left[\sum_{t=1}^{T} \inf_{f_t\in\mathcal{F}} \mathbb{E}_{x_t\sim p_t}\left[f_t(x_t)\right] - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} f(x_t) \right] && (3)\\
&= \sup_{p_1} \mathbb{E}_{x_1\sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T\sim p_T} \left[\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{T} \inf_{f_t\in\mathcal{F}} \mathbb{E}_{x_t\sim p_t}\left[f_t(x_t)\right] - \sum_{t=1}^{T} f(x_t) \right\} \right]\\
&\le \sup_{p_1} \mathbb{E}_{x_1\sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T\sim p_T} \left[\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{T} \mathbb{E}_{x_t\sim p_t}\left[f(x_t)\right] - \sum_{t=1}^{T} f(x_t) \right\} \right] && (4)
\end{aligned}$$

The last step, in fact, is the first time we deviated from keeping equalities. The upper bound is obtained by replacing each infimum by a particular choice f. Now renaming variables we have,

$$\begin{aligned}
\mathcal{V}_T(\mathcal{F}) &= \sup_{p_1} \mathbb{E}_{x_1\sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T\sim p_T} \left[\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{T} \mathbb{E}_{x'_t\sim p_t}\left[f(x'_t)\right] - \sum_{t=1}^{T} f(x_t) \right\} \right]\\
&\le \sup_{p_1} \mathbb{E}_{x_1\sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T\sim p_T} \left[\mathbb{E}_{x'_1\sim p_1} \dots \mathbb{E}_{x'_T\sim p_T} \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{T} f(x'_t) - \sum_{t=1}^{T} f(x_t) \right\} \right]\\
&\le \sup_{p_1} \mathbb{E}_{x_1,x'_1\sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T,x'_T\sim p_T} \left[\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{T} f(x'_t) - \sum_{t=1}^{T} f(x_t) \right\} \right] .
\end{aligned}$$

where the last two steps are using Jensen inequality for the supremum.

By the Key Technical Lemma (see Lemma 3 below) with $\phi(u) = u$ and $\Delta_f(x_t, x'_t) = f(x'_t) - f(x_t)$,

$$\begin{aligned}
&\sup_{p_1} \mathbb{E}_{x_1,x'_1\sim p_1} \dots \sup_{p_T} \mathbb{E}_{x_T,x'_T\sim p_T} \left[\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{T} f(x'_t) - f(x_t) \right\} \right]\\
&\qquad\qquad \le \sup_{x_1,x'_1} \left\{ \mathbb{E}_{\epsilon_1} \left[\dots \sup_{x_T,x'_T} \left\{ \mathbb{E}_{\epsilon_T} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \epsilon_t \left(f(x'_t) - f(x_t)\right) \right] \right\} \dots \right] \right\}
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathcal{V}_T(\mathcal{F}) &\le \sup_{x_1,x'_1} \left\{ \mathbb{E}_{\epsilon_1} \left[\dots \sup_{x_T,x'_T} \left\{ \mathbb{E}_{\epsilon_T} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \epsilon_t \left(f(x'_t) - f(x_t)\right) \right] \right\} \dots \right] \right\}\\
&\le \sup_{x_1,x'_1} \left\{ \mathbb{E}_{\epsilon_1} \left[\dots \sup_{x_T,x'_T} \left\{ \mathbb{E}_{\epsilon_T} \left[\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{T} \epsilon_t f(x'_t) \right\} + \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{T} -\epsilon_t f(x_t) \right\} \right] \right\} \dots \right] \right\}\\
&= 2 \sup_{x_1} \left\{ \mathbb{E}_{\epsilon_1} \left[\dots \sup_{x_T} \left\{ \mathbb{E}_{\epsilon_T} \left[\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{T} \epsilon_t f(x_t) \right\} \right] \right\} \dots \right] \right\}
\end{aligned}$$

Now, we need to move the suprema over x_t's outside. This is achieved via an idea similar to skolemization in logic. We basically exploit the identity

$$\mathbb{E}_{\epsilon_{1:t-1}}\left[\sup_{x_t} G(\epsilon_{1:t-1}, x_t)\right] = \sup_{\mathbf{x}_t} \mathbb{E}_{\epsilon_{1:t-1}}\left[G(\epsilon_{1:t-1}, \mathbf{x}_t(\epsilon_{1:t-1}))\right]$$

that holds for any $G : \{\pm 1\}^{t-1} \times \mathcal{X} \mapsto \mathbb{R}$. On the right the supremum is over functions $\mathbf{x}_t : \{\pm 1\}^{t-1} \to \mathcal{X}$. Using this identity once, we get,

$$\mathcal{V}_T(\mathcal{F}) \leq 2 \sup_{x_1, \mathbf{x}_2} \left\{ \mathbb{E}_{\epsilon_1, \epsilon_2}\left[\sup_{x_3} \ldots \sup_{x_T} \left\{ \mathbb{E}_{\epsilon_T}\left[\sup_{f\in\mathcal{F}} \left\{ \epsilon_1 f(x_1) + \epsilon_2 f(\mathbf{x}_2(\epsilon_1)) + \sum_{t=3}^{T} \epsilon_t f(x_t) \right\}\right]\right\} \ldots \right]\right\}$$

Now, use the identity $T-2$ more times to successively move the supremums over $x_3, \ldots, x_T$ outside, to get

$$\begin{aligned}\mathcal{V}_T(\mathcal{F}) &\leq 2 \sup_{x_1, \mathbf{x}_2, \ldots, \mathbf{x}_T} \mathbb{E}_{\epsilon_1, \ldots, \epsilon_T}\left[\sup_{f\in\mathcal{F}} \left\{ \epsilon_1 f(x_1) + \sum_{t=2}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1})) \right\}\right] \\ &= 2 \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_1, \ldots, \epsilon_T}\left[\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right\}\right]\end{aligned}$$

where the last supremum is over $\mathcal{X}$-valued trees of depth T. Thus we have proved the required statement. □

Theorem 2 relies on the following technical lemma, which will be used again in Section 5.3. Its proof requires considerably more work than the classical symmetrization proof [12, 26] due to the non-i.i.d. nature of the sequences.

Lemma 3 (Key Technical Lemma). *Let $(x_1, \ldots, x_T) \in \mathcal{X}^T$ be a sequence distributed according to $\mathbf{D}$ and let $(x'_1, \ldots, x'_T) \in \mathcal{X}^T$ be a tangent sequence. Let $\Delta_f(x_t, x'_t)$ be a functional $\mathcal{F} \mapsto \mathbb{R}$ such that*

$$\Delta_f(x_t, x'_t) = -\Delta_f(x'_t, x_t)$$

Let $\Phi(\Omega) = \phi\left(\sup_{f\in\mathcal{F}} \Omega(f)\right)$ or $\Phi(\Omega) = \phi\left(\sup_{f\in\mathcal{F}} |\Omega(f)|\right)$, where $\phi : \mathbb{R} \mapsto \mathbb{R}$ is some measurable real valued function and $\Omega : \mathcal{F} \mapsto \mathbb{R}$. Then

$$\sup_{p_1} \mathbb{E}_{x_1, x'_1 \sim p_1} \ldots \sup_{p_T} \mathbb{E}_{x_T, x'_T \sim p_T}\left[\Phi\left(\sum_{t=1}^{T} \Delta_f(x_t, x'_t)\right)\right] \leq \sup_{x_1, x'_1} \left\{ \mathbb{E}_{\epsilon_1}\left[\ldots \sup_{x_T, x'_T} \left\{ \mathbb{E}_{\epsilon_T}\left[\Phi\left(\sum_{t=1}^{T} \epsilon_t \Delta_f(x_t, x'_t)\right)\right]\right\} \ldots \right]\right\}$$

where $\epsilon_1, \ldots, \epsilon_T$ are independent (of each other and everything else) Rademacher random variables.

5 Covering Numbers and Combinatorial Parameters

In statistical learning theory, learnability for binary classes of functions is characterized by the Vapnik-Chervonenkis combinatorial dimension [40]. For real-valued function classes, the corresponding notions are the scale-sensitive dimensions, such as P_γ [3, 6]. For online learning, the notion characterizing learnability for binary prediction in the realizable case has been introduced by Littlestone [24] and extended to the non-realizable case of binary prediction by Shai Ben-David, Dávid Pál and Shai Shalev-Shwartz [8]. Next, we define the Littlestone's dimension [24, 8] and propose its scale-sensitive versions for real-valued function classes. In the sequel, these combinatorial parameters are shown to control the growth of covering numbers on trees. In the setting of prediction, the combinatorial parameters are shown to exactly characterize learnability (see Section 6).

Definition 6. An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is *shattered* by a function class $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$ if for all $\epsilon \in \{\pm 1\}^d$, there exists $f \in \mathcal{F}$ such that $f(\mathbf{x}_t(\epsilon)) = \epsilon_t$ for all $t \in [d]$. The *Littlestone dimension* $\mathrm{Ldim}(\mathcal{F}, \mathcal{X})$ is the largest d such that $\mathcal{F}$ shatters an $\mathcal{X}$-valued tree of depth d.

Definition 7. An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is *α-shattered* by a function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$, if there exists an $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d such that

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

The tree $\mathbf{s}$ is called the *witness to shattering*. The *fat-shattering dimension* $\mathrm{fat}_\alpha(\mathcal{F}, \mathcal{X})$ at scale α is the largest d such that $\mathcal{F}$ α-shatters an $\mathcal{X}$-valued tree of depth d.

With these definitions it is easy to see that $\mathrm{fat}_\alpha(\mathcal{F}, \mathcal{X}) = \mathrm{Ldim}(\mathcal{F}, \mathcal{X})$ for a binary-valued function class $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$ for any $0 < \alpha \leq 2$.

When $\mathcal{X}$ and/or $\mathcal{F}$ is understood from the context, we will simply write fat_α or $\mathrm{fat}_\alpha(\mathcal{F})$ instead of $\mathrm{fat}_\alpha(\mathcal{F}, \mathcal{X})$. Furthermore, we will write $\mathrm{fat}_\alpha(\mathcal{F}, \mathbf{x})$ for $\mathrm{fat}_\alpha(\mathcal{F}, \mathrm{Img}(\mathbf{x}))$. In other words, $\mathrm{fat}_\alpha(\mathcal{F}, \mathbf{x})$ is the largest d such that $\mathcal{F}$ α-shatters a tree $\mathbf{z}$ of depth d with $\mathrm{Img}(\mathbf{z}) \subseteq \mathrm{Img}(\mathbf{x})$.

Let us mention that if trees $\mathbf{x}$ are defined by constant mappings $\mathbf{x}_t(\epsilon) = x_t$, the combinatorial parameters coincide with the Vapnik-Chervonenkis dimension and with the scale-sensitive dimension P_γ. Therefore, the notions we are studying are strict "temporal" generalizations of the VC theory.

As in statistical learning theory, the combinatorial parameters are only useful if they can be shown to capture that aspect of $\mathcal{F}$ which is important for learnability. In particular, a "size" of a function class is known to be related to complexity of learning from i.i.d. data., and the classical way to measure "size" is through a cover or of a packing set. We propose the following definitions for online learning.

Definition 8. A set V of $\mathbb{R}$-valued trees of depth T is *an α-cover* (with respect to ℓ_p-norm) of $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ on a tree $\mathbf{x}$ of depth T if

$$\forall f \in \mathcal{F}, \ \forall \epsilon \in \{\pm 1\}^T \ \exists \mathbf{v} \in V \text{ s.t.} \qquad \left(\frac{1}{T} \sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - f(\mathbf{x}_t(\epsilon))|^p \right)^{1/p} \leq \alpha$$

The *covering number* of a function class $\mathcal{F}$ on a given tree $\mathbf{x}$ is defined as

$$\mathcal{N}_p(\alpha, \mathcal{F}, \mathbf{x}) = \min\{|V| : V \text{ is an } \alpha - \text{cover w.r.t. } \ell_p\text{-norm of } \mathcal{F} \text{ on } \mathbf{x}\}.$$

Further define $\mathcal{N}_p(\alpha, \mathcal{F}, T) = \sup_{\mathbf{x}} \mathcal{N}_p(\alpha, \mathcal{F}, \mathbf{x})$, the maximal ℓ_p covering number of $\mathcal{F}$ over depth T trees.

In particular, a set V of $\mathbb{R}$-valued trees of depth T is a 0-*cover* of $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ on a tree $\mathbf{x}$ of depth T if

$$\forall f \in \mathcal{F}, \ \forall \epsilon \in \{\pm 1\}^T \ \exists \mathbf{v} \in V \text{ s.t.} \quad \mathbf{v}_t(\epsilon) = f(\mathbf{x}_t(\epsilon))$$

We denote by $\mathcal{N}(0, \mathcal{F}, \mathbf{x})$ the size of a smallest 0-cover on $\mathbf{x}$ and $\mathcal{N}(0, \mathcal{F}, T) = \sup_{\mathbf{x}} \mathcal{N}(0, \mathcal{F}, \mathbf{x})$.

Let us discuss a subtle point. The 0-cover should not be mistaken for the size $|\mathcal{F}(\mathbf{x})|$ of the projection of $\mathcal{F}$ onto the tree $\mathbf{x}$, and the same care should be taken when dealing with α-covers. Let us illustrate this with an example. Consider a tree $\mathbf{x}$ of depth T and suppose for simplicity that $|\mathrm{Img}(\mathbf{x})| = 2^T - 1$, i.e. the values of $\mathbf{x}$ are all distinct. Suppose $\mathcal{F}$ consists of 2^{T-1} binary-valued functions defined as zero on all of $\mathrm{Img}(\mathbf{x})$ except for a single value of $\mathrm{Img}(\mathbf{x}_T)$. In plain words, each function is zero everywhere on the tree except for a single leaf. While the projection $\mathcal{F}(\mathbf{x})$ has 2^{T-1} distinct trees, the size of a 0-cover is only 2. It is enough to take an all-zero function g_0 along with a function g_1 which is zero on all of $\mathrm{Img}(\mathbf{x})$ except $\mathrm{Img}(\mathbf{x}_T)$ (i.e. on the leaves). It is easy to verify that $g_0(\mathbf{x})$ and $g_1(\mathbf{x})$ provide a 0-cover for $\mathcal{F}$ on $\mathbf{x}$, and therefore, unlike $|\mathcal{F}(\mathbf{x})|$, the size of the cover does not grow with T. The example is encouraging: our definition of a cover captures the fact that the function class is "simple" for any given path.

Next, we naturally propose a definition of a packing.

Definition 9. A set V of $\mathbb{R}$-valued trees of depth T is said to be α-separated if

$$\forall \mathbf{v} \in V,\ \exists \epsilon \in \{\pm 1\}^T \text{ s.t. } \ \forall \mathbf{w} \in V \setminus \{\mathbf{v}\} \qquad \left(\frac{1}{T}\sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - \mathbf{w}_t(\epsilon)|^p\right)^{1/p} > \alpha$$

The *packing number* $\mathcal{D}_p(\alpha, \mathcal{F}, \mathbf{x})$ of a function class $\mathcal{F}$ on a given tree $\mathbf{x}$ is the size of the largest α-separated subset of $\{f(\mathbf{x}) : f \in \mathcal{F}\}$.

Definition 10. A set V of $\mathbb{R}$-valued trees of depth T is said to be *strongly* α-separated if

$$\exists \epsilon \in \{\pm 1\}^T \text{ s.t. } \ \forall \mathbf{v}, \mathbf{w} \in V, \mathbf{v} \neq \mathbf{w} \qquad \left(\frac{1}{T}\sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - \mathbf{w}_t(\epsilon)|^p\right)^{1/p} > \alpha$$

The *strong packing number* $\mathcal{M}_p(\alpha, \mathcal{F}, \mathbf{x})$ of a function class $\mathcal{F}$ on a given tree $\mathbf{x}$ is the size of the largest strongly α-separated subset of $\{f(\mathbf{x}) : f \in \mathcal{F}\}$.

Note the distinction between the packing number and the strong packing number. For the former, it must be that every member of the packing is α-separated from every other member on *some* path. For the latter, there must be a path on which every member of the packing is α-separated from every other member. This distinction does not arise in the classical scenario of "batch" learning. We observe that if a tree $\mathbf{x}$ is defined by constant mappings $\mathbf{x}_t = x_t$, the two notions of packing and strong packing coincide, i.e. $\mathcal{D}_p(\alpha, \mathcal{F}, \mathbf{x}) = \mathcal{M}_p(\alpha, \mathcal{F}, \mathbf{x})$. The following lemma gives a relationship between covering numbers and the two notions of packing numbers. The form of this should be familiar, except for the distinction between the two types of packing numbers.

Lemma 4. *For any $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$, any $\mathcal{X}$-valued tree $\mathbf{x}$ of depth T, and any $\alpha > 0$*

$$\mathcal{M}_p(2\alpha, \mathcal{F}, \mathbf{x}) \le \mathcal{N}_p(\alpha, \mathcal{F}, \mathbf{x}) \le \mathcal{D}_p(\alpha, \mathcal{F}, \mathbf{x}).$$

It is important to note that the gap between the two types of packing can be as much as 2^T.

5.1 A Combinatorial Upper Bound

We now relate the combinatorial parameters introduced in the previous section to the size of a cover. In the binary case ($k = 1$ below), a reader might notice a similarity of Theorems 5 and 7 to the classical results due to Sauer [29], Shelah [33] (also, Perles and Shelah), and Vapnik and Chervonenkis [40]. There are several approaches to proving what is often called the Sauer-Shelah lemma. We opt for the inductive-style proof (e.g. Alon and Spencer [4]). Dealing with trees, however, requires more work than in the VC case.

Theorem 5. *Let $\mathcal{F} \subseteq \{0, \dots, k\}^{\mathcal{X}}$ be a class of functions with* $\mathrm{fat}_2(\mathcal{F}) = d$. *Then*

$$\mathcal{N}_\infty(1/2, \mathcal{F}, T) \le \sum_{i=0}^{d} \binom{T}{i} k^i \le (ekT)^d.$$

Furthermore, for $T \ge d$

$$\sum_{i=0}^{d} \binom{T}{i} k^i \le \left(\frac{ekT}{d}\right)^d.$$

Armed with Theorem 5, we can approach the problem of bounding the size of a cover at an α scale by a discretization trick. For the classical case of a cover based on a set points, the discretization idea appears in

[3, 27]. When passing from the combinatorial result to the cover at scale α in Corollary 6, it is crucial that Theorem 5 is in terms of $\mathrm{fat}_2(\mathcal{F})$ and not $\mathrm{fat}_1(\mathcal{F})$. This point can be seen in the proof of Corollary 6 (also see [27]): the discretization process can assign almost identical function values to discrete values which differ by 1. This explains why the combinatorial result of Theorem 5 is proved for the 2-shattering dimension.

We now show that the covering numbers are bounded in terms of the fat-shattering dimension.

Corollary 6. *Suppose $\mathcal{F}$ is a class of $[-1,1]$-valued functions on $\mathcal{X}$. Then for any $\alpha > 0$, any $T > 0$, and any $\mathcal{X}$-valued tree $\mathbf{x}$ of depth T,*

$$\mathcal{N}_1(\alpha, \mathcal{F}, \mathbf{x}) \le \mathcal{N}_2(\alpha, \mathcal{F}, \mathbf{x}) \le \mathcal{N}_\infty(\alpha, \mathcal{F}, \mathbf{x}) \le \left(\frac{2eT}{\alpha}\right)^{\mathrm{fat}_\alpha(\mathcal{F})}$$

With a proof similar to Theorem 5, a bound on the 0-cover can be proved in terms of the $\mathrm{fat}_1(\mathcal{F})$ combinatorial parameter. Of particular interest is the case $k = 1$, when $\mathrm{fat}_1(\mathcal{F}) = \mathrm{Ldim}(\mathcal{F})$.

Theorem 7. *Let $\mathcal{F} \subseteq \{0, \ldots, k\}^{\mathcal{X}}$ be a class of functions with $\mathrm{fat}_1(\mathcal{F}) = d$. Then*

$$\mathcal{N}(0, \mathcal{F}, T) \le \sum_{i=0}^{d} \binom{T}{i} k^i \le (ekT)^d .$$

Furthermore, for $T \ge d$

$$\sum_{i=0}^{d} \binom{T}{i} k^i \le \left(\frac{ekT}{d}\right)^d .$$

In particular, the result holds for binary-valued function classes ($k = 1$), in which case $\mathrm{fat}_1(\mathcal{F}) = \mathrm{Ldim}(\mathcal{F})$.

When bounding deviations of means from expectations uniformly over the function class, the usual approach proceeds by a symmetrization argument [13] followed by passing to a cover of the function class and a union bound (e.g. [26]). Alternatively, a more refined *chaining* analysis integrates over covering at different scales (e.g. [37]). By following the same path, we are able to prove a number of similar results for our setting. In the next section we present a bound similar to Massart's finite class lemma [25, Lemma 5.2], and in the following section this result will be used when integrating over different scales for the cover.

5.2 Finite Class Lemma and the Chaining Method

Lemma 8. *For any finite set V of $\mathbb{R}$-valued trees of depth T we have that*

$$\mathbb{E}_\epsilon \left[\max_{\mathbf{v} \in V} \sum_{t=1}^{T} \epsilon_t \mathbf{v}_t(\epsilon)\right] \le \sqrt{2 \log(|V|) \max_{\mathbf{v} \in V} \max_{\epsilon \in \{\pm 1\}^T} \sum_{t=1}^{T} \mathbf{v}_t(\epsilon)^2}$$

A simple consequence of the above lemma is that if $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ is a finite class, then for any given tree $\mathbf{x}$ we have that

$$\mathbb{E}_\epsilon \left[\max_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon))\right] \le \mathbb{E}_\epsilon \left[\max_{\mathbf{v} \in \mathcal{F}(\mathbf{x})} \sum_{t=1}^{T} \epsilon_t \mathbf{v}_t(\epsilon)\right] \le \sqrt{2T \log(|\mathcal{F}|)} .$$

Note that if $f \in \mathcal{F}$ is associated with an "expert", this result combined with Theorem 2 yields a bound given by the exponential weighted average forecaster algorithm (see [10]). In Section 8 we discuss this case in more detail. However, as we show next, Lemma 8 goes well beyond just finite classes and can be used to get an analog of Dudley entropy bound [11] for the online setting through a chaining argument.

Definition 11. The *Integrated complexity* of a function class $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ is defined as

$$\mathfrak{D}_T(\mathcal{F}) = \inf_{\alpha} \left\{ 4T\alpha + 12 \int_{\alpha}^{1} \sqrt{T \ \log \ \mathcal{N}_2(\delta, \mathcal{F}, T)} \, d\delta \right\} .$$

To prove the next theorem, we consider covers of the class $\mathcal{F}$ at different scales that form a geometric progression. We zoom into a given function $f \in \mathcal{F}$ using covering elements at successive scales. This zooming in procedure is visualized as forming a chain that consists of links connecting elements of covers at successive scales. The Rademacher complexity of $\mathcal{F}$ can then be bounded by controlling the Rademacher complexity of the link classes, i.e. the class consisting of differences of functions from covers at neighbouring scales. This last part of the argument is the place where our proof becomes a bit more involved than the classical case.

Theorem 9. *For any function class $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$,*

$$\mathfrak{R}_T(\mathcal{F}) \leq \mathfrak{D}_T(\mathcal{F})$$

5.3 Finite Fat-shattering Dimension Implies Uniform Convergence

In the statistical setting it can be shown that learnability of supervised learning problem is equivalent to the so called uniform Glivenko-Cantelli property of the class which says that empirical averages converge to expected value of the function for any fixed distribution (samples drawn i.i.d.) and uniformly over the function class almost surely. We define below an analogous property for dependent distributions which requires that uniformly over function class the average value of the function converges to average conditional expectation of the function values almost surely.

