Journal of Machine Learning Research 16 (2015) 155-186 Submitted 7/13; Revised 7/14; Published 2/15

Online Learning via Sequential Complexities

Alexander Rakhlin RAKHLIN@WHARTON.UPENN.EDU
Department of Statistics
University of Pennsylvania
Philadelphia, PA 19104

Karthik Sridharan SKARTHIK@WHARTON.UPENN.EDU
Department of Computer Science
Cornell University
Ithaca, NY 14853

Ambuj Tewari TEWARIA@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109

Editor: Mehryar Mohri

Abstract

We consider the problem of sequential prediction and provide tools to study the minimax value of the associated game. Classical statistical learning theory provides several useful complexity measures to study learning with i.i.d. data. Our proposed sequential complexities can be seen as extensions of these measures to the sequential setting. The developed theory is shown to yield precise learning guarantees for the problem of sequential prediction. In particular, we show necessary and sufficient conditions for online learnability in the setting of supervised learning. Several examples show the utility of our framework: we can establish learnability without having to exhibit an explicit online learning algorithm.

Keywords: online learning, sequential complexities, regret minimization

1. Introduction

This paper is concerned with sequential prediction problems where no probabilistic assumptions are made regarding the data generating mechanism. Our viewpoint is expressed well by the following quotation from Cover and Shenhar (1977):

> "We are interested in sequential prediction procedures that exploit any apparent order in the sequence. We do not assume the existence of any underlying distributions, but assume that the sequence is an outcome of a game against a malevolent intelligent nature."

We will, in fact, take the game theoretic viewpoint seriously. All our investigations will proceed by analyzing the minimax value of a repeated game between a *player* or *learner* and a "malevolent intelligent nature", or the *adversary*.

Even though we have the setting of prediction problems in mind, it will be useful to develop the theory in a somewhat abstract setting. Towards this end, fix the sets $\mathcal{F}$ and

$\mathcal{Z}$, as well as a loss function $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$, and consider the following T-round repeated two-player game, which we term the *online learning* or *sequential prediction* model. On round $t \in \{1, \ldots, T\}$, the learner chooses $f_t \in \mathcal{F}$, the adversary picks $z_t \in \mathcal{Z}$, and the learner suffers loss $\ell(f_t, z_t)$. At the end of T rounds we define *regret*

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

as the difference between the cumulative loss of the player and the cumulative loss of the best fixed decision. For the given pair $(\mathcal{F}, \mathcal{Z})$, the problem is said to be *online learnable* if there exists an algorithm for the learner such that regret grows sublinearly in the time horizon T, no matter what strategy the adversary employs.

The origin of the online learning (or sequential prediction) model can be traced back to the work of Robbins (1950) on compound statistical decision problems. Some of the earliest sequential prediction algorithms were proposed by Blackwell (1956a,b) and Hannan (1957). Blackwell's method was based on his celebrated approachability theorem whereas Hannan's was based on minimizing a randomly perturbed sum of previous losses. Hannan's ideas were to later resurface in the influential Follow-the-Perturbed-Leader family (Kalai and Vempala, 2005) of online learning algorithms. The seminal ideas in the work of Robbins, Blackwell and Hannan led to further developments in many different fields. Cover (1967), Davisson (1973), Ziv and Lempel (1977), Rissanen (1984), Feder et al. (1992), and others laid the foundation of universal coding, compression and prediction in the Information Theory literature. Within Computer Science, Littlestone and Warmuth (1994), Cesa-Bianchi et al. (1997), Vovk (1998), and others studied the online learning model and the prediction with expert advice framework. The connections between regret minimization and convergence to equilibria was studied in Economics by Foster and Vohra (1997), Hart and Mas-Colell (2000) and others.

We have no doubt left out many interesting works above. But even our partial list will convince the reader that research in online learning and sequential prediction has benefited from contributions by researchers from a variety of fields including Computer Science, Economics, Information Theory, and Statistics. For an excellent synthesis and presentation of results from these different fields we refer the reader to the book by Cesa-Bianchi and Lugosi (2006). Many of the ideas in the field are *constructive*, resulting in beautiful algorithms, or algorithmic techniques, associated with names such as Follow-the-Regularized-Leader, Follow-the-Perturbed-Leader, Weighted Majority, Hedge, and Online Gradient Descent. However, analyzing specific algorithms has obvious disadvantages. The algorithm may not be "optimal" for the task at hand. Even if it is optimal, one cannot prove that fact unless one develops tools for analyzing the inherent *complexity* of the online learning problem.