Definition 12. A function class $\mathcal{F}$ satisfies a *Universal Uniform Convergence* if for all $\alpha > 0$,

$$\lim_{n \to \infty} \sup_{\mathbf{D}} \mathbb{P}_{\mathbf{D}} \left[\sup_{T \geq n} \sup_{f \in \mathcal{F}} \frac{1}{T} \left| \sum_{t=1}^{T} \left(f(x_t) - \mathbb{E}_{t-1}[f(x_t)] \right) \right| > \alpha \right] = 0$$

where the supremum is over distributions $\mathbf{D}$ over infinite sequences $(x_1, \ldots, x_T, \ldots)$

We remark that the notion of uniform Glivenko-Cantelli classes is recovered if the supremum is taken over i.i.d. distributions. The theorem below shows that finite fat shattering dimension at all scales is a sufficient condition for *Universal Uniform Convergence.*

Theorem 10. *Let $\mathcal{F}$ be a class of $[-1,1]$-valued functions. If $\mathrm{fat}_\alpha(\mathcal{F})$ is finite for all $\alpha > 0$, then $\mathcal{F}$ satisfies Universal Uniform Convergence.*

6 Supervised Learning

In this section we study the supervised learning problem where player picks a function $f_t \in \mathbb{R}^{\mathcal{X}}$ at any time t and the adversary provides input target pair (x_t, y_t) and the player suffers loss $|f_t(x_t) - y_t|$. Note that if $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$ and each $y_t \in \{\pm 1\}$ then the problem boils down to binary classification problem. As we are interested in *prediction*, we allow f_t to be outside of $\mathcal{F}$.

Though we use the absolute loss in this section, it is easy to see that all the results hold (with modified rates) for any loss $\ell(f(x), y)$ which is such that for all f, x and y,

$$\phi(\ell(\hat{y}, y)) \leq |\hat{y} - y| \leq \Phi(\ell(\hat{y}, y))$$

where Φ and ϕ are monotonically increasing functions. For instance the squared loss is a classic example.

To formally define the value of the online supervised learning game, fix a set of labels $\mathcal{Y} \subseteq [-1,1]$. Given $\mathcal{F}$, define the associated loss class,

$$\mathcal{F}_{\mathrm{S}} = \{(x,y) \mapsto |f(x) - y| \ : \ f \in \mathcal{F}\} \ .$$

Now, the supervised game is obtained using the pair $(\mathcal{F}_{\mathrm{S}}, \mathcal{X} \times \mathcal{Y})$ and we accordingly define

$$\mathcal{V}_T^{\mathrm{S}}(\mathcal{F}) = \mathcal{V}_T(\mathcal{F}_{\mathrm{S}}, \mathcal{X} \times \mathcal{Y}) \ .$$

Binary classification is, of course, a special case when $\mathcal{Y} = \{\pm 1\}$ and $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$. In that case, we simply use $\mathcal{V}_T^{\mathrm{Binary}}$ for $\mathcal{V}_T^{\mathrm{S}}$.

Proposition 11. *For the supervised learning game played with a function class $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$, for any $T \geq 1$*

$$\frac{1}{4\sqrt{2}} \sup_{\alpha} \left\{ \alpha \sqrt{T \min\{\mathrm{fat}_\alpha, T\}} \right\} \leq \frac{1}{2} \mathcal{V}_T^S(\mathcal{F})$$
$$\leq \mathfrak{R}_T(\mathcal{F}) \leq \mathfrak{D}_T(\mathcal{F}) \leq \inf_{\alpha} \left\{ 4T\alpha + 12\sqrt{T} \int_{\alpha}^{1} \sqrt{\mathrm{fat}_\beta \log\left(\frac{2eT}{\beta}\right)} \, d\beta \right\} \quad (5)$$

The proposition above implies that finiteness of the fat-shattering dimension is necessary and sufficient for learnability of a supervised game. The next theorem makes a further claim that all the complexity notions introduced so far are within a logarithmic factor from each other whenever the problem is learnable.

Theorem 12. *For any function class $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$, the following statements are equivalent*

1. *Function class $\mathcal{F}$ is online learnable in the supervised setting.*
2. *For any $\alpha > 0$, $\mathrm{fat}_\alpha(\mathcal{F})$ is finite.*

Moreover, if the function class is online learnable, then the value of the supervised game $\mathcal{V}_T^S(\mathcal{F})$, the Sequential Rademacher complexity $\mathfrak{R}(\mathcal{F})$, and the Integrated complexity $\mathfrak{D}(\mathcal{F})$ are within a multiplicative factor of $\mathcal{O}(\log^{3/2} T)$ of each other.

Corollary 13. *For the binary classification game played with function class $\mathcal{F}$ we have that*

$$K_1 \sqrt{T \min\{\mathrm{Ldim}(\mathcal{F}), T\}} \leq \mathcal{V}_T^{Binary}(\mathcal{F}) \leq K_2 \sqrt{T \ \mathrm{Ldim}(\mathcal{F}) \log T}$$

for some universal constants K_1, K_2.

We wish to point out that the lower bound of Proposition 11 also holds for "improper" supervised learning algorithms, i.e. those that simply output a prediction $\hat{y}_t \in \mathcal{Y}$ rather than a function $f_t \in \mathcal{F}$. Formally, an improper supervised learning strategy $\tilde{\pi}$ that learns $\mathcal{F}$ using a class $\mathcal{G} \subseteq \mathcal{Y}^{\mathcal{X}}$ is defined as a sequence of mappings

$$\tilde{\pi}_t : (\mathcal{X} \times \mathcal{Y})^{t-1} \mapsto \tilde{\mathcal{Q}} \ , \ t \in [T]$$

where $\tilde{\mathcal{Q}}$ denotes probability distributions over $\mathcal{G}$. We can define the value of the improper supervised learning game as

$$\tilde{\mathcal{V}}_T^{\mathrm{S}}(\mathcal{F}; \mathcal{G}) = \inf_{q_1 \in \tilde{\mathcal{Q}}} \sup_{x_1, y_1} \mathbb{E}_{f_1 \sim q_1} \cdots \inf_{q_T \in \tilde{\mathcal{Q}}} \sup_{x_T, y_T} \mathbb{E}_{f_T \sim q_T} \left[\sum_{t=1}^{T} |g_t(x_t) - y_t| - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} |f(x_t) - y_t| \right]$$

where g_t has distribution q_t. Note that $\tilde{\mathcal{V}}_T^{\mathrm{S}}(\mathcal{F};\mathcal{F}) = \mathcal{V}_T^{\mathrm{S}}(\mathcal{F})$, the latter being the value of the "proper" learning game. We say that a class $\mathcal{F}$ is improperly online learnable in the supervised setting if

$$\limsup_{T\to\infty} \frac{\tilde{\mathcal{V}}_T^{\mathrm{S}}(\mathcal{F};\mathcal{G})}{T} = 0$$

for some $\mathcal{G}$. Since a proper learning strategy can always be used as an improper learning strategy, we trivially have that if class is online learnable in the supervised setting then it is improperly online learnable. Because of the above mentioned property of the lower bound of Proposition 11, we also have the non-trivial reverse implication: if a class is improperly online learnable in the supervised setting, it is online learnable.

It is natural to ask whether being able to learn in the online model is different from learning in a batch model (in the supervised setting). The standard example (e.g. [24, 8]) is the class of step functions on a bounded interval, which has a VC dimension 1, but is not learnable in the online setting. Indeed, it is possible to verify that the Littlestone's dimension is not bounded. Interestingly, the closely-related class of "ramp" functions (modified step functions with a Lipschitz transition between 0's and 1's) *is* learnable in the online setting (and in the batch case). We extend this example as follows. By taking a convex hull of step-up and step-down functions on a unit interval, we arrive at a class of functions of bounded variation, which is learnable in the batch model, but not in the online learning model. However, the class of *Lipschitz* functions of bounded variation is learnable in both models. Online learnability of the latter class is shown with techniques analogous to Section 8.6.

6.1 Generic Algorithm

We shall now present a generic improper learning algorithm for the supervised setting that achieves a low regret bound whenever the function class is online learnable. For any $\alpha > 0$ define an α-discretization of the $[-1,1]$ interval as $B_\alpha = \{-1+\alpha/2, -1+3\alpha/2, \ldots, -1+(2k+1)\alpha/2, \ldots\}$ for $0 \le k$ and $(2k+1)\alpha \le 4$. Also for any $a \in [-1,1]$ define $\lfloor a \rfloor_\alpha = \underset{r\in B_\alpha}{\mathrm{argmin}}\, |r-a|$. For a set of functions $V \subseteq \mathcal{F}$, any $r \in B_\alpha$ and $x \in \mathcal{X}$ define

$$V(r,x) = \{f \in V \mid f(x) \in (r-\alpha/2, r+\alpha/2]\}$$

Algorithm 1 Fat-SOA Algorithm $(\mathcal{F}, \alpha)$

```
$V_1 \leftarrow \mathcal{F}$
for $t = 1$ to $T$ do
    $R_t(x) = \{r \in B_\alpha : \mathrm{fat}_\alpha(V_t(r,x)) = \max_{r' \in B_\alpha} \mathrm{fat}_\alpha(V_t(r',x))\}$
    For each $x \in \mathcal{X}$, let $f_t(x) = \frac{1}{|R_t(x)|}\sum_{r\in R_t(x)} r$
    Play $f_t$ and receive $(x_t, y_t)$
    if $|f_t(x_t) - y_t| \le \alpha$ then
        $V_{t+1} = V_t$
    else
        $V_{t+1} = V_t(\lfloor y_t \rfloor_\alpha, x_t)$
    end if
end for
```

Lemma 14. *Let $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ be a function class with finite $\mathrm{fat}_\alpha(\mathcal{F})$. Suppose the learner is presented with a sequence $(x_1,y_1),\ldots,(x_T,y_T)$, where $y_t = f(x_t)$ for some fixed $f \in \mathcal{F}$ unknown to the player. Then for f_t's computed by the Algorithm 1 it must hold that*

$$\sum_{t=1}^{T} \mathbf{1}\left\{|f_t(x_t) - y_t| > \alpha\right\} \le \mathrm{fat}_\alpha(\mathcal{F}).$$

Lemma 14 proves a bound on the performance of Algorithm 1 in the realizable setting. We now provide an algorithm for the agnostic setting. We achieve this by generating "experts" in a way similar to [8]. Using these experts along with the exponentially weighted average (EWA) algorithm we shall provide the generic algorithm for online supervised learning. The EWA (Algorithm 3) and its regret bound are provided in the appendix for completeness (p. 45).

Algorithm 2 Expert $(\mathcal{F}, \alpha, 1 \leq i_1 < \ldots < i_L \leq T, Y_1, \ldots, Y_L)$

```
$V_1 \leftarrow \mathcal{F}$
for $t = 1$ to $T$ do
  $R_t(x) = \{r \in B_\alpha : \mathrm{fat}_\alpha(V_t(r,x)) = \max_{r' \in B_\alpha} \mathrm{fat}_\alpha(V_t(r',x))\}$
  For each $x \in \mathcal{X}$, let $f'_t(x) = \frac{1}{|R_t(x)|} \sum_{r \in R_t(x)} r$
  if $t \in \{i_1, \ldots, i_L\}$ then
    $\forall x \in \mathcal{X}, f_t(x) = Y_j$ where $j$ is s.t. $t = i_j$
    Play $f_t$ and receive $x_t$
    $V_{t+1} = V_t(f_t(x_t), x_t)$
  else
    Play $f_t = f'_t$ and receive $x_t$
    $V_{t+1} = V_t$
  end if
end for
```

For each $L \leq \mathrm{fat}_\alpha(\mathcal{F})$ and every possible choice of $1 \leq i_1 < \ldots < i_L \leq T$ and $Y_1, \ldots, Y_L \in B_\alpha$ we generate an expert. Denote this set of experts as E_T. Each expert outputs a function $f_t \in \mathcal{F}$ at every round T. Hence each expert $e \in E_T$ can be seen as a sequence $(e_1, \ldots, e_T)$ of mappings $e_t : \mathcal{X}^{t-1} \mapsto \mathcal{F}$. The total number of unique experts is clearly

$$|E_T| = \sum_{L=0}^{\mathrm{fat}_\alpha} \binom{T}{L} (|B_\alpha| - 1)^L \leq \left(\frac{2T}{\alpha}\right)^{\mathrm{fat}_\alpha}$$

Lemma 15. *For any $f \in \mathcal{F}$ there exists an expert $e \in E_T$ such that for any $t \in [T]$,*

$$|f(x_t) - e(x_{1:t-1})(x_t)| \leq \alpha$$

Proof. By Lemma 14, for any function $f \in \mathcal{F}$, the number of rounds on which $|f_t(x_t) - f(x_t)| > \alpha$ for the output of the fat-SOA algorithm f_t is bounded by $\mathrm{fat}_\alpha(\mathcal{F})$. Further on each such round there are $|B_\alpha| - 1$ other possibilities. For any possible such sequence of "mistakes", there is an expert that predicts the right label on those time steps and on the remaining time agrees with the fat-SOA algorithm for that target function. Hence we see that there is always an expert $e \in E_T$ such that

$$|f(x_t) - e(x_{1:t-1})(x_t)| \leq \alpha$$

□

Theorem 16. *For any $\alpha > 0$ if we run the exponentially weighted average (EWA) algorithm with the set E_T of experts then the expected regret of the algorithm is bounded as*

$$\mathbb{E}\left[\mathbf{R}_T\right] \leq \alpha T + \sqrt{T \mathrm{fat}_\alpha \log\left(\frac{2T}{\alpha}\right)}$$

Proof. For any $\alpha \geq 0$ if we run EWA with corresponding set of experts E_T then we can guarantee that regret w.r.t. best expert in the set E_T is bounded by $\sqrt{T \mathrm{fat}_\alpha \log\left(\frac{2T}{\alpha}\right)}$. However by Lemma 15 we have that the regret of the best expert in E_T w.r.t. best function in function class $\mathcal{F}$ is at most αT. Combining we get the required result. □

The above theorem holds for a fixed α. To provide a regret statement that optimizes over α we consider α_i's of form 2^{-i} and assign weights $p_i = \frac{6}{\pi^2} i^{-2}$ to experts generated in above theorem for each α_i and run EWA on the entire set of experts with these initial weights. Hence we get the following corollary.

Corollary 17. *Let $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$. The expected regret of the algorithm described above is bounded as*

$$\mathbb{E}\left[\mathbf{R}_T\right] \le \inf_{\alpha} \left\{ \alpha T + \sqrt{T \mathrm{fat}_\alpha \log\left(\frac{2T}{\alpha}\right)} + \sqrt{T}\left(3 + 2\log\log\left(\frac{1}{\alpha}\right)\right)\right\}$$

7 Structural Results

Being able to bound complexity of a function class by a complexity of a simpler class is of great utility for proving bounds. In statistical learning theory, such structural results are obtained through properties of Rademacher averages [26, 7]. In particular, the contraction inequality due to Ledoux and Talagrand [23, Corollary 3.17], allows one to pass from a composition of a Lipschitz function with a class to the function class itself. This wonderful property permits easy convergence proofs for a vast array of problems.

We show that the notion of Sequential Rademacher complexity also enjoys many of the same properties. In Section 8, the effectiveness of the results is illustrated on a number of examples. First, we prove the contraction inequality.

Lemma 18. *Fix a class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{Z}}$ and a function $\phi : \mathbb{R} \times \mathcal{Z} \mapsto \mathbb{R}$. Assume, for all $z \in \mathcal{Z}$, $\phi(\cdot, z)$ is a Lipschitz function with a constant L. Then*

$$\mathfrak{R}(\phi(\mathcal{F})) \le L \cdot \mathfrak{R}(\mathcal{F})$$

where $\phi(\mathcal{F}) = \{z \mapsto \phi(f(z), z) : f \in \mathcal{F}\}$.

We remark that the lemma above encompasses the case of a Lipschitz $\phi : \mathbb{R} \mapsto \mathbb{R}$, as stated in [23, 7].

The next lemma bounds the Sequential Rademacher complexity for the product of function classes.

Lemma 19. *Let $\mathcal{F} = \mathcal{F}_1 \times \ldots \times \mathcal{F}_k$ where each $\mathcal{F}_j \subset \mathbb{R}^{\mathcal{X}}$. Also let $\phi : \mathbb{R}^k \mapsto \mathbb{R}$ be L-Lipschitz w.r.t. $\|\cdot\|_\infty$ norm. Then we have that*

$$\mathfrak{R}(\phi \circ \mathcal{F}) \le L\mathcal{O}\left(\log^{3/2}(T)\right) \sum_{j=1}^{k} \mathfrak{R}(\mathcal{F}_j)$$

Corollary 20. *For a fixed binary function $b : \{\pm 1\}^k \mapsto \{\pm 1\}$ and classes $\mathcal{F}_1, \ldots, \mathcal{F}_k$ of $\{\pm 1\}$-valued functions,*

$$\mathfrak{R}(g(\mathcal{F}_1, \ldots, \mathcal{F}_k)) \le \mathcal{O}\left(\log^{3/2}(T)\right) \sum_{j=1}^{k} \mathfrak{R}(\mathcal{F}_j)$$

In the next proposition, we summarize some useful properties of Sequential Rademacher complexity (see [26, 7] for the results in the i.i.d. setting)

Proposition 21. *Sequential Rademacher complexity satisfies the following properties.*

1. *If $\mathcal{F} \subset \mathcal{G}$, then $\mathfrak{R}(\mathcal{F}) \le \mathfrak{R}(\mathcal{G})$.*
2. *$\mathfrak{R}(\mathcal{F}) = \mathfrak{R}(\mathrm{conv}(\mathcal{F}))$.*
3. *$\mathfrak{R}(c\mathcal{F}) = |c|\mathfrak{R}(\mathcal{F})$ for all $c \in \mathbb{R}$.*
4. *If $\phi : \mathbb{R} \mapsto \mathbb{R}$ is L-Lipschitz, then $\mathfrak{R}(\phi(\mathcal{F})) \le L\mathfrak{R}(\mathcal{F})$.*
5. *For any h, $\mathfrak{R}(\mathcal{F} + h) = \mathfrak{R}(\mathcal{F})$ where $\mathcal{F} + h = \{f + h : f \in \mathcal{F}\}$*

8 Examples and Applications

8.1 Example: Linear Function Classes

Suppose $\mathcal{F}_\mathcal{W}$ is a class consisting of linear functions $x \mapsto \langle w, x\rangle$ where the weight vector w comes from some set $\mathcal{W}$,

$$\mathcal{F}_\mathcal{W} = \{x \mapsto \langle w, x\rangle \,:\, w \in \mathcal{W}\} \ .$$

Often, it is possible to find a strongly convex function non-negative $\Psi(\mathbf{w})$ such that $\Psi(\mathbf{w}) \le \Psi_{\max} < \infty$ for all $\mathbf{w} \in \mathcal{W}$. Recall that a function $\Psi : \mathcal{W} \to \mathbb{R}$ is σ-strongly convex on $\mathcal{W}$ w.r.t. a norm $\|\cdot\|$ if, for all $\theta \in [0,1]$ and $w_1, w_2 \in \mathcal{W}$,

$$\Psi(\theta w_1 + (1-\theta) w_2) \le \theta \Psi(w_1) + (1-\theta)\Psi(w_2) - \frac{\sigma\,\theta\,(1-\theta)}{2}\|w_1 - w_2\|^2$$

We will give examples shortly but we first state a proposition that is useful to bound the Sequential Rademacher complexity of such linear function classes.

Proposition 22. *Let $\mathcal{W}$ be a class of weight vectors such that $0 \le \Psi(w) \le \Psi_{\max}$ for all $w \in \mathcal{W}$. Suppose that Ψ is σ-strongly convex w.r.t. a given norm $\|\cdot\|$. Then, we have,*

$$\mathfrak{R}_T(\mathcal{F}_\mathcal{W}) \le \|\mathcal{X}\|_\star \sqrt{\frac{2\,\Psi_{\max}\,T}{\sigma}} \ ,$$

where $\|\mathcal{X}\|_\star = \sup_{x\in\mathcal{X}}\ \|x\|_\star$, the maximum dual norm of any vector in the input space.

The proof of Proposition 22 is given in the appendix. It relies on the following lemma which can be found in [17]. There it is stated for i.i.d. mean zero Z_i but the proof given works even for martingale difference sequences.

Lemma 23. *Let $\Psi : \mathcal{W} \to \mathbb{R}$ be σ-strongly convex w.r.t. $\|\cdot\|$. Let $Z_t, t \ge 1$ be a martingale difference sequence w.r.t. some filtration $\{\mathcal{G}_t\}_{t\ge 1}$ (i.e. $\mathbb{E}\left[Z_t \mid \mathcal{G}_{t-1}\right] = 0$) such that $\mathbb{E}\left[\|Z_t\|_\star^2\right] \le V^2$. Define $S_t = \sum_{s\le t} Z_s$. Then, $\Psi^\star(S_t) - tV^2/2\sigma$ is a supermartingale. Furthermore, if $\inf_{w\in\mathcal{W}} \Psi(w) \ge 0$, then*

$$\mathbb{E}\left[\Psi^\star(S_T)\right] \le \frac{V^2\,T}{2\,\sigma} \ .$$

We will now show how to use the above result to derive minimax regret guarantees for *online convex optimization.* This is a particular instance of online learning where $\mathcal{F} = K \subseteq \mathbb{R}^d$ where K is a bounded closed convex set. Suppose $\|u\| \le D$ for all $u \in K$ for some norm $\|\cdot\|$. The adversary's set $\mathcal{X}$ consists of convex G-Lipschitz (w.r.t. the dual norm $\|\cdot\|_\star$) functions over K:

$$\mathcal{X} = \mathcal{X}_{\text{cvx}} = \{g : K \mapsto \mathbb{R} \,:\, g \text{ convex and } G\text{-Lipschitz w.r.t. } \|\cdot\|_\star\} \ .$$

We could directly try to bound the value $\mathcal{V}_T(\mathcal{F}, \mathcal{X}_{\text{cvx}})$ by $\mathfrak{R}_T(\mathcal{F}, \mathcal{X}_{\text{cvx}})$ but this, in fact, cannot give a non-trivial bound [31]. Instead, we use the lemma below to bound the value of the convex game with that of the linear game, i.e. one in which

$$\mathcal{X} = \mathcal{X}_{\text{lin}} = \{u \mapsto \langle u, x\rangle \,:\, \|x\|_\star \le G\} \ .$$

Lemma 24. *Suppose $\mathcal{F} = K \subseteq \mathbb{R}^d$ is a closed bounded convex set and let $\mathcal{X}_{\text{cvx}}, \mathcal{X}_{\text{lin}}$ be defined as above. Then, we have*

$$\mathcal{V}_T(\mathcal{F}, \mathcal{X}_{\text{cvx}}) = \mathcal{V}_T(\mathcal{F}, \mathcal{X}_{\text{lin}}) \ .$$

Using the above lemma in conjunction with Proposition 22 above, we can immediately conclude that

$$\mathcal{V}_T(\mathcal{F}, \mathcal{X}_{\mathrm{cvx}}) \le \mathfrak{R}_T(\mathcal{F}, \mathcal{X}_{\mathrm{lin}}) \le G\sqrt{2\,\Psi_{\max}\,T/\sigma}$$

for any non-negative function $\Psi : K \to \mathbb{R}$ that is σ-strongly w.r.t. $\|\cdot\|$. Note that, typically, $\Psi_{\max}$ will depend on D. For example, in the particular case when $\|\cdot\| = \|\cdot\|_\star = \|\cdot\|_2$, we can take $\Psi(u) = \frac{1}{2}\|u\|_2^2$ and the above regret bound becomes $GD\sqrt{T}$ and recovers the guarantee of Zinkevich for his online gradient descent algorithm. In general, for $\|\cdot\| = \|\cdot\|_p, \|\cdot\|_\star = \|\cdot\|_q$, we can use $\Psi(u) = \frac{1}{2}\|u\|_p^2$ to get a bound of $GD\sqrt{T/(p-1)}$ since Ψ is $(p-1)$-strongly convex w.r.t. $\|\cdot\|_p$. These $O(\sqrt{T})$ regret rates are not new but we rederive them to illustrate the usefulness of the tools we developed.

8.2 Example: Margin Based Regret

In the classical statistical setting, margin bounds provide guarantees on expected zero-one loss of a classifier based on the empirical margin zero-one error. These results form the basis of the theory of large margin classifiers (see [30, 22]). Recently, in the online setting, margin bounds have been shown through the concept of margin via the Littlestone dimension [8]. We show that our machinery can easily lead to margin bounds for the binary classification games for general function classes $\mathcal{F}$ based on their sequential Rademacher Complexity. We use ideas from [22] to do this.

Proposition 25. *For any function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ bounded by 1, there exists a randomized player strategy given by π such that for any sequence $z_1, \ldots, z_T$ where each $z_t = (x_t, y_t) \in \mathcal{X} \times \{\pm 1\}$, played by the adversary,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{f_t \sim \pi_t(z_{1:t-1})}\left[\mathbf{1}\left\{f_t(x_t) y_t < 0\right\}\right]\right] \le \inf_{\gamma > 0}\left\{\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathbf{1}\left\{f(x_t) y_t < \gamma\right\} + \frac{4}{\gamma}\mathfrak{R}_T(\mathcal{F}) + \sqrt{T}\left(3 + \log\log\left(\frac{1}{\gamma}\right)\right)\right\}$$

8.3 Example : Neural Networks

We provide below a bound on sequential Rademacher complexity for classic multi-layer neural networks thus showing they are learnable in the online setting. The model of neural network we consider below and the bounds we provide are analogous to the ones considered in the batch setting in [7]. We now consider a k-layer 1-norm neural network. To this end let function class $\mathcal{F}_1$ be given by

$$\mathcal{F}_1 = \left\{ x \mapsto \sum_j w_j^1 x_j \;\middle|\; \|w\|_1 \le B_1 \right\}$$

and further for each $2 \le i \le k$ define

$$\mathcal{F}_i = \left\{ x \mapsto \sum_j w_j^i \sigma\left(f_j(x)\right) \;\middle|\; \forall j \;\; f_j \in \mathcal{F}_{i-1}, \|w^i\|_1 \le B_i \right\}$$

Proposition 26. *Say $\sigma : \mathbb{R} \mapsto [-1, 1]$ is L-Lipschitz, then*

$$\mathfrak{R}_T(\mathcal{F}_k) \le \left(\prod_{i=1}^{k} 2B_i\right) L^{k-1} X_\infty \sqrt{2T \log d}$$

where X_∞ is such that $\forall x \in \mathcal{X}$, $\|x\|_\infty \le X_\infty$ and $\mathcal{X} \subset \mathbb{R}^d$

8.4 Example: Decision Trees

We consider here the supervised learning game where adversary provides instances from instance space $\mathcal{X}$ and binary labels ± 1 corresponding to the instances and the player plays decision trees of depth no more than d with decision functions from set $\mathcal{H} \subset \{\pm 1\}^{\mathcal{X}}$ of binary valued functions. The following proposition shows that there exists a player strategy which under certain circumstances could have low regret for the supervised learning (binary) game played with class of decision trees of depth at most d with decision functions from $\mathcal{H}$. The proposition is analogous to the one in [7] considered in the batch (classical) setting.

Proposition 27. *Denote by $\mathcal{T}$ the class of decision trees of depth at most d with decision functions in $\mathcal{H}$. There exists a randomized player strategy π such that for any sequence of instances $z_1 = (x_1, y_1), \ldots, z_T = (x_T, y_T) \in (\mathcal{X} \times \{\pm 1\})^T$ played by the adversary,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{f_t \sim \pi_t(z_{1:t-1})}\left[\mathbf{1}\left\{f_t(x_t) \neq y_t\right\}\right]\right] \leq \inf_{t \in \mathcal{T}} \sum_{t=1}^{T} \mathbf{1}\left\{t(x_t) \neq y_t\right\} + \mathcal{O}\left(\sum_{l} \min\left(\tilde{C}_T(l), d \log^{3/2}(T)\ \mathfrak{R}(\mathcal{H})\right) + \sqrt{T}\,(3 + 2\log(N_{leaf}))\right)$$

where $\tilde{C}_T(l)$ denotes the number of instances which reach the leaf l and are correctly classified in the decision tree t that minimizes $\sum_{t=1}^{T} \mathbf{1}\left\{t(x_t) \neq y_t\right\}$ and let N_{leaf} be the number of leaves in this tree.

8.5 Example: Transductive Learning and Prediction of Individual Sequences

Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. Let

$$\widehat{\mathcal{N}}_\infty(\alpha, \mathcal{F}) = \min\left\{|G| : G \subseteq \mathbb{R}^{\mathcal{X}} \text{ s.t. } \forall f \in \mathcal{F}\ \ \exists g \in G\ \text{ satisfying }\ \|f - g\|_\infty \leq \alpha\right\}. \tag{6}$$

be the ℓ_∞ covering number at scale α, where the cover is pointwise on all of $\mathcal{X}$. It is easy to verify that

$$\forall T, \quad \mathcal{N}_\infty(\alpha, \mathcal{F}, T) \leq \widehat{\mathcal{N}}_\infty(\alpha, \mathcal{F}) \tag{7}$$

Indeed, let G be a minimal cover of $\mathcal{F}$ at scale α. We claim that the set $V = \{\mathbf{v}^g = g(\mathbf{x}) : g \in G\}$ of $\mathbb{R}$-valued trees is an ℓ_∞ cover of $\mathcal{F}$ on $\mathbf{x}$. Fix any $\epsilon \in \{\pm 1\}^T$ and $f \in \mathcal{F}$, and let $g \in G$ be such that $\|f - g\|_\infty \leq \alpha$. Then clearly $|\mathbf{v}^g_t(\epsilon) - f(\mathbf{x}_t(\epsilon))|$ for any $1 \leq t \leq T$, which concludes the proof.

This simple observation can be applied in several situations. First, consider the problem of *transductive learning*, where the set $\mathcal{X} = \{z_1, \ldots, z_n\}$ is a finite set. To ensure online learnability, it is sufficient to consider an assumption on the dependence of $\widehat{\mathcal{N}}_\infty(\alpha, \mathcal{F})$ on α. An obvious example of such a class is a VC-type class with $\widehat{\mathcal{N}}_\infty(\alpha, \mathcal{F}) \leq (c/\alpha)^d$ for some c which can depend on n. Assume that $\mathcal{F} \subset [-1, 1]^{\mathcal{X}}$. Substituting this bound on the covering number into

$$\mathfrak{D}_T(\mathcal{F}) = \inf_{\alpha}\left\{4T\alpha + \int_{\alpha}^{1} \sqrt{T\ \log\ \mathcal{N}_2(\delta, \mathcal{F}, T)}\, d\delta\right\}$$

and choosing $\alpha = 0$, we observe that the value of the supervised game is upper bounded by $2\mathfrak{D}_T(\mathcal{F}) \leq 4\sqrt{dT \log c}$ by Proposition 11. It is easy to see that if n is fixed and the problem is learnable in the batch (e.g. PAC) setting, then the problem is learnable in the online transductive model.

In the transductive setting considered by Kakade and Kalai [16], it is assumed that $n \leq T$ and $\mathcal{F}$ are binary-valued. If $\mathcal{F}$ is a class with VC dimension d, the Sauer-Shelah lemma ensures that the ℓ_∞ cover is smaller than $(en/d)^d \leq (eT/d)^d$. Using the previous argument with $c = eT$, we obtain a bound of $4\sqrt{dT \log(eT)}$ for the value of the game, matching [16] up to a constant 2.

We also consider the problem of prediction of individual sequences, which has been studied both in information theory and in learning theory. In particular, in the case of binary prediction, Cesa-Bianchi and Lugosi [9] proved upper bounds on the value of the game in terms of the (classical) Rademacher complexity and the (classical) Dudley integral. The particular assumption made in [9] is that experts are *static*. That is, their prediction only depends on the current round, not on the past information. Formally, we define static experts as mappings $f : \{1, \dots, T\} \mapsto \mathcal{Y} = [-1, 1]$, and let $\mathcal{F}$ denote a class of such experts. Defining $\mathcal{X} = \{1, \dots, T\}$ puts us in the setting considered earlier with $n = T$. We immediately obtain $4\sqrt{dT \log(eT)}$, matching the results on [9, p. 1873]. We mention that the upper bound in Theorem 4 in [9] is tighter by a $\log T$ factor if a sharper bound on the ℓ_2 cover is considered. Finally, for the case of a finite number of experts, clearly $\widehat{\mathcal{N}}_\infty \leq N$ which gives the classical $O(\sqrt{T \log N})$ bound on the value of the game [10].

8.6 Example: Isotron

Recently, Kalai and Sastry [18] introduced a method called *Isotron* for learning Single Index Models (SIM). These models generalize linear and logistic regression, generalized linear models, and classification by linear threshold functions. For brevity, we only describe the Idealized SIM problem from [18]. In its "batch" version, we assume that the data is revealed at once as a set $\{(x_i, y_i)\}_{t=1}^T \in \mathbb{R}^n \times \mathbb{R}$ where $y_t = u(\langle w, x_i \rangle)$ for some unknown $w \in \mathbb{R}^n$ of bounded norm and an unknown non-decreasing $u : \mathbb{R} \mapsto \mathbb{R}$ with a bounded Lipschitz constant. Given this data, the goal is to iteratively find the function u and the direction w, making as few mistakes as possible. The error is measured as $\frac{1}{T}\sum_{t=1}^T (f_i(x_t) - y_t)^2$, where $f_i(x) = u_i(\langle w_i, x \rangle)$ is the iterative approximation found by the algorithm on the ith round. The elegant computationally efficient method presented in [18] is motivated by Perceptron, and a natural open question posed by the authors is whether there is an online variant of Isotron. Before even attempting a quest for such an algorithm, we can ask a more basic question: is the (Idealized) SIM problem even learnable in the online framework? After all, most online methods deal with convex functions, but u is only assumed to be Lipschitz and non-decreasing. We answer the question easily with the tools we have developed.

We are interested in online learnability in the supervised setting of the following class of functions

$$\mathcal{H} = \{f(x, y) = (y - u(\langle w, x \rangle))^2 \mid u : [-1, 1] \mapsto [-1, 1] \text{ 1-Lipschitz }, \|w\|_2 \leq 1\} \tag{8}$$

over $\mathcal{X} = B_2$ (the unit Euclidean ball in $\mathbb{R}^d$) and $\mathcal{Y} = [-1, 1]$, where both u and w range over the possibilities. In particular, we prove the result for Lipschitz, but not necessarily non-decreasing functions. It is evident that $\mathcal{H}$ is a composition with three levels: the squared loss, the Lipschitz non-decreasing function, and the linear function. The proof of the following Proposition boils down to showing that the covering number of the class does not increase much under these compositions.

Proposition 28. *The class $\mathcal{H}$ defined in* (8) *is online learnable in the supervised setting. Moreover,*

$$\mathcal{V}_T(\mathcal{H}, \mathcal{X} \times \mathcal{Y}) = O(\sqrt{T} \log^{3/2} T).$$

Proof. First, by the classical result of Kolmogorov and Tihomirov [20], the class $\mathcal{G}$ of all bounded Lipschitz functions has small metric entropy: $\log \widehat{\mathcal{N}}_\infty(\alpha, \mathcal{G}) = \Theta(1/\alpha)$. For the particular class of non-decreasing 1-Lipschitz functions, it is trivial to verify that the entropy is in fact bounded by $2/\alpha$.

Next, consider the class $\mathcal{F} = \{\langle w, x \rangle \mid \|w\|_2 \leq 1\}$ over the Euclidean ball. By Proposition 22, $\mathfrak{R}_T(\mathcal{F}) \leq \sqrt{2T}$. Using the lower bound of Proposition 11, $\text{fat}_\alpha \leq 64/\alpha^2$ whenever $\alpha > 8/\sqrt{T}$. This implies that $\mathcal{N}_\infty(\alpha, \mathcal{F}, T) \leq (2eT/\alpha)^{64/\alpha^2}$ whenever $\alpha > 8/\sqrt{T}$. Note that this bound does not depend on the ambient dimension of $\mathcal{X}$.

Next, we show that a composition of $\mathcal{G}$ with any small class $\mathcal{F} \subset [-1, 1]^{\mathcal{X}}$ also has a small cover. To this end, suppose $\mathcal{N}_\infty(\alpha, \mathcal{F}, T)$ is the covering number for $\mathcal{F}$. Fix a particular tree $\mathbf{x}$ and let $V = \{\mathbf{v}_1, \dots, \mathbf{v}_N\}$ be an ℓ_∞ cover of $\mathcal{F}$ on $\mathbf{x}$ at scale α. Analogously, let $W = \{g_1, \dots, g_M\}$ be an ℓ_∞ cover of $\mathcal{G}$ with

$M = \widehat{\mathcal{N}}_\infty(\alpha, \mathcal{G})$. Consider the class $\mathcal{G} \circ \mathcal{F} = \{g \circ f : g \in \mathcal{G}, f \in \mathcal{F}\}$. The claim is that $\{g(\mathbf{v}) : \mathbf{v} \in V, g \in W\}$ provides an ℓ_∞ cover for $\mathcal{G} \circ \mathcal{F}$ on $\mathbf{x}$. Fix any $f \in \mathcal{F}, g \in \mathcal{G}$ and $\epsilon \in \{\pm 1\}^T$. Let $\mathbf{v} \in V$ be such that $\max_{t \in [T]} |f(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon)| \leq \alpha$, and let $g' \in W$ be such that $\|g - g'\|_\infty \leq \alpha$. Then, using the fact that functions in $\mathcal{G}$ are 1-Lipschitz, for any $t \in [T]$,

$$|g(f(\mathbf{x}_t(\epsilon))) - g'(\mathbf{v}_t(\epsilon))| \leq |g(f(\mathbf{x}_t(\epsilon))) - g'(f(\mathbf{x}_t(\epsilon))| + |g'(f(\mathbf{x}_t(\epsilon)) - g'(\mathbf{v}_t(\epsilon))| \leq 2\alpha \ .$$

Hence, $\mathcal{N}_\infty(2\alpha, \mathcal{G} \circ \mathcal{F}, T) \leq \widehat{\mathcal{N}}_\infty(\alpha, \mathcal{G}) \times \mathcal{N}_\infty(\alpha, \mathcal{F}, T)$.

Finally, we put all the pieces together. By Lemma 18, the Sequential Rademacher complexity of $\mathcal{H}$ is bounded by 4 times the Sequential Rademacher complexity of the class

$$\mathcal{G} \circ \mathcal{F} = \{u(\langle w, x \rangle) \mid u : [-1, 1] \mapsto [-1, 1] \text{ is 1-Lipschitz },\ \|w\|_2 \leq 1\}$$

since the squared loss is 4-Lipschitz on the space of possible values. The latter complexity is then bounded by

$$\mathfrak{D}(\mathcal{G} \circ \mathcal{F}) \leq 32\sqrt{T} + 12 \int_{8/\sqrt{T}}^{1} \sqrt{T \ \log\ \mathcal{N}(\delta, \mathcal{G} \circ \mathcal{F}, T)}\, d\delta \leq 32\sqrt{T} + 12\sqrt{T} \int_{8/\sqrt{T}}^{1} \sqrt{\frac{2}{\delta} + \frac{64}{\delta^2} \log(2eT)} d\delta \ .$$

We conclude that the value of the game $\mathcal{V}_T(\mathcal{H}, \mathcal{X} \times \mathcal{Y}) = O(\sqrt{T} \log^{3/2} T)$. □

Acknowledgements

We would like to thank J. Michael Steele for helpful discussions. We thank Sham Kakade for bringing to our attention the Isotron algorithm. A. Rakhlin gratefully acknowledges the support of NSF under grant CAREER DMS-0954737.

References

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

[2] J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 414–424. Omnipress, 2008.

[3] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44:615–631, 1997.

[4] N. Alon and J. Spencer. *The Probabilistic Method.* John Wiley & Sons, 2nd edition, 2000.

[5] P. L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems 20*, 2007.

[6] P. L. Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996. (special issue on COLT‘94).

[7] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2003.

[8] S. Ben-David, D. Pal, and S. Shalev-Shwartz. Agnostic online learning. In *Proceedings of the 22th Annual Conference on Learning Theory*, 2009.

[9] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Annals of Statistics*, pages 1865–1895, 1999.

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

[11] R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

[12] R. M. Dudley. *Uniform Central Limit Theorems.* Cambridge University Press, 1999.

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

[14] J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

[15] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[16] S. M. Kakade and A. Kalai. From batch to transductive online learning. In *NIPS*, 2005.

[17] S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21*, pages 793–800. MIT Press, 2009.

[18] A. Tauman Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. In *Proceedings of the 22th Annual Conference on Learning Theory*, 2009.

[19] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

[20] A.N. Kolmogorov and V.M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

[21] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability II*, 47:443–459, 2000.

[22] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1):1–50, 2002.

[23] M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer-Verlag, New York, 1991.

[24] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 04 1988.

[25] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX(2):245–303, 2000.

[26] S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures in Machine Learning, LNCS 2600, Machine Learning Summer School 2002, Canberra, Australia, February 11-22*, pages 1–40. Springer, 2003.

[27] S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152:37–55, 2003.

[28] D. Pollard. *Empirical Processes: Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics.* Institute of Mathematical Statistics, Hayward, CA, 1990.

[29] N. Sauer. On the density of families of sets. *J. Combinatorial Theory*, 13:145–147, 1972.

[30] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, pages 322–330, 1997.

[31] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Conference on Learning Theory*, 2009.

[32] S. Shalev-Shwartz and Y. Singer. Convex repeated games and fenchel duality. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1265–1272. MIT Press, Cambridge, MA, 2007.

[33] S. Shelah. A combinatorial problem: Stability and order for models and theories in infinitary languages. *Pac. J. Math*, 4:247–261, 1972.

[34] K. Sridharan and A. Tewari. Convex games in Banach spaces. In *Proceedings of the 23nd Annual Conference on Learning Theory*, 2010.

[35] A. W. Van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes : With Applications to Statistics.* Springer Series, March 1996.

[36] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[37] S.A. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2000.

[38] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data (Springer Series in Statistics).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1982.

[39] V. N. Vapnik. Inductive principles of the search for empirical dependences (methods based on weak convergence of probability measures). In *Annual Workshop on Computational Learning Theory*, 1989.

[40] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[41] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.

A Proofs

Proof of Theorem 1. For simplicity, denote $\psi(x_{1:T}) = \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} f(x_t)$. The first step in the proof is to appeal to the minimax theorem for every couple of inf and sup:

$$\begin{aligned}
&\inf_{q_1}\sup_{x_1}\mathbb{E}_{f_1\sim q_1}\ldots\inf_{q_T}\sup_{x_T}\mathbb{E}_{f_T\sim q_T}\left[\sum_{t=1}^{T} f_t(x_t) - \psi(x_{1:T})\right]\\
&\quad= \inf_{q_1\in\mathcal{Q}}\sup_{p_1}\mathbb{E}_{\substack{f_1\sim q_1\\ x_1\sim p_1}}\ldots\inf_{q_T\in\mathcal{Q}}\sup_{p_T}\mathbb{E}_{\substack{f_T\sim q_T\\ x_T\sim p_T}}\left[\sum_{t=1}^{T} f_t(x_t) - \psi(x_{1:T})\right]\\
&\quad= \sup_{p_1}\inf_{q_1\in\mathcal{Q}}\mathbb{E}_{\substack{f_1\sim q_1\\ x_1\sim p_1}}\ldots\sup_{p_T}\inf_{q_T\in\mathcal{Q}}\mathbb{E}_{\substack{f_T\sim q_T\\ x_T\sim p_T}}\left[\sum_{t=1}^{T} f_t(x_t) - \psi(x_{1:T})\right] \qquad \text{(by Minimax theorem)}\\
&\quad= \sup_{p_1}\inf_{f_1}\mathbb{E}_{x_1\sim p_1}\ldots\sup_{p_T}\inf_{f_T}\mathbb{E}_{x_T\sim p_T}\left[\sum_{t=1}^{T} f_t(x_t) - \psi(x_{1:T})\right]
\end{aligned}$$

From now on, it will be understood that x_t has distribution p_t. By moving the expectation with respect to x_T and then the infimum with respect to f_T inside the expression, we arrive at

$$\begin{aligned}
&\sup_{p_1}\inf_{f_1}\mathbb{E}_{x_1}\dots\sup_{p_{T-1}}\inf_{f_{T-1}}\mathbb{E}_{x_{T-1}}\sup_{p_T}\left[\sum_{t=1}^{T-1}f_t(x_t)+\left[\inf_{f_T}\mathbb{E}_{x_T}f_T(x_T)\right]-\mathbb{E}_{x_T}\psi(x_{1:T})\right]\\
&=\sup_{p_1}\inf_{f_1}\mathbb{E}_{x_1}\dots\sup_{p_{T-1}}\inf_{f_{T-1}}\mathbb{E}_{x_{T-1}}\sup_{p_T}\mathbb{E}_{x_T}\left[\sum_{t=1}^{T-1}f_t(x_t)+\left[\inf_{f_T}\mathbb{E}_{x_T}f_T(x_T)\right]-\psi(x_{1:T})\right]
\end{aligned}$$

Let us now repeat the procedure for step $T-1$. The above expression is equal to

$$\begin{aligned}
&\sup_{p_1}\inf_{f_1}\mathbb{E}_{x_1}\dots\sup_{p_{T-1}}\inf_{f_{T-1}}\mathbb{E}_{x_{T-1}}\left[\sum_{t=1}^{T-1}f_t(x_t)+\sup_{p_T}\mathbb{E}_{x_T}\left[\inf_{f_T}\mathbb{E}_{x_T}f_T(x_T)-\psi(x_{1:T})\right]\right]\\
&=\sup_{p_1}\inf_{f_1}\mathbb{E}_{x_1}\dots\sup_{p_{T-1}}\left[\sum_{t=1}^{T-2}f_t(x_t)+\left[\inf_{f_{T-1}}\mathbb{E}_{x_{T-1}}f_{T-1}\right]+\mathbb{E}_{x_{T-1}}\sup_{p_T}\mathbb{E}_{x_T}\left[\inf_{f_T}\mathbb{E}_{x_T}f_T(x_T)-\psi(x_{1:T})\right]\right]\\
&=\sup_{p_1}\inf_{f_1}\mathbb{E}_{x_1}\dots\sup_{p_{T-1}}\mathbb{E}_{x_{T-1}}\sup_{p_T}\mathbb{E}_{x_T}\left[\sum_{t=1}^{T-2}f_t(x_t)+\left[\inf_{f_{T-1}}\mathbb{E}_{x_{T-1}}f_{T-1}\right]+\left[\inf_{f_T}\mathbb{E}_{x_T}f_T(x_T)\right]-\psi(x_{1:T})\right]
\end{aligned}$$

Continuing in this fashion for $T-2$ and all the way down to $t=1$ proves the theorem. □

Proof of the Key Technical Lemma (Lemma 3). We start by noting that since x_T, x'_T are both drawn from p_T,

$$\begin{aligned}
\mathbb{E}_{x_T,x'_T\sim p_T}\left[\Phi\left(\sum_{t=1}^{T}\Delta_f(x_t,x'_t)\right)\right]&=\mathbb{E}_{x_T,x'_T\sim p_T}\left[\Phi\left(\sum_{t=1}^{T-1}\Delta_f(x_t,x'_t)+\Delta_f(x_T,x'_T)\right)\right]\\
&=\mathbb{E}_{x'_T,x_T\sim p_T}\left[\Phi\left(\sum_{t=1}^{T-1}\Delta_f(x_t,x'_t)+\Delta_f(x_T,x'_T)\right)\right]\\
&=\mathbb{E}_{x_T,x'_T\sim p_T}\left[\Phi\left(\sum_{t=1}^{T-1}\Delta_f(x_t,x'_t)-\Delta_f(x_T,x'_T)\right)\right]\ ,
\end{aligned}$$

where the last line is by antisymmetry of Δ_f. Since the first and last lines are equal, they are both equal to their average and hence

$$\mathbb{E}_{x_T,x'_T\sim p_T}\left[\Phi\left(\sum_{t=1}^{T-1}\Delta_f(x_t,x'_t)\right)\right]=\mathbb{E}_{x_T,x'_T\sim p_T}\left[\mathbb{E}_{\epsilon_T}\left[\Phi\left(\sum_{t=1}^{T-1}\Delta_f(x_t,x'_t)+\epsilon_T\Delta_f(x_T,x'_T)\right)\right]\right]\ .$$

Hence we conclude that

$$\begin{aligned}
\sup_{p_T}\mathbb{E}_{x_T,x'_T\sim p_T}\left[\Phi\left(\sum_{t=1}^{T}\Delta_f(x_t,x'_t)\right)\right]&=\sup_{p_T}\mathbb{E}_{x_T,x'_T\sim p_T}\left[\mathbb{E}_{\epsilon_T}\left[\Phi\left(\sum_{t=1}^{T-1}\Delta_f(x_t,x'_t)+\epsilon_T\Delta_f(x_T,x'_T)\right)\right]\right]\\
&\le\sup_{x_T,x'_T}\mathbb{E}_{\epsilon_T}\left[\Phi\left(\sum_{t=1}^{T-1}\Delta_f(x_t,x'_t)+\epsilon_T\Delta_f(x_T,x'_T)\right)\right]\ .
\end{aligned}$$

Using the above and noting that x_{T-1}, x'_{T-1} are both drawn from p_{T-1} and hence similar to previous step introducing Rademacher variable ϵ_{T-1} we get that

$$\begin{aligned}
&\sup_{p_{T-1}} \mathbb{E}_{x_{T-1},x'_{T-1}\sim p_{T-1}} \sup_{p_T} \mathbb{E}_{x_T,x'_T\sim p_T} \left[\Phi\left(\sum_{t=1}^{T} \Delta_f(x_t,x'_t)\right)\right] \\
&\qquad \leq \sup_{p_{T-1}} \mathbb{E}_{x_{T-1},x'_{T-1}\sim p_{T-1}} \left[\sup_{x_T,x'_T} \mathbb{E}_{\epsilon_T}\left[\Phi\left(\sum_{t=1}^{T-1} \Delta_f(x_t,x'_t) + \epsilon_T \Delta_f(x_T,x'_T)\right)\right]\right] \\
&\qquad = \sup_{p_{T-1}} \mathbb{E}_{x_{T-1},x'_{T-1}\sim p_T} \mathbb{E}_{\epsilon_{T-1}} \left[\sup_{x_T,x'_T} \mathbb{E}_{\epsilon_T}\left[\Phi\left(\sum_{t=1}^{T-2} \Delta_f(x_t,x'_t) + \epsilon_{T-1}\Delta_f(x_{T-1},x'_{T-1}) + \epsilon_T \Delta_f(x_T,x'_T)\right)\right]\right] \\
&\qquad \leq \sup_{x_{T-1},x'_{T-1}} \mathbb{E}_{\epsilon_{T-1}} \left[\sup_{x_T,x'_T} \mathbb{E}_{\epsilon_T}\left[\Phi\left(\sum_{t=1}^{T-2} \Delta_f(x_t,x'_t) + \epsilon_{T-1}\Delta_f(x_{T-1},x'_{T-1}) + \epsilon_T \Delta_f(x_T,x'_T)\right)\right]\right] \ .
\end{aligned}$$

Proceeding in similar fashion introducing Rademacher variables all the way upto ϵ_1 we finally get the required statement that

$$\sup_{p_1} \mathbb{E}_{x_1,x'_1\sim p_1} \ldots \sup_{p_T} \mathbb{E}_{x_T,x'_T\sim p_T} \left[\Phi\left(\sum_{t=1}^{T} \Delta_f(x_t,x'_t)\right)\right] \leq \sup_{x_1,x'_1} \left\{\mathbb{E}_{\epsilon_1}\left[\ldots \sup_{x_T,x'_T}\left\{\mathbb{E}_{\epsilon_T}\left[\Phi\left(\sum_{t=1}^{T} \epsilon_t \Delta_f(x_t,x'_t)\right)\right]\right\}\ldots\right]\right\}$$

□

Proof of Lemma 8. For any $\lambda > 0$, we invoke Jensen's inequality to get

$$\begin{aligned}
M(\lambda) := \exp\left\{\lambda \mathbb{E}_\epsilon\left[\max_{\mathbf{v}\in V} \sum_{t=1}^{T} \epsilon_t \mathbf{v}_t(\epsilon)\right]\right\} &\leq \mathbb{E}_\epsilon\left[\exp\left\{\lambda \max_{\mathbf{v}\in V} \sum_{t=1}^{T} \epsilon_t \mathbf{v}_t(\epsilon)\right\}\right] \\
&\leq \mathbb{E}_\epsilon\left[\max_{\mathbf{v}\in V} \exp\left\{\lambda \sum_{t=1}^{T} \epsilon_t \mathbf{v}_t(\epsilon)\right\}\right] \leq \mathbb{E}_\epsilon\left[\sum_{\mathbf{v}\in V} \exp\left\{\lambda \sum_{t=1}^{T} \epsilon_t \mathbf{v}_t(\epsilon)\right\}\right]
\end{aligned}$$

With the usual technique of peeling from the end,

$$\begin{aligned}
M(\lambda) &\leq \sum_{\mathbf{v}\in V} \mathbb{E}_{\epsilon_1,\ldots,\epsilon_T}\left[\prod_{t=1}^{T} \exp\left\{\lambda \epsilon_t \mathbf{v}_t(\epsilon_{1:t-1})\right\}\right] \\
&= \sum_{\mathbf{v}\in V} \mathbb{E}_{\epsilon_1,\ldots,\epsilon_{T-1}}\left[\prod_{t=1}^{T-1} \exp\left\{\lambda \epsilon_t \mathbf{v}_t(\epsilon_{1:t-1})\right\} \times \left(\frac{\exp\left\{\lambda \mathbf{v}_T(\epsilon_{1:T-1})\right\} + \exp\left\{-\lambda \mathbf{v}_T(\epsilon_{1:T-1})\right\}}{2}\right)\right] \\
&\leq \sum_{\mathbf{v}\in V} \mathbb{E}_{\epsilon_1,\ldots,\epsilon_{T-1}}\left[\prod_{t=1}^{T-1} \exp\left\{\lambda \epsilon_t \mathbf{v}_t(\epsilon_{1:t-1})\right\} \times \exp\left\{\frac{\lambda^2 \mathbf{v}_T(\epsilon_{1:T-1})^2}{2}\right\}\right]
\end{aligned}$$

where we used the inequality $\frac{1}{2}\left\{\exp(a) + \exp(-a)\right\} \leq \exp(a^2/2)$, valid for all $a \in \mathbb{R}$. Peeling off the second term is a bit more involved:

$$\begin{aligned}
M(\lambda) \leq \sum_{\mathbf{v}\in V} \mathbb{E}_{\epsilon_1,\ldots,\epsilon_{T-2}}\Bigg[&\prod_{t=1}^{T-2} \exp\left\{\lambda \epsilon_t \mathbf{v}_t(\epsilon_{1:t-1})\right\} \times \\
&\frac{1}{2}\left(\exp\left\{\lambda \mathbf{v}_{T-1}(\epsilon_{1:T-2})\right\} \exp\left\{\frac{\lambda^2 \mathbf{v}_T((\epsilon_{1:T-2},1))^2}{2}\right\}\right. \\
&\qquad \left. + \exp\left\{-\lambda \mathbf{v}_{T-1}(\epsilon_{1:T-2})\right\} \exp\left\{\frac{\lambda^2 \mathbf{v}_T((\epsilon_{1:T-2},-1))^2}{2}\right\}\right)\Bigg]
\end{aligned}$$

Consider the term inside:

$$
\begin{aligned}
&\frac{1}{2}\left(\exp\left\{\lambda \mathbf{v}_{T-1}(\epsilon_{1:T-2})\right\}\exp\left\{\frac{\lambda^2 \mathbf{v}_T((\epsilon_{1:T-2},1))^2}{2}\right\} + \exp\left\{-\lambda \mathbf{v}_{T-1}(\epsilon_{1:T-2})\right\}\exp\left\{\frac{\lambda^2 \mathbf{v}_T((\epsilon_{1:T-2},-1))^2}{2}\right\}\right)\\
&\le \max_{\epsilon_{T-1}}\left(\exp\left\{\frac{\lambda^2 \mathbf{v}_T((\epsilon_{1:T-2},\epsilon_{T-1}))^2}{2}\right\}\right)\frac{\exp\left\{\lambda \mathbf{v}_{T-1}(\epsilon_{1:T-2})\right\}+\exp\left\{-\lambda \mathbf{v}_{T-1}(\epsilon_{1:T-2})\right\}}{2}\\
&\le \max_{\epsilon_{T-1}}\left(\exp\left\{\frac{\lambda^2 \mathbf{v}_T((\epsilon_{1:T-2},\epsilon_{T-1}))^2}{2}\right\}\right)\exp\left\{\frac{\lambda^2 \mathbf{v}_{T-1}(\epsilon_{1:T-2})^2}{2}\right\}\\
&= \exp\left\{\frac{\lambda^2 \max_{\epsilon_{T-1}\in\{\pm 1\}}\left(\mathbf{v}_{T-1}(\epsilon_{1:T-2})^2 + \mathbf{v}_T(\epsilon_{1:T-1})^2\right)}{2}\right\}
\end{aligned}
$$

Repeating the last steps, we show that for any i,

$$
M(\lambda) \le \sum_{\mathbf{v}\in V}\mathbb{E}_{\epsilon_1,\ldots,\epsilon_{i-1}}\left[\prod_{t=1}^{i-1}\exp\left\{\lambda\epsilon_t \mathbf{v}_t(\epsilon_{1:t-1})\right\}\times\exp\left\{\frac{\lambda^2 \max_{\epsilon_i\ldots\epsilon_{T-1}\in\{\pm1\}}\sum_{t=i}^{T}\mathbf{v}_t(\epsilon_{1:t-1})^2}{2}\right\}\right]
$$

We arrive at

$$
\begin{aligned}
M(\lambda) &\le \sum_{\mathbf{v}\in V}\exp\left\{\frac{\lambda^2 \max_{\epsilon_1\ldots\epsilon_{T-1}\in\{\pm1\}}\sum_{t=1}^{T}\mathbf{v}_t(\epsilon_{1:t-1})^2}{2}\right\}\\
&\le |V|\exp\left\{\frac{\lambda^2 \max_{\mathbf{v}\in V}\max_{\epsilon\in\{\pm1\}^T}\sum_{t=1}^{T}\mathbf{v}_t(\epsilon)^2}{2}\right\}
\end{aligned}
$$

Taking logarithms on both sides, dividing by λ and setting $\lambda = \sqrt{\frac{2\log(|V|)}{\max_{\mathbf{v}\in V}\max_{\epsilon\in\{\pm1\}^T}\sum_{t=1}^{T}\mathbf{v}_t(\epsilon)^2}}$ we conclude that

$$
\mathbb{E}_{\epsilon_1,\ldots,\epsilon_T}\left[\max_{\mathbf{v}\in V}\sum_{t=1}^{T}\epsilon_t\mathbf{v}_t(\epsilon)\right] \le \sqrt{2\log(|V|)\max_{\mathbf{v}\in V}\max_{\epsilon\in\{\pm1\}^T}\sum_{t=1}^{T}\mathbf{v}_t(\epsilon)^2}
$$

□

Proof of Lemma 4. We prove the first inequality. Let $\{\mathbf{w}^1,\ldots,\mathbf{w}^M\}$ be a largest strongly 2α-separated set of $\mathcal{F}(\mathbf{x})$ with $M = \mathcal{M}_p(2\alpha,\mathcal{F},\mathbf{x})$. Let $\{\mathbf{v}^1,\ldots,\mathbf{v}^N\}$ be a smallest α-cover of $\mathcal{F}$ on $\mathbf{x}$ with $N = \mathcal{N}_p(\alpha,\mathcal{F},\mathbf{x})$. For the sake of contradiction, assume $M > N$. Consider a path $\epsilon \in \{\pm1\}^T$ on which all the trees $\{\mathbf{w}^1,\ldots,\mathbf{w}^M\}$ are (2α)-separated. By the definition of a cover, for any $\mathbf{w}^i$ there exists a tree $\mathbf{v}^j$ such that

$$
\left(\frac{1}{T}\sum_{t=1}^{T}|\mathbf{v}_t^j(\epsilon)-\mathbf{w}_t^i(\epsilon)|^p\right)^{1/p} \le \alpha.
$$

Since $M > N$, there must exist distinct $\mathbf{w}^i$ and $\mathbf{w}^k$, for which the covering tree $\mathbf{v}^j$ is the same for the given path ϵ. By triangle inequality,

$$
\left(\frac{1}{T}\sum_{t=1}^{T}|\mathbf{w}_t^i(\epsilon)-\mathbf{w}_t^k(\epsilon)|^p\right)^{1/p} \le 2\alpha,
$$

which is a contradiction. We conclude that $M \le N$.

Now, we prove the second inequality. Consider a maximal α-packing $V \subseteq \mathcal{F}(\mathbf{x})$ of size $\mathcal{D}_p(\alpha,\mathcal{F},\mathbf{x})$. Since this is a *maximal* α-packing, for any $f \in \mathcal{F}$, there is no path on which $f(\mathbf{x})$ is α-separated from every member of the packing. In other words, for every path $\epsilon \in \{\pm 1\}^T$, there is a member of the packing $\mathbf{v} \in V$ such that

$$\left(\frac{1}{T}\sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - f(\mathbf{x}_t(\epsilon))|^p\right)^{1/p} \leq \alpha$$

which means that the packing V is a cover. □

Proof of Theorem 5. For any $d \geq 0$ and $T \geq 0$, define the function

$$g_k(d,T) = \sum_{i=0}^{d} \binom{T}{i} k^i.$$

It is not difficult to verify that this function satisfies the recurrence

$$g_k(d,T) = g_k(d,T-1) + kg_k(d-1,T-1)$$

for all $d, T \geq 1$. To visualize this recursion, consider a $k \times T$ matrix and ask for ways to choose at most d columns followed by a choice among the k rows for each chosen column. The task can be decomposed into (a) making the d column choices out of the first $T-1$ columns, followed by picking rows (there are $g_k(d,T-1)$ ways to do it) or (b) choosing $d-1$ columns (followed by row choices) out of the first $T-1$ columns and choosing a row for the Tth column (there are $kg_k(d-1,T-1)$ ways to do it). This gives the recursive formula.

In what follows, we shall refer to an L_∞ cover at scale $1/2$ simply as a $1/2$-cover. The theorem claims that the size of a minimal $1/2$-cover is at most $g_k(d,T)$. The proof proceeds by induction on $T+d$.

Base: For $d=1$ and $T=1$, there is only one node in the tree, i.e. the tree is defined by the constant $\mathbf{x}_1 \in \mathcal{X}$. Functions in $\mathcal{F}$ can take up to $k+1$ values on $\mathbf{x}_1$, i.e. $\mathcal{N}(0,\mathcal{F},1) \leq k+1$ (and, thus, also for the $1/2$-cover). Using the convention $\binom{T}{0} = 1$, we indeed verify that $g_k(1,1) = 1+k = k+1$. The same calculation gives the base case for $T=1$ and any $d \in \mathbb{N}$. Furthermore, for any $T \in \mathbb{N}$ if $d=0$, then there is no point which is 2-shattered by $\mathcal{F}$. This means that functions in $\mathcal{F}$ differ by at most 1 on any point of $\mathcal{X}$. Thus, there is a $1/2$ cover of size $1 = g_k(0,T)$, verifying this base case.

Induction step: Suppose by the way of induction that the statement holds for $(d, T-1)$ and $(d-1, T-1)$. Consider any tree $\mathbf{x}$ of depth T with $\mathrm{fat}_2(\mathcal{F},\mathbf{x}) = d$. Define the partition $\mathcal{F} = \mathcal{F}_0 \cup \ldots \cup \mathcal{F}_k$ with $\mathcal{F}_i = \{f \in \mathcal{F} : f(\mathbf{x}_1) = i\}$ for $i \in \{0,\ldots,k\}$, where $\mathbf{x}_1$ is the root of $\mathbf{x}$. Let $n = |\{i : \mathrm{fat}_2(\mathcal{F}_i,\mathbf{x}) = d\}|$.

Suppose first, for the sake of contradiction, that $\mathrm{fat}_2(\mathcal{F}_i,\mathbf{x}) = \mathrm{fat}_2(\mathcal{F}_j,\mathbf{x}) = d$ for $|i-j| \geq 2$. Then there exist two trees $\mathbf{z}$ and $\mathbf{v}$ of depth d which are 2-shattered by $\mathcal{F}_i$ and $\mathcal{F}_j$, respectively, and with $\mathrm{Img}(\mathbf{z}), \mathrm{Img}(\mathbf{v}) \subseteq \mathrm{Img}(\mathbf{x})$. Since functions within each subset $\mathcal{F}_i$ take on the same values on $\mathbf{x}_1$, we conclude that $\mathbf{x}_1 \notin \mathrm{Img}(\mathbf{z}), \mathbf{x}_1 \notin \mathrm{Img}(\mathbf{v})$. This follows immediately from the definition of shattering. We now *join* the two shattered $\mathbf{z}$ and $\mathbf{v}$ trees with $\mathbf{x}_1$ at the root and observe that $\mathcal{F}_i \cup \mathcal{F}_j$ 2-shatters this resulting tree of depth $d+1$, which is a contradiction. Indeed, the witness $\mathbb{R}$-valued tree $\mathbf{s}$ is constructed by joining the two witnesses for the 2-shattered trees $\mathbf{z}$ and $\mathbf{v}$ and by defining the root as $\mathbf{s}_1 = (i+j)/2$. It is easy to see that $\mathbf{s}$ is a witness to the shattering. Given any $\epsilon \in \{\pm 1\}^{d+1}$, there is a function $f^i \in \mathcal{F}_i$ which realizes the desired separation under the signs $(\epsilon_2,\ldots,\epsilon_{d+1})$ for the tree $\mathbf{z}$ and there is a function $f^j \in \mathcal{F}_j$ which does the same for $\mathbf{v}$. Depending on $\epsilon_1 = +1$ or $\epsilon_1 = -1$, either f^i or f^j realize the separation over ϵ.

We conclude that the number of subsets of $\mathcal{F}$ with fat-shattering dimension equal to d cannot be more than two (for otherwise at least two indices will be separated by 2 or more). We have three cases: $n=0$, $n=1$, or $n=2$, and in the last case it must be that the indices of the two subsets differ by 1.

First, consider any $\mathcal{F}_i$ with $\mathrm{fat}_2(\mathcal{F}_i,\mathbf{x}) \leq d-1$, $i \in \{0,\ldots,k\}$. By induction, there are $1/2$-covers V^ℓ and V^r of $\mathcal{F}_i$ on the subtrees $\mathbf{x}^\ell$ and $\mathbf{x}^r$, respectively, both of size at most $g_k(d-1,T-1)$. Informally, out of these $1/2$-covers we can create a $1/2$-cover V for $\mathcal{F}_i$ on $\mathbf{x}$ by pairing the $1/2$-covers in V^ℓ and V^r. The resulting cover of $\mathcal{F}_i$ will be of size $g_k(d-1, T-1)$. Formally, consider a set of pairs $(\mathbf{v}^\ell, \mathbf{v}^r)$ of trees, with $\mathbf{v}^\ell \in V^\ell$, $\mathbf{v}^r \in V^r$ and such that each tree in V^ℓ and V^r appears in at least one of the pairs. Clearly, this can be done using at most $g_k(d-1, T-1)$ pairs, and such a pairing is not unique. We join the subtrees in every pair $(\mathbf{v}^\ell, \mathbf{v}^r)$ with a constant i as the root, thus creating a set V of trees, $|V| \leq g_k(d-1, T-1)$. We claim that V is a 1/2-cover for $\mathcal{F}_i$ on $\mathbf{x}$. Note that all the functions in $\mathcal{F}_i$ take on the same value i on $\mathbf{x}_1$ and by construction $\mathbf{v}_1 = i$ for any $\mathbf{v} \in V$. Now, consider any $f \in \mathcal{F}_i$ and $\epsilon \in \{\pm 1\}^T$. Without loss of generality, assume $\epsilon_1 = -1$. By assumption, there is a $\mathbf{v}^\ell \in V^\ell$ such that $|\mathbf{v}^\ell_t(\epsilon_{2:T}) - f(\mathbf{x}_{t+1}(\epsilon_{1:T}))| \leq 1/2$ for any $t \in [T-1]$. By construction $\mathbf{v}^\ell$ appears as a left subtree of at least one tree in V, which, therefore, matches the values of f for $\epsilon_{1:T}$. The same argument holds for $\epsilon_1 = +1$ by finding an appropriate subtree in V^r. We conclude that V is a 1/2-cover of $\mathcal{F}_i$ on $\mathbf{x}$, and this holds for any $i \in \{0, \ldots, k\}$ with $\mathrm{fat}_2(\mathcal{F}_i, \mathbf{x}) \leq d-1$. Therefore, the total size of a 1/2-cover for the union $\cup_{i:\mathrm{fat}_2(\mathcal{F}_i,\mathbf{x}) \leq d-1} \mathcal{F}_i$ is at most $(k+1-n) g_k(d-1, T-1)$. If $n = 0$, the induction step is proven because $g_k(d-1, T-1) \leq g_k(d, T-1)$ and so the total size of the constructed cover is at most

$$(k+1) g_k(d-1, T-1) \leq g_k(d, T-1) + k g_k(d-1, T-1) = g_k(d, T).$$

Now, consider the case $n = 1$ and let $\mathrm{fat}_2(\mathcal{F}_i, \mathbf{x}) = d$. An argument exactly as above yields a 1/2-cover for $\mathcal{F}_i$, and this cover is of size at most $g_k(d, T-1)$ by induction. The total 1/2-cover is therefore of size at most

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

Lastly, for $n = 2$, suppose $\mathrm{fat}_2(\mathcal{F}_i, \mathbf{x}) = \mathrm{fat}_2(\mathcal{F}_j, \mathbf{x}) = d$ for $|i - j| = 1$. Let $\mathcal{F}' = \mathcal{F}_i \cup \mathcal{F}_j$. Note that $\mathrm{fat}_2(\mathcal{F}', \mathbf{x}) = d$. Just as before, the 1/2-covering for $\mathbf{x}$ can be constructed by considering the 1/2-covers for the two subtrees. However, when joining any $(\mathbf{v}^\ell, \mathbf{v}^r)$, we take $(i+j)/2$ as the root. It is straightforward to check that the resulting cover is indeed an 1/2-cover, thanks to the relation $|i - j| = 1$. The size of the constructed cover is, by induction, $g_k(d, T-1)$, and the induction step follows. This concludes the induction proof, yielding the main statement of the theorem.

Finally, the upper bound on $g_k(d, T)$ is

$$\sum_{i=1}^{d} \binom{T}{i} k^i \leq \left(\frac{kT}{d}\right)^d \sum_{i=1}^{d} \binom{T}{i} \left(\frac{d}{T}\right)^i \leq \left(\frac{kT}{d}\right)^d \left(1 + \frac{d}{T}\right)^T \leq \left(\frac{ekT}{d}\right)^d$$

whenever $T \geq d$.

□

Proof of Theorem 7. The proof is very close to the proof of Theorem 5, with a few key differences. As before, for any $d \geq 0$ and $T \geq 0$, define the function $g_k(d, T) = \sum_{i=0}^{d} \binom{T}{i} k^i$.

The theorem claims that the size of a minimal 0-cover is at most $g_k(d, T)$. The proof proceeds by induction on $T + d$.

Base: For $d = 1$ and $T = 1$, there is only one node in the tree, i.e. the tree is defined by the constant $\mathbf{x}_1 \in \mathcal{X}$. Functions in $\mathcal{F}$ can take up to $k+1$ values on $\mathbf{x}_1$, i.e. $\mathcal{N}(0, \mathcal{F}, 1) \leq k+1$. Using the convention $\binom{T}{0} = 1$, we indeed verify that $g_k(1, 1) = 1 + k = k + 1$. The same calculation gives the base case for $T = 1$ and any $d \in \mathbb{N}$. Furthermore, for any $T \in \mathbb{N}$ if $d = 0$, then there is no point which is 1-shattered by $\mathcal{F}$. This means that all functions in $\mathcal{F}$ are identical, proving that there is a 0-cover of size $1 = g_k(0, T)$.

Induction step: Suppose by the way of induction that the statement holds for $(d, T-1)$ and $(d-1, T-1)$. Consider any tree $\mathbf{x}$ of depth T with $\mathrm{fat}_1(\mathcal{F}, \mathbf{x}) = d$. Define the partition $\mathcal{F} = \mathcal{F}_0 \cup \ldots \cup \mathcal{F}_k$ with $\mathcal{F}_i = \{f \in \mathcal{F} : f(\mathbf{x}_1) = i\}$ for $i \in \{0, \ldots, k\}$, where $\mathbf{x}_1$ is the root of $\mathbf{x}$.

We first argue that $\mathrm{fat}_1(\mathcal{F}_i, \mathbf{x}) = d$ for at most one value $i \in \{0, \ldots, k\}$. By the way of contradiction, suppose we do have $\mathrm{fat}_1(\mathcal{F}_i, \mathbf{x}) = \mathrm{fat}_1(\mathcal{F}_j, \mathbf{x}) = d$ for $i \neq j$. Then there exist two trees $\mathbf{z}$ and $\mathbf{v}$ of depth d 1-shattered by $\mathcal{F}_i$ and $\mathcal{F}_j$, respectively, and with $\mathrm{Img}(\mathbf{z}), \mathrm{Img}(\mathbf{v}) \subseteq \mathrm{Img}(\mathbf{x})$. Since functions within each subset $\mathcal{F}_i$ take on the same values on $\mathbf{x}_1$, we conclude that $\mathbf{x}_1 \notin \mathrm{Img}(\mathbf{z}), \mathbf{x}_1 \notin \mathrm{Img}(\mathbf{v})$. This follows immediately from the definition of shattering. We now *join* the two shattered $\mathbf{z}$ and $\mathbf{v}$ trees with $\mathbf{x}_1$ at the root and observe that $\mathcal{F}_i \cup \mathcal{F}_j$ 1-shatters this resulting tree of depth $d+1$, which is a contradiction. Indeed, the witness $\mathbb{R}$-valued tree $\mathbf{s}$ is constructed by joining the two witnesses for the 1-shattered trees $\mathbf{z}$ and $\mathbf{v}$ and by defining the root as $\mathbf{s}_1 = (i+j)/2$. It is easy to see that $\mathbf{s}$ is a witness to the shattering. Given any $\epsilon \in \{\pm 1\}^{d+1}$, there is a function $f^i \in \mathcal{F}_i$ which realizes the desired separation under the signs $(\epsilon_2, \ldots, \epsilon_{d+1})$ for the tree $\mathbf{z}$ and there is a function $f^j \in \mathcal{F}_j$ which does the same for $\mathbf{v}$. Depending on $\epsilon_1 = +1$ or $\epsilon_1 = -1$, either f^i or f^j realize the separation over ϵ.

We conclude that $\mathrm{fat}_1(\mathcal{F}_i, \mathbf{x}) = d$ for at most one $i \in \{0, \ldots, k\}$. Without loss of generality, assume $\mathrm{fat}_1(\mathcal{F}_0, \mathbf{x}) \le d$ and $\mathrm{fat}_1(\mathcal{F}_i, \mathbf{x}) \le d-1$ for $i \in \{1, \ldots, k\}$. By induction, for any $\mathcal{F}_i$, $i \in \{1, \ldots, k\}$, there are 0-covers V^ℓ and V^r of $\mathcal{F}_i$ on the subtrees $\mathbf{x}^\ell$ and $\mathbf{x}^r$, respectively, both of size at most $g_k(d-1, T-1)$. Out of these 0-covers we can create a 0-cover V for $\mathcal{F}_i$ on $\mathbf{x}$ by pairing the 0-covers in V^ℓ and V^r. Formally, consider a set of pairs $(\mathbf{v}^\ell, \mathbf{v}^r)$ of trees, with $\mathbf{v}^\ell \in V^\ell$, $\mathbf{v}^r \in V^r$ and such that each tree in V^ℓ and V^r appears in at least one of the pairs. Clearly, this can be done using at most $g_k(d-1, T-1)$ pairs, and such a pairing is not unique. We join the subtrees in every pair $(\mathbf{v}^\ell, \mathbf{v}^r)$ with a constant i as the root, thus creating a set V of trees, $|V| \le g_k(d-1, T-1)$. We claim that V is a 0-cover for $\mathcal{F}_i$ on $\mathbf{x}$. Note that all the functions in $\mathcal{F}_i$ take on the same value i on $\mathbf{x}_1$ and by construction $\mathbf{v}_1 = i$ for any $\mathbf{v} \in V$. Now, consider any $f \in \mathcal{F}_i$ and $\epsilon \in \{\pm 1\}^T$. Without loss of generality, assume $\epsilon_1 = -1$. By assumption, there is a $\mathbf{v}^\ell \in V^\ell$ such that $\mathbf{v}^\ell_t(\epsilon_{2:T}) = f(\mathbf{x}_{t+1}(\epsilon_{1:T}))$ for any $t \in [T-1]$. By construction $\mathbf{v}^\ell$ appears as a left subtree of at least one tree in V, which, therefore, matches the values of f for $\epsilon_{1:T}$. The same argument holds for $\epsilon_1 = +1$ by finding an appropriate subtree in V^r. We conclude that V is a 0-cover of $\mathcal{F}_i$ on $\mathbf{x}$, and this holds for any $i \in \{1, \ldots, k\}$.

Therefore, the total size of a 0-cover for $\mathcal{F}_1 \cup \ldots \cup F_k$ is at most $k g_k(d-1, T-1)$. A similar argument yields a 0-cover for $\mathcal{F}_0$ on $\mathbf{x}$ of size at most $g_k(d, T-1)$ by induction. Thus, the size of the resulting 0-cover of $\mathcal{F}$ on $\mathbf{x}$ is at most

$$g_k(d, T-1) + k g_k(d-1, T-1) = g_k(d, T),$$

completing the induction step and yielding the main statement of the theorem.

The upper bound on $g_k(d, T)$ appears in the proof of Theorem 5.

□

Proof of Corollary 6. The first two inequalities follow by simple comparison of norms. It remains to prove the bound for the ℓ_∞ covering. For any $\alpha > 0$ define an α-discretization of the $[-1, 1]$ interval as $B_\alpha = \{-1 + \alpha/2, -1 + 3\alpha/2, \ldots, -1 + (2k+1)\alpha/2, \ldots\}$ for $0 \le k$ and $(2k+1)\alpha \le 4$. Also for any $a \in [-1, 1]$ define $\lfloor a \rfloor_\alpha = \operatorname*{argmin}_{r \in B_\alpha} |r - a|$ with ties being broken by choosing the smaller discretization point. For a function $f : \mathcal{X} \mapsto [-1, 1]$ let the function $\lfloor f \rfloor_\alpha$ be defined pointwise as $\lfloor f(x) \rfloor_\alpha$, and let $\lfloor \mathcal{F} \rfloor_\alpha = \{\lfloor f \rfloor_\alpha : f \in \mathcal{F}\}$. First, we prove that $\mathcal{N}_\infty(\alpha, \mathcal{F}, \mathbf{x}) \le \mathcal{N}_\infty(\alpha/2, \lfloor \mathcal{F} \rfloor_\alpha, \mathbf{x})$. Indeed, suppose the set of trees V is a minimal $\alpha/2$-cover of $\lfloor \mathcal{F} \rfloor_\alpha$ on $\mathbf{x}$. That is,

$$\forall f_\alpha \in \lfloor \mathcal{F} \rfloor_\alpha,\ \forall \epsilon \in \{\pm 1\}^T\ \exists \mathbf{v} \in V \text{ s.t. } \quad |\mathbf{v}_t(\epsilon) - f_\alpha(\mathbf{x}_t(\epsilon))| \le \alpha/2$$

Pick any $f \in \mathcal{F}$ and let $f_\alpha = \lfloor f \rfloor_\alpha$. Then $\|f - f_\alpha\|_\infty \le \alpha/2$. Then for all $\epsilon \in \{\pm 1\}^T$ and any $t \in [T]$

$$|f(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon)| \le |f(\mathbf{x}_t(\epsilon)) - f_\alpha(\mathbf{x}_t(\epsilon))| + |f_\alpha(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon)| \le \alpha,$$

and so V also provides an L_∞ cover at scale α.

We conclude that $\mathcal{N}_\infty(\alpha, \mathcal{F}, \mathbf{x}) \le \mathcal{N}_\infty(\alpha/2, \lfloor \mathcal{F} \rfloor_\alpha, \mathbf{x}) = \mathcal{N}_\infty(1/2, \mathcal{G}, \mathbf{x})$ where $G = \frac{1}{\alpha} \lfloor \mathcal{F} \rfloor_\alpha$. The functions of $\mathcal{G}$ take on a discrete set of at most $\lfloor 2/\alpha \rfloor + 1$ values. Obviously, by adding a constant to all the functions in $\mathcal{G}$, we can make the set of values to be $\{0, \ldots, \lfloor 2/\alpha \rfloor\}$. We now apply Theorem 5 with an upper bound $\sum_{i=0}^d \binom{T}{i} k^i \le (ekT)^d$ which holds for any $T > 0$. This yields $\mathcal{N}_\infty(1/2, \mathcal{G}, \mathbf{x}) \le (2eT/\alpha)^{\mathrm{fat}_2(\mathcal{G})}$.

It remains to prove $\mathrm{fat}_2(\mathcal{G}) \le \mathrm{fat}_\alpha(\mathcal{F})$, or, equivalently (by scaling) $\mathrm{fat}_{2\alpha}(\lfloor \mathcal{F} \rfloor_\alpha) \le \mathrm{fat}_\alpha(\mathcal{F})$. To this end, suppose there exists an $\mathbb{R}$-valued tree $\mathbf{x}$ of depth $d = \mathrm{fat}_{2\alpha}(\lfloor \mathcal{F} \rfloor_\alpha)$ such that there is an witness tree $\mathbf{s}$ with

$$\forall \epsilon \in \{\pm 1\}^d, \ \exists f_\alpha \in \lfloor \mathcal{F} \rfloor_\alpha \quad \text{s.t. } \forall t \in [d], \ \epsilon_t (f_\alpha(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) \ge \alpha$$

Using the fact that for any $f \in \mathcal{F}$ and $f_\alpha = \lfloor f \rfloor_\alpha$ we have $\|f - f_\alpha\|_\infty \le \alpha/2$, it follows that

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

That is, $\mathbf{s}$ is a witness to α-shattering by $\mathcal{F}$. Thus for any $\mathbf{x}$,

$$\mathcal{N}_\infty(\alpha, \mathcal{F}, \mathbf{x}) \le \mathcal{N}_\infty(\alpha/2, \lfloor \mathcal{F} \rfloor_\alpha, \mathbf{x}) \le \left(\frac{2eT}{\alpha}\right)^{\mathrm{fat}_{2\alpha}(\lfloor \mathcal{F} \rfloor_\alpha)} \le \left(\frac{2eT}{\alpha}\right)^{\mathrm{fat}_\alpha(\mathcal{F})}$$

□

Proof of Theorem 9. Define $\beta_0 = 1$ and $\beta_j = 2^{-j}$. For a fixed tree $\mathbf{x}$ of depth T, let V_j be an ℓ_2-cover at scale β_j. For any path $\epsilon \in \{\pm 1\}^T$ and any $f \in \mathcal{F}$, let $\mathbf{v}[f,\epsilon]^j \in V_j$ the element of the cover such that

$$\sqrt{\frac{1}{T}\sum_{t=1}^T |\mathbf{v}[f,\epsilon]_t^j(\epsilon) - f(\mathbf{x}_t(\epsilon))|^2} \le \beta_j$$

By the definition such a $\mathbf{v}[f,\epsilon]^j \in V_j$ exists, and we assume for simplicity this element is unique (ties can be broken in an arbitrary manner). Thus, $f \mapsto \mathbf{v}[f,\epsilon]^j$ is a well-defined mapping for any fixed ϵ and j. As before, $\mathbf{v}[f,\epsilon]_t^j$ denotes the t-th mapping of $\mathbf{v}[f,\epsilon]^j$. For any $t \in [T]$, we have

$$f(\mathbf{x}_t(\epsilon)) = f(\mathbf{x}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t^N(\epsilon) + \sum_{j=1}^N (\mathbf{v}[f,\epsilon]_t^j(\epsilon) - \mathbf{v}[f,\epsilon]_t^{j-1}(\epsilon))$$

where $\mathbf{v}[f,\epsilon]_t^0(\epsilon) = 0$. Hence,

$$\begin{aligned}
\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^T \epsilon_t f(\mathbf{x}_t(\epsilon))\right] &= \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^T \epsilon_t \left(f(\mathbf{x}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t^N(\epsilon) + \sum_{j=1}^N (\mathbf{v}[f,\epsilon]_t^j(\epsilon) - \mathbf{v}[f,\epsilon]_t^{j-1}(\epsilon))\right)\right] \\
&= \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^T \epsilon_t \left(f(\mathbf{x}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t^N(\epsilon)\right) + \sum_{t=1}^T \epsilon_t \left(\sum_{j=1}^N (\mathbf{v}[f,\epsilon]_t^j(\epsilon) - \mathbf{v}[f,\epsilon]_t^{j-1}(\epsilon))\right)\right] \\
&\le \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^T \epsilon_t \left(f(\mathbf{x}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t^N(\epsilon)\right)\right] + \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^T \epsilon_t \left(\sum_{j=1}^N (\mathbf{v}[f,\epsilon]_t^j(\epsilon) - \mathbf{v}[f,\epsilon]_t^{j-1}(\epsilon))\right)\right]
\end{aligned} \tag{9}$$

The first term above can be bounded via the Cauchy-Schwarz inequality as

$$\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^T \epsilon_t \left(f(\mathbf{x}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t^N(\epsilon)\right)\right] \le T\ \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^T \frac{\epsilon_t}{\sqrt{T}} \frac{\left(f(\mathbf{x}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t^N(\epsilon)\right)}{\sqrt{T}}\right] \le T\ \beta_N.$$

The second term in (9) is bounded by considering successive refinements of the cover. The argument, however, is more delicate than in the classical case, as the trees $\mathbf{v}[f,\epsilon]^j$, $\mathbf{v}[f,\epsilon]^{j-1}$ depend on the particular path. Consider all possible pairs of $\mathbf{v}^s \in V_j$ and $\mathbf{v}^r \in V_{j-1}$, for $1 \le s \le |V_j|$, $1 \le r \le |V_{j-1}|$, where we assumed an arbitrary enumeration of elements. For each pair $(\mathbf{v}^s, \mathbf{v}^r)$, define a real-valued tree $\mathbf{w}^{(s,r)}$ by

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

for all $t \in [T]$ and $\epsilon \in \{\pm 1\}^T$. It is crucial that $\mathbf{w}^{(s,r)}$ can be non-zero only on those paths ϵ for which $\mathbf{v}^s$ and $\mathbf{v}^r$ are indeed the members of the covers (at successive resolutions) close to $f(\mathbf{x}(\epsilon))$ (in the ℓ_2 sense) *for some* $f \in \mathcal{F}$. It is easy to see that $\mathbf{w}^{(s,r)}$ is well-defined. Let the set of trees W_j be defined as

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

Now, the second term in (9) can be written as

$$\begin{aligned}
\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \epsilon_t \sum_{j=1}^N (\mathbf{v}[f,\epsilon]^j_t(\epsilon) - \mathbf{v}[f,\epsilon]^{j-1}_t(\epsilon)) \right] &\le \sum_{j=1}^N \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \epsilon_t (\mathbf{v}[f,\epsilon]^j_t(\epsilon) - \mathbf{v}[f,\epsilon]^{j-1}_t(\epsilon)) \right] \\
&\le \sum_{j=1}^N \mathbb{E}_\epsilon \left[\max_{\mathbf{w} \in W_j} \sum_{t=1}^T \epsilon_t \mathbf{w}_t(\epsilon) \right]
\end{aligned}$$

The last inequality holds because for any $j \in [N]$, $\epsilon \in \{\pm 1\}^T$ and $f \in \mathcal{F}$ there is some $\mathbf{w}^{(s,r)} \in W_j$ with $\mathbf{v}[f,\epsilon]^j = \mathbf{v}^s$, $\mathbf{v}[f,\epsilon]^{j-1} = \mathbf{v}^r$ and

$$\mathbf{v}^s_t(\epsilon) - \mathbf{v}^r_t(\epsilon) = \mathbf{w}^{(s,r)}_t(\epsilon) \quad \forall t \le T.$$

Clearly, $|W_j| \le |V_j| \cdot |V_{j-1}|$. To invoke Lemma 8, it remains to bound the magnitude of all $\mathbf{w}^{(s,r)} \in W_j$ along all paths. For this purpose, fix $\mathbf{w}^{(s,r)}$ and a path ϵ. If there exists $f \in \mathcal{F}$ for which $\mathbf{v}^s = \mathbf{v}[f,\epsilon]^j$ and $\mathbf{v}^r = \mathbf{v}[f,\epsilon]^{j-1}$, then $\mathbf{w}^{(s,r)}_t(\epsilon) = \mathbf{v}[f,\epsilon]^j_t - \mathbf{v}[f,\epsilon]^{j-1}_t$ for any $t \in [T]$. By triangle inequality

$$\sqrt{\sum_{t=1}^T \mathbf{w}^{(s,r)}_t(\epsilon)^2} \le \sqrt{\sum_{t=1}^T (\mathbf{v}[f,\epsilon]^j_t(\epsilon) - f(\mathbf{x}_t(\epsilon)))^2} + \sqrt{\sum_{t=1}^T (\mathbf{v}[f,\epsilon]^{j-1}_t(\epsilon) - f(\mathbf{x}_t(\epsilon)))^2} \le \sqrt{T}(\beta_j + \beta_{j-1}) = 3\sqrt{T}\beta_j.$$

If there exists no such $f \in \mathcal{F}$ for the given ϵ and (s,r), then $\mathbf{w}^{(s,r)}_t(\epsilon)$ is zero for all $t \ge t_o$, for some $1 \le t_o < T$, and thus

$$\sqrt{\sum_{t=1}^T \mathbf{w}^{(s,r)}_t(\epsilon)^2} \le \sqrt{\sum_{t=1}^T \mathbf{w}^{(s,r)}_t(\epsilon')^2}$$

for any other path ϵ' which agrees with ϵ up to t_o. Hence, the bound

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

holds for all $\epsilon \in \{\pm 1\}^T$ and all $\mathbf{w}^{(s,r)} \in W_j$.

Now, back to (9), we put everything together and apply Lemma 8:

$$\begin{aligned}
\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \epsilon_t f(\mathbf{x}_t(\epsilon)) \right] &\le T\ \beta_N + \sqrt{T}\ \sum_{j=1}^N 3\beta_j \sqrt{2\log(|V_j|\ |V_{j-1}|)} \\
&\le T\ \beta_N + \sqrt{T}\ \sum_{j=1}^N 6\beta_j \sqrt{\log(|V_j|)} \\
&\le T\ \beta_N + 12\ \sqrt{T}\ \sum_{j=1}^N (\beta_j - \beta_{j+1}) \sqrt{\log \mathcal{N}_2(\beta_j, \mathcal{F}, \mathbf{x})} \\
&\le T\ \beta_N + 12 \int_{\beta_{N+1}}^{\beta_0} \sqrt{T\ \log\ \mathcal{N}_2(\delta, \mathcal{F}, \mathbf{x})}\, d\delta
\end{aligned}$$

where the last but one step is because $2(\beta_j - \beta_{j+1}) = \beta_j$. Now for any $\alpha > 0$, pick $N = \sup\{j : \beta_j > 2\alpha\}$. In this case we see that by our choice of N, $\beta_{N+1} \le 2\alpha$ and so $\beta_N = 2\beta_{N+1} \le 4\alpha$. Also note that since $\beta_N > 2\alpha$, $\beta_{N+1} = \frac{\beta_N}{2} > \alpha$. Hence we conclude that

$$\mathfrak{R}_T(\mathcal{F}) \le \inf_{\alpha} \left\{ 4T\alpha + 12 \int_{\alpha}^{1} \sqrt{T \ \log \ \mathcal{N}_2(\delta, \mathcal{F}, T)}\, d\delta \right\}$$

□

Proof of Theorem 10. Let $(x'_1, \ldots, x'_T)$ be a sequence tangent to $(x_1, \ldots, x_T)$. Recall the notation $\mathbb{E}_{t-1}\left[f(x'_t)\right] = \mathbb{E}\left\{f(x'_t)|x_1, \ldots, x_{t-1}\right\}$. By Chebychev's inequality, for any $f \in \mathcal{F}$,

$$\begin{aligned}
\mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\left|\sum_{t=1}^{T}\left(f(x'_t) - \mathbb{E}_{t-1}\left[f(x'_t)\right]\right)\right| > \alpha/2 \middle| \, x_1, \ldots, x_T\right] &\le \frac{\mathbb{E}\left[\left(\sum_{t=1}^{T}\left(f(x'_t) - \mathbb{E}_{t-1}\left[f(x'_t)\right]\right)\right)^2 \middle| x_1, \ldots, x_T\right]}{T^2\alpha^2/4} \\
&= \frac{\sum_{t=1}^{T}\mathbb{E}\left[\left(f(x'_t) - \mathbb{E}_{t-1}\left[f(x'_t)\right]\right)^2 \middle| x_1, \ldots, x_T\right]}{T^2\alpha^2/4} \\
&\le \frac{4T}{T^2\alpha^2/4} = \frac{16}{T\alpha^2}.
\end{aligned}$$

The second step is due to the fact that the cross terms are zero:

$$\mathbb{E}\left\{\left(f(x'_t) - \mathbb{E}_{t-1}\left[f(x'_t)\right]\right)\left(f(x'_s) - \mathbb{E}_{s-1}\left[f(x'_s)\right]\right) \middle| x_1, \ldots, x_T\right\} = 0 \ .$$

Hence

$$\inf_{f\in\mathcal{F}} \mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\left|\sum_{t=1}^{T}\left(f(x'_t) - \mathbb{E}_{t-1}[f(x'_t)]\right)\right| \le \alpha/2 \,\middle|\, x_1, \ldots, x_T\right] \ge 1 - \frac{16}{T\alpha^2}$$

Whenever $\alpha^2 \ge \frac{32}{T}$ we can conclude that

$$\inf_{f\in\mathcal{F}} \mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\left|\sum_{t=1}^{T}\left(f(x'_t) - \mathbb{E}_{t-1}[f(x'_t)]\right)\right| \le \alpha/2 \,\middle|\, x_1, \ldots, x_T\right] \ge \frac{1}{2}$$

Now given a fixed $x_1, ..., x_T$ let f^* be the function that maximizes $\frac{1}{T}\left|\sum_{t=1}^{T}\left(f(x_t) - \mathbb{E}_{t-1}[f(x'_t)]\right)\right|$. Note that f^* is a deterministic choice given $x_1, ..., x_T$. Hence

$$\begin{aligned}
\frac{1}{2} &\le \inf_{f\in\mathcal{F}} \mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\left|\sum_{t=1}^{T}\left(f(x'_t) - \mathbb{E}_{t-1}[f(x'_t)]\right)\right| \le \alpha/2 \,\middle|\, x_1, \ldots, x_T\right] \\
&\le \mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\left|\sum_{t=1}^{T}\left(f^*(x'_t) - \mathbb{E}_{t-1}[f^*(x'_t)]\right)\right| \le \alpha/2 \middle| \, x_1, \ldots, x_T\right]
\end{aligned}$$

Let $A = \left\{(x_1, \ldots, x_T)\middle| \frac{1}{T}\sup_{f\in\mathcal{F}}|\sum_{t=1}^{T} f(x_t) - \mathbb{E}_{t-1}\left[f(x'_t)\right]| > \alpha\right\}$. Since the above inequality holds for any $x_1, \ldots, x_T$ we can assert that

$$\frac{1}{2} \le \mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\left|\sum_{t=1}^{T}\left(f^*(x'_t) - \mathbb{E}_{t-1}[f^*(x'_t)]\right)\right| \le \alpha/2 \middle|(x_1, \ldots, x_T) \in A\right]$$

Hence we conclude that

$$
\begin{aligned}
&\frac{1}{2}\mathbb{P}_{\mathbf{D}}\left[\sup_{f\in\mathcal{F}}\frac{1}{T}\left|\sum_{t=1}^{T}\left(f(x_t)-\mathbb{E}_{t-1}[f(x'_t)]\right)\right|>\alpha\right] \\
&\quad\le \mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\left|\sum_{t=1}^{T}\left(f^*(x'_t)-\mathbb{E}_{t-1}[f^*(x'_t)]\right)\right|\le\alpha/2 \,\middle|\, (x_1,\dots,x_T)\in A\right] \\
&\qquad\qquad\times \mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\left(f(x_t)-\mathbb{E}_{t-1}[f(x'_t)]\right)\right|>\alpha\right] \\
&\quad\le \mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\left|\sum_{t=1}^{T}\left(f^*(x_t)-f^*(x'_t)\right)\right|>\alpha/2\right] \\
&\quad\le \mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\left(f(x_t)-f(x'_t)\right)\right|>\alpha/2\right]
\end{aligned}
$$

Now we apply Lemma 3 with $\phi(u)=\mathbf{1}\{u>\alpha/2\}$ and $\Delta_f(x_t,x'_t)=f(x_t)-f(x'_t)$,

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}f(x_t)-f(x'_t)\right|\ge\alpha/2\right\}\right] \\
&\qquad\le \sup_{x_1,x'_1}\left\{\mathbb{E}_{\epsilon_1}\left[\dots\sup_{x_T,x'_T}\left\{\mathbb{E}_{\epsilon_T}\left[\mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t\left(f(x_t)-f(x'_t)\right)\right|\ge\alpha/2\right\}\right]\right\}\dots\right]\right\}
\end{aligned} \tag{10}
$$

The next few steps are similar to the proof of Theorem 2. Since

$$
\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t\left(f(x_t)-f(x'_t)\right)\right|\le\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t f(x_t)\right|+\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t f(x'_t)\right|
$$

it is true that

$$
\mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t\left(f(x_t)-f(x'_t)\right)\right|\ge\alpha/2\right\}\le\mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t f(x_t)\right|\ge\alpha/4\right\}+\mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t f(x'_t)\right|\ge\alpha/4\right\}
$$

The right-hand side of Eq. (10) then splits into two equal parts:

$$
\begin{aligned}
&\sup_{x_1}\left\{\mathbb{E}_{\epsilon_1}\left[\dots\sup_{x_T}\left\{\mathbb{E}_{\epsilon_T}\left[\mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t f(x_t)\right|\ge\alpha/4\right\}\right]\right\}\dots\right]\right\} \\
&+\sup_{x'_1}\left\{\mathbb{E}_{\epsilon_1}\left[\dots\sup_{x'_T}\left\{\mathbb{E}_{\epsilon_T}\left[\mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t f(x'_t)\right|\ge\alpha/4\right\}\right]\right\}\dots\right]\right\} \\
&=2\sup_{x_1}\left\{\mathbb{E}_{\epsilon_1}\left[\dots\sup_{x_T}\left\{\mathbb{E}_{\epsilon_T}\left[\mathbf{1}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t f(x_t)\right|\ge\alpha/4\right\}\right]\right\}\dots\right]\right\}
\end{aligned}
$$

Moving to the tree representation,

$$
\begin{aligned}
\mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\left(f(x_t)-f(x'_t)\right)\right|>\alpha/2\right]&\le2\sup_{\mathbf{x}}\mathbb{E}_{\epsilon}\left[\mathbf{1}\left\{\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t f(\mathbf{x}_t(\epsilon))\right|>\alpha/4\right\}\right] \\
&=2\sup_{\mathbf{x}}\mathbb{P}_{\epsilon}\left[\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t f(\mathbf{x}_t(\epsilon))\right|>\alpha/4\right]
\end{aligned}
$$

We can now conclude that

$$\mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\left(f(x_t)-\mathbb{E}_{t-1}[f(x_t)]\right)\right|>\alpha\right]\le 4\sup_{\mathbf{x}}\ \mathbb{P}_{\epsilon}\left[\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t f(\mathbf{x}_t(\epsilon))\right|>\alpha/4\right]$$

Fix an $\mathcal{X}$-valued tree $\mathbf{x}$ of depth T. By assumption $\mathrm{fat}_{\alpha}(\mathcal{F})<\infty$ for any $\alpha>0$. Let V be a minimum ℓ_1-cover of $\mathcal{F}$ over $\mathbf{x}$ at scale $\alpha/8$. Corollary 6 ensures that

$$|V|=\mathcal{N}_1(\alpha/8,\mathcal{F},\mathbf{x})\le\left(\frac{16eT}{\alpha}\right)^{\mathrm{fat}_{\frac{\alpha}{8}}}$$

and for any $f\in\mathcal{F}$ and $\epsilon\in\{\pm1\}^T$, there exists $\mathbf{v}[f,\epsilon]\in V$ such that

$$\frac{1}{T}\sum_{t=1}^{T}|f(\mathbf{x}_t(\epsilon))-\mathbf{v}[f,\epsilon]_t(\epsilon)|\le\alpha/8$$

on the given path ϵ. Hence

$$\begin{aligned}
&\mathbb{P}_{\epsilon}\left[\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t f(\mathbf{x}_t(\epsilon))\right|>\alpha/4\right]\\
&\quad=\mathbb{P}_{\epsilon}\left[\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t\left(f(\mathbf{x}_t(\epsilon))-\mathbf{v}[f,\epsilon]_t(\epsilon)+\mathbf{v}[f,\epsilon]_t(\epsilon)\right)\right|>\alpha/4\right]\\
&\quad\le\mathbb{P}_{\epsilon}\left[\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t\left(f(\mathbf{x}_t(\epsilon))-\mathbf{v}[f,\epsilon]_t(\epsilon)\right)\right|+\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t\mathbf{v}[f,\epsilon]_t(\epsilon)\right|>\alpha/4\right]\\
&\quad\le\mathbb{P}_{\epsilon}\left[\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t\mathbf{v}[f,\epsilon]_t(\epsilon)\right|>\alpha/8\right]
\end{aligned}$$

For fixed $\epsilon=(\epsilon_1,\ldots,\epsilon_T)$,

$$\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t\mathbf{v}[f,\epsilon]_t(\epsilon)\right|>\alpha/8\qquad\Longrightarrow\qquad\frac{1}{T}\max_{\mathbf{v}\in V}\left|\sum_{t=1}^{T}\epsilon_t\mathbf{v}_t(\epsilon)\right|>\alpha/8$$

and, therefore, for any $\mathbf{x}$,

$$\begin{aligned}
\mathbb{P}_{\epsilon}\left[\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\epsilon_t f(\mathbf{x}_t(\epsilon))\right|>\alpha/4\right]\quad&\le\quad\mathbb{P}_{\epsilon}\left[\frac{1}{T}\max_{\mathbf{v}\in V}\left|\sum_{t=1}^{T}\epsilon_t\mathbf{v}_t(\epsilon)\right|>\alpha/8\right]\\
\le\sum_{\mathbf{v}\in V}\mathbb{P}_{\epsilon}\left[\frac{1}{T}\left|\sum_{t=1}^{T}\epsilon_t\mathbf{v}_t(\epsilon)\right|>\alpha/8\right]\quad&\le\quad 2|V|e^{-T\alpha^2/128}\quad\le\quad 2\left(\frac{16eT}{\alpha}\right)^{\mathrm{fat}_{\alpha/8}}e^{-T\alpha^2/128}
\end{aligned}$$

Hence we conclude that for any $\mathbf{D}$

$$\mathbb{P}_{\mathbf{D}}\left[\frac{1}{T}\sup_{f\in\mathcal{F}}\left|\sum_{t=1}^{T}\left(f(x_t)-\mathbb{E}_{t-1}[f(x_t)]\right)\right|>\alpha\right]\le 8\left(\frac{16eT}{\alpha}\right)^{\mathrm{fat}_{\alpha/8}}e^{-T\alpha^2/128}$$

Now applying Borel-Cantelli lemma proves the required result as

$$\sum_{T=1}^{\infty}8\left(\frac{16eT}{\alpha}\right)^{\mathrm{fat}_{\alpha/8}}e^{-T\alpha^2/128}<\infty\ .$$

□

Proof of Proposition 11. For the upper bound, we start by using Theorem 2 to bound the value of the game by Sequential Rademacher complexity,

$$\mathcal{V}_T^{\mathrm{S}}(\mathcal{F}) \leq 2\mathfrak{R}(\mathcal{F}_{\mathrm{S}}) \ .$$

Using the Lipschitz composition lemma (Lemma 18) with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\phi(t,(x,y)) = |t-y|$, we have $\mathfrak{R}(\mathcal{F}_{\mathrm{S}}) \leq \mathfrak{R}(\mathcal{F})$. This is because $|t-y|$ is 1-Lipschitz in t for any y. Hence, $\mathcal{V}_T^{\mathrm{S}}(\mathcal{F}) \leq 2\mathfrak{R}(\mathcal{F})$. We combine Theorem 9 and Corollary 6 to obtain the upper bound.

For the lower bound, we use a construction similar to [8]. We construct a particular distribution which induces a lower bound on regert for any algorithm. For any $\alpha \geq 0$ by definition of fat-shattering dimension, there exists a tree $\mathbf{x}$ of depth $d = \mathrm{fat}_\alpha(\mathcal{F})$ that can be α-shattered by $\mathcal{F}$. For simplicity, we assume $T = kd$ where k is some non-negative integer, and the case $T \leq d$ is discussed at the end of the proof. Now, define the jth block of time $T_j = \{(j-1)k+1, \ldots, jk\}$.

Now the strategy of Nature (Adversary) is to first pick $\tilde{\epsilon} \in \{\pm 1\}^T$ independently and uniformly at random. Further let $\epsilon \in \{\pm 1\}^d$ be defined as $\epsilon_j = \mathrm{sign}\left(\sum_{t \in T_j} \tilde{\epsilon}_t\right)$ for $1 \leq j \leq d$, the block-wise modal sign of $\tilde{\epsilon}$. Now note that by definition of α-shattering, there exists a witness tree $\mathbf{s}$ such that for any $\epsilon \in \{\pm 1\}^d$ there exists $f_\epsilon \in \mathcal{F}$ with $\epsilon_j(f_\epsilon(\mathbf{x}_j(\epsilon)) - \mathbf{s}_j(\epsilon)) \geq \alpha/2$ for all $1 \leq j \leq d$. Now let the random sequence $(x_1, y_1), \ldots, (x_T, y_T)$ be defined by $x_t = \mathbf{x}_j(\epsilon)$ for all $t \in T_j$ and $j \in \{1, \ldots, d\}$ and $y_t = \tilde{\epsilon}_t$. In the remainder of the proof we show that any algorithm suffers large expected regret.

Now consider any player strategy (possibly randomized) making prediction $\hat{y}_t \in [-1,1]$ at round t. Note that if we consider block j, $y_t = \tilde{\epsilon}_t$ is ± 1 uniformly at random. This means that irrespective of what $\hat{y}_t$ the player plays, the expectation over $\tilde{\epsilon}_t$ of the loss the player suffers at round t is

$$\mathbb{E}_{\tilde{\epsilon}_t} |\hat{y}_t - y_t| = 1$$

Hence on block j, the expected loss accumulated by any player is k and so for any player strategy (possibly randomized),

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] = \sum_{j=1}^{d} k = dk = T \tag{11}$$

On the other hand since $x_t = \mathbf{x}_j(\epsilon)$, we know that there always exists a function for any $\epsilon \in \{\pm 1\}^d$, say f_ϵ such that $\epsilon_j(f_\epsilon(\mathbf{x}_j(\epsilon)) - \mathbf{s}_j(\epsilon)) \geq \alpha/2$. Hence

$$\begin{aligned}
\mathbb{E}\left[\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} |f(x_t) - y_t|\right] &\leq \sum_{j=1}^{d} \mathbb{E}\left[\sum_{t \in T_j} |f_\epsilon(x_t) - y_t|\right] \\
&= \sum_{j=1}^{d} \mathbb{E}\left[\sum_{t \in T_j} |f_\epsilon(\mathbf{x}_j(\epsilon)) - y_t|\right] \\
&\leq \sum_{j=1}^{d} \mathbb{E}\left[\max_{c_j \in [\mathbf{s}_j(\epsilon) + \epsilon_j \frac{\alpha}{2}, \epsilon_j]} \sum_{t \in T_j} |c_j - y_t|\right]
\end{aligned}$$

where the last step is because for all of block j, $f_\epsilon(\mathbf{x}_j(\epsilon))$ does not depend on t and lies in the interval[1] $[\mathbf{s}_j(\epsilon) + \epsilon_j \frac{\alpha}{2}, \epsilon_j]$ (i.e. the majority side) and so by replacing it by the maximal c_j in the same interval for that block we only make the quantity bigger. Now for a block j, define the number of labels that match the sign of ϵ_j (the majority) as $M_j = \sum_{t \in T_j} \mathbf{1}\{y_t = \epsilon_j\}$. Since $y_t = \tilde{\epsilon}_t \in \{\pm 1\}$, observe that the function

[1] We use the convention that $[a,b]$ stands for $[b,a]$ whenever $a > b$.

$g(c_j) = \sum_{t\in T_j} |c_j - y_t|$ is linear on the interval $[-1, 1]$ with its minimum at the majority sign ϵ_j. Hence, the maximum over $[\mathbf{s}_j(\epsilon) + \epsilon_j \frac{\alpha}{2}, \epsilon_j]$ must occur at $c_j = \mathbf{s}_j(\epsilon) + \epsilon_j \frac{\alpha}{2}$. Substituting,

$$\begin{aligned}
\max_{c_j \in [\mathbf{s}_j(\epsilon)+\epsilon_j \frac{\alpha}{2}, \epsilon_j]} \sum_{t\in T_j} |c_j - y_t| &= M_j \left|\mathbf{s}_j(\epsilon) + \epsilon_j \frac{\alpha}{2} - \epsilon_j\right| + (k - M_j) \left|\mathbf{s}_j(\epsilon) + \epsilon_j \frac{\alpha}{2} + \epsilon_j\right| \\
&= M_j \left|\epsilon_j \mathbf{s}_j(\epsilon) + \frac{\alpha}{2} - 1\right| + (k - M_j) \left|\epsilon_j \mathbf{s}_j(\epsilon) + \frac{\alpha}{2} + 1\right| \\
&= M_j \left(1 - \epsilon_j \mathbf{s}_j(\epsilon) - \frac{\alpha}{2}\right) + (k - M_j) \left(1 + \epsilon_j \mathbf{s}_j(\epsilon) + \frac{\alpha}{2}\right) \\
&= k + (k - 2M_j) \left(\epsilon_j \mathbf{s}_j(\epsilon) + \frac{\alpha}{2}\right)
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\inf_{f\in\mathcal{F}} \sum_{t=1}^{T} |f(x_t) - y_t|\right] &\le dk + \sum_{j=1}^{d} \mathbb{E}\left[\epsilon_j \mathbf{s}_j(\epsilon)(k - 2M_j) + \frac{\alpha}{2}(k - 2M_j)\right] \\
&= dk + \sum_{j=1}^{d} \mathbb{E}\left[\epsilon_j \mathbf{s}_j(\epsilon)(k - 2M_j)\right] + \frac{\alpha}{2} \sum_{j=1}^{d} \mathbb{E}\left[k - 2M_j\right]
\end{aligned}$$

Further note that $k - 2M_j = -|\sum_{t\in T_j} \tilde{\epsilon}_t|$ and so $\epsilon_j(k - 2M_j) = -\sum_{t\in T_j} \tilde{\epsilon}_t$ and so the expectation

$$\mathbb{E}\left[\epsilon_j \mathbf{s}_j(\epsilon)(k - 2M_j)\right] = \mathbb{E}\left[\mathbb{E}_{\tilde{\epsilon}_{k(j-1)+1:jk}}\left[\epsilon_j \mathbf{s}_j(\epsilon)(k - 2M_j)\right]\right] = 0$$

because $\mathbf{s}_j(\epsilon)$ is independent of $\tilde{\epsilon}_t$ for $t \in T_j$. Hence we see that

$$\mathbb{E}\left[\inf_{f\in\mathcal{F}} \sum_{t=1}^{T} |f(x_t) - y_t|\right] \le dk + \frac{\alpha}{2} \sum_{j=1}^{d} \mathbb{E}\left[k - 2M_j\right] \tag{12}$$

Combining Equations (11) and (12) we can conclude that for any player strategy,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] - \mathbb{E}\left[\inf_{f\in\mathcal{F}} \sum_{t=1}^{T} |f(x_t) - y_t|\right] \quad &\ge \quad \frac{\alpha}{2} \sum_{j=1}^{d} \mathbb{E}\left[2M_j - k\right] \\
= \quad \frac{\alpha}{2} \mathbb{E}\left[\sum_{j=1}^{d} \left|\sum_{t\in T_j} \tilde{\epsilon}_t\right|\right] \quad = \quad \frac{\alpha}{2} \sum_{j=1}^{d} \mathbb{E}\left|\sum_{t\in T_j} \tilde{\epsilon}_t\right| \quad &\ge \quad \frac{\alpha d}{2}\sqrt{\frac{k}{2}} \quad = \quad \alpha\sqrt{\frac{Td}{8}} \quad = \quad \alpha\sqrt{\frac{T \,\mathrm{fat}_\alpha}{8}}
\end{aligned}$$

by Khinchine's inequality (e.g. [10, Lemma A.9]), yielding the theorem statement for $T \ge \mathrm{fat}_\alpha$. For the case of $T < \mathrm{fat}_\alpha$, the proof is the same with $k = 1$ and the depth of the shattered tree being T, yielding a lower bound of $\alpha T/\sqrt{8}$. This completes the proof. □

Proof of Theorem 12. The equivalence of *1* and *2* follows directly from Proposition 11. First, suppose that fat_α is infinite for some $\alpha > 0$. Then, the lower bound says that $\mathcal{V}_T^{\mathrm{S}}(\mathcal{F}) \ge \alpha T/2\sqrt{2}$ and hence $\limsup_{T\to\infty} \mathcal{V}_T^{\mathrm{S}}(\mathcal{F})/T \ge \alpha/2\sqrt{2}$. Thus, the class $\mathcal{F}$ is not online learnable in the supervised setting. Now, assume that fat_α is finite for all α. Fix an $\epsilon > 0$ and choose $\alpha = \epsilon/16$. Using the upper bound, we have

$$\begin{aligned}
\mathcal{V}_T^{\mathrm{S}}(\mathcal{F}) &\le 8T\alpha + 24\sqrt{T} \int_{\alpha}^{1} \sqrt{\mathrm{fat}_\beta \log\left(\frac{2eT}{\beta}\right)}\, d\beta \\
&\le 8T\alpha + 24\sqrt{T}(1 - \alpha)\sqrt{\mathrm{fat}_\alpha \log\left(\frac{2eT}{\alpha}\right)} \\
&\le \epsilon T/2 + \epsilon T/2
\end{aligned}$$

for T large enough. Thus, $\limsup_{T\to\infty} \mathcal{V}^{\mathrm{S}}_T(\mathcal{F})/T \le \epsilon$. Since $\epsilon > 0$ was arbitrary, this proves that $\mathcal{F}$ is online learnable in the supervised setting.

Let us now prove that if the problem is learnable, then all the complexities are close to each other. In the upper bounds below, the following calculation will be used several times. For any $b > 1$ and $\alpha \in (0,1)$

$$\int_\alpha^1 \frac{1}{\beta}\sqrt{\log(b/\beta)}d\beta = \int_b^{b/\alpha} \frac{1}{x}\sqrt{\log x}dx = \frac{2}{3}\log^{3/2}(x)\Big|_b^{b/\alpha} \le \frac{2}{3}\log^{3/2}(b/\alpha) \tag{13}$$

where we performed a change of variable with $x = b/\beta$.

First, if $\mathrm{fat}_\beta = \mathcal{O}(1/\beta^2)$, the problem is simple. Indeed, choosing $\alpha = 1/\sqrt{T}$, we appeal to Proposition 11 with the particular choice of α, followed by applying Eq. (13):

$$\begin{aligned}\mathfrak{D}(\mathcal{F}) &\le \inf_\alpha \left\{4T\alpha + 12\sqrt{T}\int_\alpha^1 \sqrt{\mathrm{fat}_\beta \log(2eT/\beta)}\ d\beta\right\} \\ &\le \sqrt{T}\cdot\mathcal{O}\left(1 + \int_{1/\sqrt{T}}^1 \frac{1}{\beta}\sqrt{\log(2eT/\beta)}\ d\beta\right) = \mathcal{O}\left(\sqrt{T}\log^{3/2}(T)\right),\end{aligned}$$

By choosing $\alpha = 1/2$ in the lower bound in Proposition 11, we conclude that for $T > \mathrm{fat}_{\frac{1}{2}}$,

$$\frac{1}{8\sqrt{2}}\sqrt{T\mathrm{fat}_{\frac{1}{2}}} \le \frac{1}{2}\mathcal{V}^{\mathrm{S}}_T(\mathcal{F}) \le \mathfrak{R}_T(\mathcal{F}) \le \mathfrak{D}_T(\mathcal{F}) \le c\sqrt{T}\log^{3/2}(T)$$

for some constant c, yielding the statement of the theorem for the simple case.

Now, for the case when fat_β grows faster than $1/\beta^2$ as β decays, let $\alpha = \hat\alpha$ be the solution of $T = \mathrm{fat}_\alpha$. The lower bound of Proposition 11 then becomes

$$c\hat\alpha\sqrt{T\mathrm{fat}_{\hat\alpha}} = c\hat\alpha T \le \mathcal{V}^{\mathrm{S}}_T(\mathcal{F}) \qquad \text{with } c = 1/(2\sqrt{2}) \tag{14}$$

Note that $\mathrm{fat}_\beta \le \mathrm{fat}_{\hat\alpha}$ for $\beta \ge \hat\alpha$, and the fact that fat_β grows at least as fast as $1/\beta^2$ implies that $\beta\sqrt{\mathrm{fat}_\beta} \le \hat\alpha\sqrt{\mathrm{fat}_{\hat\alpha}}$. From (14), $\sqrt{\mathrm{fat}_\beta} \le \frac{\mathcal{V}^{\mathrm{S}}_T(\mathcal{F})}{c\beta\sqrt{T}}$ for $\beta > \hat\alpha$ and $\hat\alpha \le \frac{\mathcal{V}^{\mathrm{S}}_T(\mathcal{F})}{cT}$. Using Eq. (13), for some absolute constant c',

$$\begin{aligned}\mathfrak{D}(\mathcal{F}) &\le 4T\hat\alpha + 12\sqrt{T}\int_{\hat\alpha}^1 \left(\frac{\mathcal{V}^{\mathrm{S}}_T(\mathcal{F})}{c\beta\sqrt{T}}\right)\sqrt{\log(2eT/\beta)}\ d\beta \\ &\le c'\left\{\mathcal{V}^{\mathrm{S}}_T(\mathcal{F}) + \mathcal{V}^{\mathrm{S}}_T(\mathcal{F})\int_{\hat\alpha}^1 \frac{1}{\beta}\sqrt{\log(2eT/\beta)}\ d\beta\right\} \\ &= \mathcal{V}^{\mathrm{S}}_T(\mathcal{F})\cdot\mathcal{O}\left(\log^{3/2}(T/\hat\alpha)\right)\end{aligned}$$

By our assumption, fat_α grows at least as fast as $1/\alpha^2$, and thus $T/\hat\alpha = \mathcal{O}(T^{3/2})$, allowing us to conclude that $\mathfrak{D}_T(\mathcal{F}) \le \mathcal{V}^{\mathrm{S}}_T(\mathcal{F})\cdot\mathcal{O}(\log^{3/2}(T))$. We conclude

$$\mathfrak{D}_T(\mathcal{F}) \le \mathcal{V}^{\mathrm{S}}_T(\mathcal{F})\cdot\mathcal{O}(\log^{3/2}(T)) \quad \text{and} \quad \mathcal{V}^{\mathrm{S}}_T(\mathcal{F}) \le 2\mathfrak{R}_T(\mathcal{F}) \le 2\mathfrak{D}_T(\mathcal{F}).$$

□

Proof of Lemma 14. First, we claim that for any $x \in \mathcal{X}$, $\mathrm{fat}_\alpha(V_t(r,x)) = \mathrm{fat}_\alpha(V_t)$ for at most two $r, r' \in B_\alpha$.[2] Further if there are two such $r, r' \in B_\alpha$ then r, r' are consecutive elements of B_α (i.e. $|r - r'| = \alpha$). Suppose, for the sake of contradiction, that $\mathrm{fat}_\alpha(V_t(r,x)) = \mathrm{fat}_\alpha(V_t(r',x)) = \mathrm{fat}_\alpha(V_t)$ for distinct $r, r' \in B_\alpha$

[2] The argument should be compared to the combinatorial argument in Theorem 5.

that are not consecutive (i.e. $|r - r'| \geq 2\alpha$). Then let $s = (r + r')/2$ and without loss of generality suppose $r > r'$. By definition for any $f \in V_t(r, x)$,

$$f(x) \geq r - \alpha/2 = (r' + r)/2 + (r - r')/2 - \alpha/2 \geq s + \alpha/2$$

Also for any $g \in V_t(r', x)$ we also have,

$$g(x) \leq r' + \alpha/2 = (r' + r)/2 + (r' - r)/2 + \alpha/2 \leq s - \alpha/2$$

Let $\mathbf{v}$ and $\mathbf{v}'$ be trees of depth $\text{fat}_\alpha(V_t)$ α-shattered by $V_t(r, x)$ and $V_t(r', x)$, respectively. To get a contradiction, form a new tree of depth $\text{fat}_\alpha(V_t) + 1$ by joining $\mathbf{v}$ and $\mathbf{v}'$ with the constant function $\mathbf{x}_1 = x$ as the root. It is straightforward that this tree is shattered by $V_t(r, x) \cup V_t(r', x)$, a contradiction.

Notice that the times $t \in [T]$ for which $|f_t(x_t) - y_t| > \alpha$ are exactly those times when we update current set V_{t+1}. We shall show that whenever an update is made, $\text{fat}_\alpha(V_{t+1}) < \text{fat}_\alpha(V_t)$ and hence claim that the total number of times $|f_t(x_t) - y_t| > \alpha$ is bounded by $\text{fat}_\alpha(\mathcal{F})$.

At any round we have three possibilities. First is when $\text{fat}_\alpha(V_t(r, x_t)) < \text{fat}_\alpha(V_t)$ for all $r \in B_\alpha$. In this case, clearly, an update results in $\text{fat}_\alpha(V_{t+1}) = \text{fat}_\alpha(V_t(\lfloor y_t \rfloor_\alpha, x_t)) < \text{fat}_\alpha(V_t)$.

The second case is when $\text{fat}_\alpha(V_t(r, x_t)) = \text{fat}_\alpha(V_t)$ for exactly one $r \in B_\alpha$. In this case the algorithm chooses $f_t(x_t) = r$. If the update is made, $|f_t(x_t) - y_t| > \alpha$ and thus $\lfloor y_t \rfloor_\alpha \neq f_t(x_t)$. We can conclude that

$$\text{fat}_\alpha(V_{t+1}) = \text{fat}_\alpha(V_t(\lfloor y_t \rfloor_\alpha, x_t)) < \text{fat}_\alpha(V_t(f_t(x_t), x_t)) = \text{fat}_\alpha(V_t)$$

The final case is when $\text{fat}_\alpha(V_t(r, x_t)) = \text{fat}_\alpha(V_t(r', x_t)) = \text{fat}_\alpha(V_t)$ and $|r - r'| = \alpha$. In this case, the algorithm chooses $f_t(x_t) = \frac{r+r'}{2}$. Whenever y_t falls in either of these two consecutive intervals given by r or r', we have $|f_t(x_t) - y_t| \leq \alpha$, and hence no update is made. Thus, if an update is made, $\lfloor y_t \rfloor_\alpha \neq r$ and $\lfloor y_t \rfloor_\alpha \neq r'$. However, for any element or B_α other than r, r', the fat shattering dimension is less than that of V_t. That is

$$\text{fat}_\alpha(V_{t+1}) = \text{fat}_\alpha(V_t(\lfloor y_t \rfloor_\alpha, x_t)) < \text{fat}_\alpha(V_t(r, x_t)) = \text{fat}_\alpha(V_t(r', x_t)) = \text{fat}_\alpha(V_t).$$

We conclude that whenever we update, $\text{fat}_\alpha(V_{t+1}) < \text{fat}_\alpha(V_t)$, and so we can conclude that algorithm's prediction is more than α away from y_t on at most $\text{fat}_\alpha(\mathcal{F})$ number of rounds. □

Proof of Corollary 17. For the choice of weights $p_i = \frac{6}{\pi^2} i^{-2}$ we see from Proposition 29 that for any i,

$$\mathbb{E}[\mathbf{R}_T] \leq \alpha_i T + \sqrt{T \text{fat}_{\alpha_i} \log\left(\frac{2T}{\alpha_i}\right)} + \sqrt{T}\left(3 + 2\log(i)\right)$$

Now for any $\alpha > 0$ let i_α be such that $\alpha \leq 2^{-i_\alpha}$ and for any $i < i_\alpha$, $\alpha > 2^{-i_\alpha}$. Using the above bound on expected regret we have that

$$\mathbb{E}[\mathbf{R}_T] \leq \alpha_{i_\alpha} T + \sqrt{T \text{fat}_{\alpha_{i_\alpha}} \log\left(\frac{2T}{\alpha_{i_\alpha}}\right)} + \sqrt{T}\left(3 + 2\log(i_\alpha)\right)$$

However for our choice of i_α we see that $i_\alpha \leq \log(1/\alpha)$ and further $\alpha_{i_\alpha} \leq \alpha$. Hence we conclude that

$$\mathbb{E}[\mathbf{R}_T] \leq \alpha T + \sqrt{T \text{fat}_\alpha \log\left(\frac{2T}{\alpha}\right)} + \sqrt{T}\left(3 + 2\log\log\left(\frac{1}{\alpha}\right)\right)$$

Since choice of α was arbitrary we take infimum and get the result. □

Proof of Lemma 18. Without loss of generality assume $L = 1$. The general case follow from this by simply scaling ϕ appropriately. We will also use the shorthand $\phi(f, z)$ to denote $\phi(f(z), z)$. We have

$$\mathfrak{R}(\phi(\mathcal{F})) = \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t \phi(f, \mathbf{x}_t(\epsilon)) \right] .$$

The proof proceeds by sequentially using the Lipschitz property of $\phi(f(\mathbf{x}_t(\epsilon)))$ for increasing t. Towards this end, define

$$R_t = \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{s=1}^{t} \epsilon_s f(\mathbf{x}_s(\epsilon)) + \sum_{s=t+1}^{T} \epsilon_s \phi(f, \mathbf{x}_s(\epsilon)) \right] .$$

Note that $R_0 = \mathfrak{R}(\phi(\mathcal{F}))$ and $R_T = \mathfrak{R}(\mathcal{F})$. We need to show $R_0 \le R_T$ and we will show this by proving $R_{t-1} \le R_t$ for all $t \in [T]$. So, let us fix $t \in [T]$ and start with R_{t-1}:

$$R_{t-1} = \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s f(\mathbf{x}_s(\epsilon)) + \sum_{s=t}^{T} \epsilon_s \phi(f, \mathbf{x}_s(\epsilon)) \right] .$$

We can write the above supremum as,

$$\begin{aligned} R_{t-1} &= \sup_{x_1 \in \mathcal{Z}} \mathbb{E}_{\epsilon_1} \dots \sup_{x_t \in \mathcal{Z}} \mathbb{E}_{\epsilon_t} \sup_{\tilde{\mathbf{x}}} \mathbb{E}_{\epsilon_{t+1:T}} \left[\sup_{f \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s f(x_s) + \epsilon_t \phi(f, x_t) + \sum_{s=t+1}^{T} \epsilon_s \phi(f, \tilde{\mathbf{x}}_{s-t}(\epsilon_{t+1:T})) \right] \\ &= \sup_{x_1 \in \mathcal{Z}} \mathbb{E}_{\epsilon_1} \dots \sup_{x_t \in \mathcal{Z}} S^\phi(x_{1:t}, \epsilon_{1:t-1}) , \end{aligned}$$

where we have simply defined

$$S^\phi(x_{1:t}, \epsilon_{1:t-1}) = \mathbb{E}_{\epsilon_t} \sup_{\tilde{\mathbf{x}}} \mathbb{E}_{\epsilon_{t+1:T}} \left[\sup_{f \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s f(x_s) + \epsilon_t \phi(f, x_t) + \sum_{s=t+1}^{T} \epsilon_s \phi(f, \tilde{\mathbf{x}}_{s-t}(\epsilon_{t+1:T})) \right] .$$

Here, $\tilde{\mathbf{x}}$ ranges over all $\mathcal{Z}$-valued trees of depth $T - t$.

Similarly, R_t can be written as,

$$\begin{aligned} R_t &= \sup_{x_1 \in \mathcal{Z}} \mathbb{E}_{\epsilon_1} \dots \sup_{x_t \in \mathcal{Z}} \mathbb{E}_{\epsilon_t} \sup_{\tilde{\mathbf{x}}} \mathbb{E}_{\epsilon_{t+1:T}} \left[\sup_{f \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s f(x_s) + \epsilon_t f(x_t) + \sum_{s=t+1}^{T} \epsilon_s \phi(f, \tilde{\mathbf{x}}_{s-t}(\epsilon_{t+1:T})) \right] \\ &= \sup_{x_1 \in \mathcal{Z}} \mathbb{E}_{\epsilon_1} \dots \sup_{x_t \in \mathcal{Z}} S(x_{1:t}, \epsilon_{1:t-1}) , \end{aligned}$$

where we have defined

$$S(x_{1:t}, \epsilon_{1:t-1}) = \mathbb{E}_{\epsilon_t} \sup_{\tilde{\mathbf{x}}} \mathbb{E}_{\epsilon_{t+1:T}} \left[\sup_{f \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s f(x_s) + \epsilon_t f(x_t) + \sum_{s=t+1}^{T} \epsilon_s \phi(f, \tilde{\mathbf{x}}_{s-t}(\epsilon_{t+1:T})) \right] .$$

Thus, to prove $R_{t-1} \le R_t$ it suffices to prove $S^\phi(x_{1:t}, \epsilon_{1:t-1}) \le S(x_{1:t}, \epsilon_{1:t-1})$ for all $x_{1:t} \in \mathcal{Z}^t$ and $\epsilon_{1:t-1} \in \{\pm 1\}^{t-1}$. Fix $x_{1:t}, \epsilon_{1:t-1}$. By explicitly taking expectation w.r.t. ϵ_t in the definition of S^ϕ, we have

$$\begin{aligned} 2S^\phi &= \sup_{\tilde{\mathbf{x}}} \mathbb{E}_{\epsilon_{t+1:T}} \left[\sup_{f \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s f(x_s) + \phi(f, x_t) + \sum_{s=t+1}^{T} \epsilon_s \phi(f, \tilde{\mathbf{x}}_{s-t}(\epsilon_{t+1:T})) \right] \\ &\quad + \sup_{\tilde{\mathbf{x}}'} \mathbb{E}_{\epsilon_{t+1:T}} \left[\sup_{g \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s g(x_s) - \phi(g, x_t) + \sum_{s=t+1}^{T} \epsilon_s \phi(g, \tilde{\mathbf{x}}'_{s-t}(\epsilon_{t+1:T})) \right] \\ &= \sup_{\tilde{\mathbf{x}}, \tilde{\mathbf{x}}'} \mathbb{E}_{\epsilon_{t+1:T}} \left[\sup_{f, g \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s (f(x_s) + g(x_s)) + \phi(f, x_t) - \phi(g, x_t) + \sum_{s=t+1}^{T} \epsilon_s (\phi(f, \tilde{\mathbf{x}}_{s-t}(\epsilon_{t+1:T})) + \phi(g, \tilde{\mathbf{x}}'_{s-t}(\epsilon_{t+1:T})) \right] . \end{aligned}$$

Now $\phi(f, x_t) - \phi(g, x_t) = \phi(f(x_t), x_t) - \phi(g(x_t), x_t)$ is upper bounded by $|f(x_t) - g(x_t)|$ because $\phi(\cdot, z)$ is 1-Lipschitz for any z. Hence,

$$2S^{\phi} \leq \sup_{\tilde{\mathbf{x}}, \tilde{\mathbf{x}}'} \mathbb{E}_{\epsilon_{t+1:T}} \left[\sup_{f,g \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s (f(x_s) + g(x_s)) + |f(x_t) - g(x_t)| + \sum_{s=t+1}^{T} \epsilon_s(\phi(f, \tilde{\mathbf{x}}_{s-t}(\epsilon_{t+1:T})) + \phi(g, \tilde{\mathbf{x}}'_{s-t}(\epsilon_{t+1:T})) \right] .$$

Since, for any $\epsilon_{t+1:T}$, the first and last sum above are unchanged if we simultaneously exchange f with g and $\tilde{\mathbf{x}}$ with $\tilde{\mathbf{x}}'$, the above supremum is actually equal to one where the absolute value in the middle term is absent. That is,

$$\begin{aligned}
2\, S^{\phi} &\leq \sup_{\tilde{\mathbf{x}}, \tilde{\mathbf{x}}'} \mathbb{E}_{\epsilon_{t+1:T}} \left[\sup_{f,g \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s (f(x_s) + g(x_s)) + f(x_t) - g(x_t) + \sum_{s=t+1}^{T} \epsilon_s(\phi(f, \tilde{\mathbf{x}}_{s-t}(\epsilon_{t+1:T})) + \phi(g, \tilde{\mathbf{x}}'_{s-t}(\epsilon_{t+1:T})) \right] \\
&= \sup_{\tilde{\mathbf{x}}} \mathbb{E}_{\epsilon_{t+1:T}} \left[\sup_{f \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s f(x_s) + f(x_t) + \sum_{s=t+1}^{T} \epsilon_s \phi(f, \tilde{\mathbf{x}}_{s-t}(\epsilon_{t+1:T})) \right] \\
&\quad + \sup_{\tilde{\mathbf{x}}'} \mathbb{E}_{\epsilon_{t+1:T}} \left[\sup_{g \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s g(x_s) - g(x_t) + \sum_{s=t+1}^{T} \epsilon_s \phi(g, \tilde{\mathbf{x}}'_{s-t}(\epsilon_{t+1:T})) \right] \\
&= 2\, \mathbb{E}_{\epsilon_t} \sup_{\tilde{\mathbf{x}}} \mathbb{E}_{\epsilon_{t+1:T}} \left[\sup_{f \in \mathcal{F}} \sum_{s=1}^{t-1} \epsilon_s f(x_s) + \epsilon_t f(x_t) + \sum_{s=t+1}^{T} \epsilon_s \phi(f, \tilde{\mathbf{x}}_{s-t}(\epsilon_{t+1:T})) \right] \\
&= 2\, S(x_{1:t}, \epsilon_{1:t-1}) \ .
\end{aligned}$$

□

Proof of Lemma 19. Without loss of generality assume that the Lipschitz constant $L = 1$ because the general case follows by scaling ϕ. Now note that by Theorem 9 we have that

$$\mathfrak{R}(\phi \circ \mathcal{F}) \leq \inf_{\alpha} \left\{ 4T\alpha + 12 \int_{\alpha}^{1} \sqrt{T \ \log \ \mathcal{N}_2(\delta, \phi \circ \mathcal{F}, T)}\, d\delta \right\} \tag{15}$$

Now we claim that we can bound

$$\log \ \mathcal{N}_2(\delta, \phi \circ \mathcal{F}, T) \leq \sum_{j=1}^{k} \log \ \mathcal{N}_\infty(\delta, \mathcal{F}_j, T)$$

To see this we first start by noting that

$$\begin{aligned}
\sqrt{\frac{1}{T} \sum_{t=1}^{T} \left(\phi(f(\mathbf{x}_t(\epsilon))) - \phi(\mathbf{v}_t(\epsilon))\right)^2} \leq \sqrt{\frac{1}{T} \sum_{t=1}^{T} \max_j \left(f_j(\mathbf{x}_t(\epsilon))) - \mathbf{v}_t^j(\epsilon) \right)^2} \quad &\leq \sqrt{\max_{t \in [T]} \max_j \left(f_j(\mathbf{x}_t(\epsilon))) - \mathbf{v}_t^j(\epsilon) \right)^2} \\
&\leq \max_{j \in [k], t \in [T]} |f_j(\mathbf{x}_t(\epsilon))) - \mathbf{v}_t^j(\epsilon)|
\end{aligned}$$

This means that if we have $V_1, \ldots, V_k$ that are minimal L_∞ covers for $\mathcal{F}_1, \ldots, \mathcal{F}_k$ at level δ then if we construct a cover $V = V_1 \times \ldots \times V_k$ for $\mathcal{F}$ then for any $f = (f_1, \ldots, f_k) \in \mathcal{F}$ and any $\epsilon \in \{\pm 1\}^T$, there exists $\mathbf{v} = (\mathbf{v}^1, \ldots, \mathbf{v}^k) \in V$ such that

$$\sqrt{\frac{1}{T} \sum_{t=1}^{T} \left(\phi(f(\mathbf{x}_t(\epsilon))) - \phi(\mathbf{v}_t(\epsilon))\right)^2} \leq \max_{j \in [k]} \max_{t \in [T]} |f_j(\mathbf{x}_t(\epsilon))) - \mathbf{v}_t^j(\epsilon)| \leq \delta$$

Hence we see that V is an ℓ_∞ cover at scale δ for $\phi \circ \mathcal{F}$. Hence

$$\log \mathcal{N}_2(\delta, \phi \circ \mathcal{F}, T) \leq \log \mathcal{N}_\infty(\delta, \phi \circ \mathcal{F}, T) \leq \log(|V|) = \sum_{j=1}^{k} \log(|V_j|) = \sum_{j=1}^{k} \log \mathcal{N}_\infty(\delta, \mathcal{F}_j, T)$$

as claimed. Now using this in Equation 15 we have that

$$\begin{aligned}\mathfrak{R}(\phi \circ \mathcal{F}) &\leq \inf_\alpha \left\{ 4T\alpha + 12 \int_\alpha^1 \sqrt{T \sum_{j=1}^{k} \log \mathcal{N}_\infty(\delta, \mathcal{F}_j, T)\ d\delta} \right\} \\ &\leq \inf_\alpha \left\{ 4T\alpha + 12 \sum_{j=1}^{k} \int_\alpha^1 \sqrt{T \log \mathcal{N}_\infty(\delta, \mathcal{F}_j, T)}\ d\delta \right\} \\ &\leq \sum_{j=1}^{k} \inf_\alpha \left\{ 4T\alpha + 12 \int_\alpha^1 \sqrt{T \log \mathcal{N}_\infty(\delta, \mathcal{F}_j, T)}\ d\delta \right\}\end{aligned}$$

Now applying Theorem 12 we conclude as required that

$$\mathfrak{R}(\phi \circ \mathcal{F}) \leq \mathcal{O}\left(\log^{3/2}(T)\right) \sum_{j=1}^{k} \mathfrak{R}(\mathcal{F}_j)$$

□

Proof of Corollary 20. We first extend the binary function b to a function $\bar{b}$ to any $x \in \mathbb{R}^k$ as follows :

$$\bar{b}(x) = \begin{cases} (1 - \|x - a\|_\infty) b(a) & \text{if } \|x - a\|_\infty < 1 \text{ for some } a \in \{\pm 1\}^k \\ 0 & \text{otherwise} \end{cases}$$

First note that $\bar{b}$ is well-defined since all points in the k-cube are separated by L_∞ distance 2. Further note that $\bar{b}$ is 1-Lipschitz w.r.t. the L_∞ norm and so applying Lemma 19 we conclude the statement of the corollary. □

Proof of Proposition 21. The most difficult of these is Part 4, which follows immediately by Lemma 18 by taking $\phi(\cdot, z)$ there to be simply $\phi(\cdot)$. The other items follow similarly to Theorem 15 in [26] and we provide the proofs for completeness. Note that, unlike Rademacher complexity defined in [26], Sequential Rademacher complexity does not have the absolute value around the sum.

Part 1 is immediate because for any fixed tree $\mathbf{x}$ and fixed realization of $\{\epsilon_i\}$,

$$\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon)) \leq \sup_{f \in \mathcal{G}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon)) ,$$

Now taking expectation over ϵ and supremum over $\mathbf{x}$ completes the argument.

To show Part 2, first observe that, according to Part 1,

$$\mathfrak{R}(\mathcal{F}) \leq \mathfrak{R}(\text{conv}(\mathcal{F})) .$$

Now, any $h \in \text{conv}(\mathcal{F})$ can be written as $h = \sum_{j=1}^m \alpha_j f_j$ with $\sum_{j=1}^m \alpha_j = 1$, $\alpha_j \geq 0$. Then, for fixed tree $\mathbf{x}$ and sequence ϵ,

$$\sum_{t=1}^{T} \epsilon_t h(\mathbf{x}_t(\epsilon) = \sum_{j=1}^{m} \alpha_j \sum_{t=1}^{T} \epsilon_t f_j(\mathbf{x}_t(\epsilon) \leq \sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon))$$

and thus

$$\sup_{h\in\mathrm{conv}(\mathcal{F})}\sum_{t=1}^{T}\epsilon_t h(\mathbf{x}_t(\epsilon) \leq \sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\epsilon_t f(\mathbf{x}_t(\epsilon)\ .$$

Taking expectation over ϵ and supremum over $\mathbf{x}$ completes the proof.

To prove Part 3, first observe that the statement is easily seem to hold for $c \geq 0$. That is, $\mathfrak{R}(c\mathcal{F}) = c\mathfrak{R}(\mathcal{F})$ follows directly from the definition. Hence, it remains to convince ourselves of the statement for $c = -1$. That is, $\mathfrak{R}(-\mathcal{F}) = \mathfrak{R}(\mathcal{F})$. To prove this, consider a tree $\mathbf{x}^R$ that is a reflection of $\mathbf{x}$. That is, $\mathbf{x}_t^R(\epsilon) = \mathbf{x}_t(-\epsilon)$ for all $t \in [T]$. It is then enough to observe that

$$\begin{aligned}\mathbb{E}_\epsilon\left[\sup_{f\in-\mathcal{F}}\sum_{t=1}^{T}\epsilon_t f(\mathbf{x}_t(\epsilon))\right] &= \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}-\epsilon_t f(\mathbf{x}_t(\epsilon))\right]\\ = \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\epsilon_t f(\mathbf{x}_t(-\epsilon))\right] &= \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\epsilon_t f(\mathbf{x}_t^R(\epsilon))\right]\end{aligned}$$

where we used the fact that ϵ and $-\epsilon$ have the same distribution. As $\mathbf{x}$ varies over all trees, $\mathbf{x}^R$ also varies over all trees. Hence taking the supremum over $\mathbf{x}$ above finishes the argument.

Finally, for Part 5,

$$\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{T}\epsilon_t\left(f+h\right)\left(\mathbf{x}_t(\epsilon)\right)\right\} = \left\{\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\epsilon_t f(\mathbf{x}_t(\epsilon))\right\} + \left\{\sum_{t=1}^{T}\epsilon_t h(\mathbf{x}_t(\epsilon))\right\}$$

Note that, since $h(\mathbf{x}_t(\epsilon))$ only depends on $\epsilon_{1:t-1}$, we have

$$\mathbb{E}_\epsilon\left[\epsilon_t h(\mathbf{x}_t(\epsilon))\right] = \mathbb{E}_{\epsilon_{1:t-1}}\left[\mathbb{E}\left[\epsilon_t|\epsilon_{1:t-1}\right]h(\mathbf{x}_t(\epsilon)\right] = 0\ .$$

Thus,

$$\mathfrak{R}(\mathcal{F}+h) = \mathfrak{R}(\mathcal{F})\ .$$

□

Proof of Proposition 22. We use linearity of the functions in $\mathcal{F}_\mathcal{W}$ to write

$$\begin{aligned}\mathfrak{R}_T(\mathcal{F}_\mathcal{W}) &= \sup_{\mathbf{x}}\mathbb{E}_\epsilon\left[\sup_{w\in\mathcal{W}}\sum_{t=1}^{T}\epsilon_t\left\langle w, \mathbf{x}_t(\epsilon)\right\rangle\right]\\ &= \sup_{\mathbf{x}}\mathbb{E}_\epsilon\left[\sup_{w\in\mathcal{W}}\left\langle w, \sum_{t=1}^{T}\epsilon_t\mathbf{x}_t(\epsilon)\right\rangle\right]\end{aligned}$$

Let $\Psi^\star$ be the Fenchel conjugate of Ψ. By Fenchel-Young inequality, for any $\lambda > 0$,

$$\left\langle w, \sum_{t=1}^{T}\epsilon_t\mathbf{x}_t(\epsilon)\right\rangle \leq \frac{\Psi(w)}{\lambda} + \frac{\Psi^\star\left(\sum_{t=1}^{T}\lambda\,\epsilon_t\,\mathbf{x}_t(\epsilon)\right)}{\lambda}\ .$$

Taking supremum over $w \in \mathcal{W}$, we get,

$$\sup_{w\in\mathcal{W}}\left\langle w, \sum_{t=1}^{T}\epsilon_t\mathbf{x}_t(\epsilon)\right\rangle \leq \frac{\Psi_{\max}}{\lambda} + \frac{\Psi^\star\left(\sum_{t=1}^{T}\lambda\,\epsilon_t\,\mathbf{x}_t(\epsilon)\right)}{\lambda}\ .$$

Now, taking expectation w.r.t. ϵ, we get

$$\mathbb{E}_\epsilon \left[\sup_{w \in \mathcal{W}} \left\langle w, \sum_{t=1}^T \epsilon_t \mathbf{x}_t(\epsilon) \right\rangle \right] \le \frac{\Psi_{\max}}{\lambda} + \frac{\mathbb{E}_\epsilon \left[\Psi^\star \left(\sum_{t=1}^T \lambda \, \epsilon_t \, \mathbf{x}_t(\epsilon) \right) \right]}{\lambda} .$$

Now we use Lemma 23 below with $Z_t = \lambda \, \epsilon_t \, \mathbf{x}_t(\epsilon)$. Note that $\mathbb{E}\left[Z_t \mid \epsilon_{1:t-1} \right] = 0$ and $\|Z_t\|_\star \le \lambda \|\mathcal{X}\|_\star$ almost surely. Thus, we get,

$$\mathbb{E}_\epsilon \left[\sup_{w \in \mathcal{W}} \left\langle w, \sum_{t=1}^T \epsilon_t \mathbf{x}_t(\epsilon) \right\rangle \right] \le \frac{\Psi_{\max}}{\lambda} + \frac{\lambda^2 \, \|\mathcal{X}\|_\star^2 \, T}{2 \, \sigma \, \lambda} .$$

Simplifying and optimizing over $\lambda > 0$, gives

$$\mathfrak{R}_T(\mathcal{F}_{\mathcal{W}}) \le \|\mathcal{X}\|_\star \sqrt{\frac{2 \, \Psi_{\max} \, T}{\sigma}} .$$

□

Proof of Lemma 24. Consider the game $(\mathcal{F}, \mathcal{X}_{\text{cvx}})$ and fix a randomized strategy π of the player. Then, the expected regret of Π against any adversary playing $g_1, \ldots, g_T$ can be bounded as

$$\begin{aligned}
\mathbf{R}(\pi, g_{1:T}) &= \sum_{t=1}^T \mathbb{E}_{u_t \sim \pi_t(g_{1:t-1})} \left[g_t(u_t) \right] - \inf_{u \in \mathcal{F}} \sum_{t=1}^T g_t(u) \\
&\le \sum_{t=1}^T g_t \left(\mathbb{E}_{u_t \sim \pi_t(g_{1:t-1})} \left[u_t \right] \right) - \inf_{u \in \mathcal{F}} \sum_{t=1}^T g_t(u) \\
&= \mathbf{R}(\pi', g_{1:T}) .
\end{aligned}$$

Here we used Jensen's inequality in the second line and π' is simply the *deterministic* strategy obtained from π that, on round t, plays

$$\mathbb{E}_{u_t \sim \pi_t(g_{1:t-1})} \left[u_t \right] .$$

This means that $\mathcal{V}_T(\mathcal{F}, \mathcal{X}_{\text{cvx}}) = \mathcal{V}_T^{\text{det}}(\mathcal{F}, \mathcal{X}_{\text{cvx}})$ where $\mathcal{V}_T^{\text{det}}$ is defined as the minimax regret obtainable only using deterministic player strategies. Now, we appeal to Theorem 14 in [2] that says $\mathcal{V}_T^{\text{det}}(\mathcal{F}, \mathcal{X}_{\text{cvx}}) = \mathcal{V}_T^{\text{det}}(\mathcal{F}, \mathcal{X}_{\text{lin}})$. Since $\mathcal{X}_{\text{lin}}$ also consists of convex (in fact, linear) functions, the above argument again gives $\mathcal{V}_T^{\text{det}}(\mathcal{F}, \mathcal{X}_{\text{lin}}) = \mathcal{V}_T(\mathcal{F}, \mathcal{X}_{\text{lin}})$. This finishes the proof of the lemma. □

Proof of Proposition 25. Fix a $\gamma > 0$ and use loss

$$\ell(\hat{y}, y) = \begin{cases} 1 & \hat{y}y \le 0 \\ 1 - \hat{y}y/\gamma & 0 < \hat{y}y < \gamma \\ 0 & \hat{y}y \ge \gamma \end{cases}$$

First note that since the loss is $1/\gamma$-Lipschitz, we can use Theorem 2 and the Rademacher contraction Lemma 18 to show that for each $\gamma > 0$ there exists a randomized strategy π^γ such that

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{f_t \sim \pi_t^\gamma(z_{1:t-1})} \left[\ell(f_t(x_t), y_t) \right] \right] \le \inf_{f \in \mathcal{F}} \sum_{t=1}^T \ell(f(x_t), y_t) + \frac{2}{\gamma} \mathfrak{R}_T(\mathcal{F})$$

Now note that the loss is lower bounded by the Zero-one loss $\mathbf{1}\left\{ \hat{y}y < 0 \right\}$ and is upper bounded by the margin Zero-one loss $\mathbf{1}\left\{ \hat{y}y < \gamma \right\}$. Hence we see that for this strategy,

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{f_t \sim \pi_t^\gamma(z_{1:t-1})} \left[\mathbf{1}\left\{ f_t(x_t) y_t < 0 \right\} \right] \right] \le \inf_{f \in \mathcal{F}} \sum_{t=1}^T \mathbf{1}\left\{ f(x_t) y_t < \gamma \right\} + \frac{2}{\gamma} \mathfrak{R}_T(\mathcal{F}) \tag{16}$$

Hence for each fixed γ for randomized strategy given by π^{γ} we have the above bound. Now we discretize over γ's as $\gamma_i = 1/2^i$ and using the output of the randomized strategies $\pi^{\gamma_1}, \pi^{\gamma_2}, \ldots$ that attain the regret bounds given in (16) as experts and running experts algorithm given in Algorithm 3 with initial weight for expert i as $p_i = \frac{6}{\pi^2 i^2}$ then using Proposition 29 we get that for this randomized strategy π, such that for any i

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{f_t \sim \pi_t(z_{1:t-1})}\left[\mathbf{1}\left\{f_t(x_t) y_t < 0\right\}\right]\right] \le \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathbf{1}\left\{f(x_t) y_t < \gamma_i\right\} + \frac{2}{\gamma_i}\mathfrak{R}_T(\mathcal{F}) + \sqrt{T}\left(1 + 2\log\left(\frac{i\pi}{\sqrt{6}}\right)\right)$$

Now for any $\gamma > 0$ let i_γ be such that $\gamma \le 2^{-i_\gamma}$ and for any $i < i_\gamma$, $\gamma > 2^{-i_\gamma}$. Then using the above bound we see that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{f_t \sim \pi_t(z_{1:t-1})}\left[\mathbf{1}\left\{f_t(x_t) y_t < 0\right\}\right]\right] \le \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathbf{1}\left\{f(x_t) y_t < 2\gamma\right\} + \frac{2}{\gamma}\mathfrak{R}_T(\mathcal{F}) + \sqrt{T}\left(1 + 2\log\left(\frac{i\pi}{\sqrt{6}}\right)\right)$$

However note that $i_\gamma \le \log(1/\gamma)$ and so we can conclude that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{f_t \sim \pi_t(z_{1:t-1})}\left[\mathbf{1}\left\{f_t(x_t) y_t < 0\right\}\right]\right] \le \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathbf{1}\left\{f(x_t) y_t < 2\gamma\right\} + \frac{2}{\gamma}\mathfrak{R}_T(\mathcal{F}) + \sqrt{T}\left(1 + 2\log\left(\frac{\pi \log(1/\gamma)}{\sqrt{6}}\right)\right)$$

□

***Proof of Proposition 26*.** We shall prove that for any $i \in [k]$,

$$\mathfrak{R}_T(\mathcal{F}_i) \le 2LB_i\mathfrak{R}_T(\mathcal{F}_{i-1})$$

To see this note that

$$\begin{aligned}
\mathfrak{R}_T(\mathcal{F}_i) &= \sup_{\mathbf{x}} \mathbb{E}_{\epsilon}\left[\sup_{\substack{w^i : \|w^i\|_1 \le B_i \\ \forall j f_j \in \mathcal{F}_{i-1}}} \sum_{t=1}^{T} \epsilon_t \left(\sum_j w^i_j \sigma\left(f_j(\mathbf{x}_t(\epsilon))\right)\right)\right] \\
&\le \sup_{\mathbf{x}} \mathbb{E}_{\epsilon}\left[\sup_{\substack{w^i : \|w^i\|_1 \le B_i \\ \forall j f_j \in \mathcal{F}_{i-1}}} \|w^i\|_1 \max_j \left|\sum_{t=1}^{T} \epsilon_t \sigma\left(f_j(\mathbf{x}_t(\epsilon))\right)\right|\right] && \text{(Hölder's inequality)} \\
&\le \sup_{\mathbf{x}} \mathbb{E}_{\epsilon}\left[B_i \sup_{f \in \mathcal{F}_{i-1}} \left|\sum_{t=1}^{T} \epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right)\right|\right] \\
&= \sup_{\mathbf{x}} \mathbb{E}_{\epsilon}\left[B_i \sup_{f \in \mathcal{F}_{i-1}} \max\left\{\sum_{t=1}^{T} \epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right), -\sum_{t=1}^{T} \epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right)\right\}\right] \\
&= \sup_{\mathbf{x}} \mathbb{E}_{\epsilon}\left[B_i \max\left\{\sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^{T} \epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right), \sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^{T} -\epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right)\right\}\right] \\
&\le \sup_{\mathbf{x}} \mathbb{E}_{\epsilon}\left[B_i \sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^{T} \epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right)\right] + \sup_{\mathbf{x}} \mathbb{E}_{\epsilon}\left[B_i \sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^{T} -\epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right)\right] && (\sigma(0) = 0 \text{ and } 0 \in \mathcal{F}_i) \\
&= 2B_i \sup_{\mathbf{x}} \mathbb{E}_{\epsilon}\left[\sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^{T} \epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right)\right] && \text{(Proposition 21)} \\
&\le 2B_i L \sup_{\mathbf{x}} \mathbb{E}_{\epsilon}\left[\sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon))\right] && \text{(Lemma 18)} \\
&= 2B_i L \mathfrak{R}_T(\mathcal{F}_{i-1}) && (17)
\end{aligned}$$

To finish the proof we note that

$$
\begin{aligned}
\mathfrak{R}_T(\mathcal{F}_1) &= \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{w \in \mathbb{R}^d : \|w\|_1 \le B_1} \sum_{t=1}^T \epsilon_t w^\top \mathbf{x}_t(\epsilon) \right] \\
&\le \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{w \in \mathbb{R}^d : \|w\|_1 \le B_1} \|w\|_1 \left\| \sum_{t=1}^T \epsilon_t \mathbf{x}_t(\epsilon) \right\|_\infty \right] \\
&\le B_1 \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\max_{i \in [d]} \left\{ \sum_{t=1}^T \epsilon_t \mathbf{x}_t(\epsilon)[i] \right\} \right]
\end{aligned}
$$

Note that the instances $x \in \mathcal{X}$ are vectors in $\mathbb{R}^d$ and so for a given instance tree $\mathbf{x}$, for any $i \in [d]$, $\mathbf{x}[i]$ given by only taking the i^{th} co-ordinate is a valid real valued tree. Hence using Lemma 8 we conclude that

$$
\begin{aligned}
\mathfrak{R}_T(\mathcal{F}_1) &\le B_1 \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\max_{i \in [d]} \left\{ \sum_{t=1}^T \epsilon_t \mathbf{x}_t(\epsilon)[i] \right\} \right] \\
&\le B_1 \sqrt{2 T X_\infty^2 \log d}
\end{aligned}
$$

Using the above and Equation 17 we conclude the proof. □

Proof of Proposition 27. For a tree of depth d, the indicator function of a leaf is a conjunction of no more than d decision functions. More specifically, if the decision tree consists of decision nodes chosen from a class $\mathcal{H}$ of binary-valued functions, the indicator function of leaf l (which takes value 1 at a point x if x reaches l, and 0 otherwise) is a conjunction of d_l functions from $\mathcal{H}$, where d_l is the depth of leaf l. We can represent the function computed by the tree as the sign of

$$
g(x) = \sum_l w_l \sigma_l \bigwedge_{i=1}^{d_l} h_{l,i}(x)
$$

where the sum is over all leaves l, $w_l > 0$, $\sum_l w_l = 1$, $\sigma_l \in \{\pm 1\}$ is the label of leaf l, $h_{l,i} \in \mathcal{H}$, and the conjunction is understood to map to $\{0, 1\}$. Let $\mathcal{F}$ be this class of functions. Now note that if we fix some $L > 0$ then we see that the loss

$$
\phi_L(\alpha) = \begin{cases} 1 & \text{if } \alpha \le 0 \\ 1 - L\alpha & \text{if } 0 < \alpha \le 1/L \\ 0 & \text{otherwise} \end{cases}
$$

is L-Lipschitz and so by Theorem 2 and Lemma 18 we have that for every $L > 0$, there exists a randomized strategy π^L for the player, such that for any sequence $z_1 = (x_1, y_1), \ldots, z_T = (x_T, y_T)$,

$$
\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{f_t \sim \pi_t^L(z_{1:t-1})} \left[\phi_L(y_t f_t(x_t)) \right] \right] \le \inf_{t \in \mathcal{T}} \sum_{t=1}^T \phi_L(y_t t(x_t)) + L\mathfrak{R}(\mathcal{T})
$$

Now note that ϕ_L upper bounds the step function and so

$$
\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{f_t \sim \pi_t^L(z_{1:t-1})} \left[\mathbf{1}\left\{ f_t(x_t) \neq y_t \right\} \right] \right] \le \inf_{t \in \mathcal{T}} \sum_{t=1}^T \phi_L(y_t t(x_t)) + L\mathfrak{R}(\mathcal{T})
$$

Now say $t^* \in \mathcal{T}$ is the minimizer of $\sum_{t=1}^{T} \mathbf{1}\left\{t(x_t) \neq y_t\right\}$ then note that

$$\begin{aligned}
\sum_{t=1}^{T} \phi_L(y_t t^*(x_t)) &= \sum_{t=1}^{T} \mathbf{1}\left\{t(x_t) \neq y_t\right\} + \sum_{l} \tilde{C}_T(l)\phi_L(w_l) \\
&\leq \sum_{t=1}^{T} \mathbf{1}\left\{t^*(x_t) \neq y_t\right\} + \sum_{l} \tilde{C}_T(l) \max(0, 1 - Lw_l) \\
&\leq \sum_{t=1}^{T} \mathbf{1}\left\{t^*(x_t) \neq y_t\right\} + \sum_{l} \max\left(0, (1 - Lw_l)\tilde{C}_T(l)\right) \\
&= \inf_{t \in \mathcal{T}} \sum_{t=1}^{T} \mathbf{1}\left\{t(x_t) \neq y_t\right\} + \sum_{l} \max\left(0, (1 - Lw_l)\tilde{C}_T(l)\right)
\end{aligned}$$

Hence we see that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{f_t \sim \pi_t^L(z_{1:t-1})}\left[\mathbf{1}\left\{f_t(x_t) \neq y_t\right\}\right]\right] \leq \inf_{t \in \mathcal{T}} \sum_{t=1}^{T} \mathbf{1}\left\{t(x_t) \neq y_t\right\} + \sum_{l} \max\left(0, (1 - Lw_l)\tilde{C}_T(l)\right)$$

Now if we discretize over L as $L_i = i$ for all $i \in \mathbb{N}$ and run experts algorithm 3 with $\pi_1, \pi_2, \ldots$ as our experts and weight of expert π_i is $p_i = \frac{6}{\pi^2} i^{-2}$ so that $\sum_i p_i = 1$ then we get that for this randomized strategy P, we have from Proposition 29 that for all $L \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{f_t \sim \pi_t(z_{1:t-1})}\left[\mathbf{1}\left\{f_t(x_t) \neq y_t\right\}\right]\right] \\
&\leq \inf_{t \in \mathcal{T}} \sum_{t=1}^{T} \mathbf{1}\left\{t(x_t) \neq y_t\right\} + \sum_{l} \max\left(0, (1 - Lw_l)\tilde{C}_T(l)\right) + L\mathfrak{R}(\mathcal{T}) + \sqrt{T} + 2\sqrt{T}\log(L\pi/\sqrt{6})
\end{aligned}$$

Now we pick $L = |\{l : \tilde{C}_T(l) > 2\mathfrak{R}(\mathcal{T})\}| =: N_{\text{leaf}}$ and also pick $w_l = 0$ if $\tilde{C}_T(l) \leq 2\mathfrak{R}(\mathcal{T})$ and $w_l = 1/L$ otherwise. Hence we see that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{f_t \sim \pi_t(z_{1:t-1})}\left[\mathbf{1}\left\{f_t(x_t) \neq y_t\right\}\right]\right] &\leq \inf_{t \in \mathcal{T}} \sum_{t=1}^{T} \mathbf{1}\left\{t(x_t) \neq y_t\right\} + \sum_{l} \tilde{C}_T(l)\mathbf{1}\left\{\tilde{C}_T(l) \leq 2\mathfrak{R}(\mathcal{T})\right\} \\
&\quad + 2\mathfrak{R}(\mathcal{T}) \sum_{l} \mathbf{1}\left\{\tilde{C}_T(l) > 2\mathfrak{R}(\mathcal{T})\right\} + \sqrt{T} + 2\sqrt{T}\log(N_{\text{leaf}}\pi/\sqrt{6}) \\
&= \inf_{t \in \mathcal{T}} \sum_{t=1}^{T} \mathbf{1}\left\{t(x_t) \neq y_t\right\} + \sum_{l} \min(\tilde{C}_T(l), 2\mathfrak{R}(\mathcal{T})) + \sqrt{T}\left(1 + 2\log(N_{\text{leaf}}\pi/\sqrt{6})\right)
\end{aligned}$$

Now finally we can apply Corollary 20 to bound $\mathfrak{R}(\mathcal{T}) \leq d\mathcal{O}(\log^{3/2} T)\ \mathfrak{R}(\mathcal{H})$ and thus conclude the proof by plugging this into the above. □

Exponentially Weighted Average (EWA) Algorithm on Countable Experts

We consider here a version of the exponentially weighted experts algorithm for countable (possibly infinite) number of experts and provide a bound on the expected regret of the randomized algorithm. The proof of the result closely follows the finite case (e.g. [10, Theorem 2.2]).

Say we are provided with countable experts $E_1, E_2, \ldots$ where each expert can herself be thought of as a randomized/deterministic player strategy which, given history, produces an element of $\mathcal{F}$ at round t. Here

Algorithm 3 EWA $(E_1, E_2, \ldots, p_1, p_2, \ldots)$

```
Initialize each $w_i^1 \leftarrow p_i$
for $t = 1$ to $T$ do
  Pick randomly an expert $i$ with probability $w_i^t$
  Play $f_t = f_i^t$
  Receive $x_t$
  Update for each $i$, $w_i^{t+1} = \frac{w_i^t e^{-\eta f_i^t(x_t)}}{\sum_i w_i^t e^{-\eta f_i^t(x_t)}}$
end for
```

we also assume that $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ contains only non-negative functions. Denote by f_t^i the function output by expert i at round t given the history. The EWA algorithm we consider needs access to the countable set of experts and also needs an initial weighting on each expert $p_1, p_2, \ldots$ such that $\sum_i p_i = 1$.

Proposition 29. *For the exponentially weighted average forecaster (Algorithm 3) with $\eta = T^{-1/2}$ yields*

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t)\right] \le \sum_{t=1}^{T} f_i^t(x_t) + \frac{\sqrt{T}}{8} + \sqrt{T}\log\left(1/p_i\right)$$

for any $i \in \mathbb{N}$.

Proof. Define $W_t = \sum_i p_i e^{-\eta \sum_{j=1}^{t} f_i^j(x_t)}$. Then note that

$$\log\left(\frac{W_t}{W_{t-1}}\right) = \log\left(\frac{\sum_i p_i e^{-\eta \sum_{j=1}^{t} f_i^j(x_t)}}{W_{t-1}}\right) = \log\left(\sum_i w_i^{t-1} e^{-\eta f_i^t(x_t)}\right)$$

Now using Hoeffding's inequality (see [10, Lemma 2.2]) we have that

$$\log\left(\frac{W_t}{W_{t-1}}\right) \le -\eta \sum_i w_i^{t-1} f_i^t(x_t) + \frac{\eta^2}{8} = -\eta \mathbb{E}\left[f_t(x_t)\right] + \frac{\eta^2}{8}$$

Summing over t we get

$$\log(W_T) - \log(W_0) = \sum_{t=1}^{T} \log\left(\frac{W_t}{W_{t-1}}\right) \le -\eta \mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t)\right] + \frac{T\eta^2}{8} \tag{18}$$

Note that $W_0 = \sum_i p_i = 1$ and so $\log(W_0) = 0$. Also note that for any $i \in \mathbb{N}$,

$$\log(W_T) = \log\left(\sum_i p_i e^{-\eta \sum_{t=1}^{T} f_i^t(x_t)}\right) \ge \log\left(p_i^{-\eta \sum_{t=1}^{T} f_i^t(x_t)}\right) = \log(p_i) - \eta \sum_{t=1}^{T} f_i^t(x_t)$$

Hence using this with Equation 18 we see that

$$\log(p_i) - \eta \sum_{t=1}^{T} f_i^t(x_t) \le -\eta \mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t)\right] + \frac{T\eta^2}{8}$$

Rearranging we get

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t)\right] \le \sum_{t=1}^{T} f_i^t(x_t) + \frac{\eta T}{8} + \frac{1}{\eta}\log\left(\frac{1}{p_i}\right)$$

Using $\eta = \frac{1}{\sqrt{T}}$ we get the desired bound. □