Our goal is precisely to provide such tools. We will begin by defining the minimax value of the game underlying the abstract online learning model. Then we will develop tools for controlling the minimax value resulting in a theory that parallels statistical learning theory. In particular, we develop analogues of combinatorial dimensions, covering numbers, and Rademacher complexities. We will also provide results relating these complexities.

Note that our approach is *non-constructive*: controlling the sequential complexities mentioned above will only guarantee the *existence* of a good online learning algorithm but

will not explicitly create one. However, it turns out that that the minimax point of view can indeed lead to constructive algorithms as shown by Rakhlin et al. (2012).

2. Minimax Value and Online Learnability

To proceed further in our analysis of the minimax value of the repeated game between the learner and the adversary, we need to make a few technical assumptions. We assume that $\mathcal{F}$ is a subset of a separable metric space. Let $\mathcal{Q}$ be the set of probability measures on $\mathcal{F}$ and assume that $\mathcal{Q}$ is weakly compact. In order to allow randomized prediction, we allow the learner to choose a distribution $q_t \in \mathcal{Q}$ on every round. The minimax value of the game is then defined as

$$\mathcal{V}_T(\mathcal{F},\mathcal{Z}) \triangleq \inf_{q_1\in\mathcal{Q}} \sup_{z_1\in\mathcal{Z}} \mathop{\mathbb{E}}_{f_1\sim q_1} \cdots \inf_{q_T\in\mathcal{Q}} \sup_{z_T\in\mathcal{Z}} \mathop{\mathbb{E}}_{f_T\sim q_T} \left[\sum_{t=1}^{T} \ell(f_t,z_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f,z_t)\right] . \tag{1}$$

Henceforth, the notation $\mathbb{E}_{f\sim q}$ stands for the expectation operator integrating out the random variable f with distribution q. We consider here the *adaptive* adversary who gets to choose each z_t based on the history of moves $f_{1:t-1}$ and $z_{1:t-1}$.

The first key step in the study of the value of the game is to appeal to the minimax theorem and exchange the pairs of infima and suprema in (1). This dual formulation is easier to analyze because the choice of the player comes *after* the choice of the mixed strategy of the adversary. We remark that the minimax theorem holds under a very general assumption of weak compactness of $\mathcal{Q}$ and lower semi-continuity of the loss function.[1] Under these conditions, we can appeal to Theorem 1 stated below, which is adapted for our needs from the work of Abernethy et al. (2009).

Theorem 1 *Let $\mathcal{F}$ and $\mathcal{Z}$ 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 measures (mixed strategies) on $\mathcal{F}$ and $\mathcal{Z}$, respectively. Then*

$$\mathcal{V}_T(\mathcal{F},\mathcal{Z}) = \sup_{p_1} \mathop{\mathbb{E}}_{z_1\sim p_1} \cdots \sup_{p_T} \mathop{\mathbb{E}}_{z_T\sim p_T} \left[\sum_{t=1}^{T} \inf_{f_t\in\mathcal{F}} \mathop{\mathbb{E}}_{z_t\sim p_t} [\ell(f_t,z_t)] - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f,z_t)\right], \tag{2}$$

where suprema over p_t range over all distributions in $\mathcal{P}$.

The question of learnability in the online learning model is now reduced to the study of $\mathcal{V}_T(\mathcal{F},\mathcal{Z})$, taking (2) as the starting point.

Definition 2 *A class $\mathcal{F}$ is said to be* online learnable *with respect to the given $\mathcal{Z}$ and ℓ if*

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

Note that our notion of learnability is related to, but distinct from, *Hannan consistency* (Hannan, 1957; Cesa-Bianchi and Lugosi, 2006). The latter notion requires the iterated game to go on for an infinite number of rounds and is formulated in terms of *almost sure*

1. We refer to Appendix A for a precise statement of the minimax theorem, as well as sufficient conditions.

convergence. In contrast, we consider a distinct game for each T and look at *expected* regret. Nevertheless, it is possible to obtain Hannan consistency using the techniques developed in this paper by considering a slightly different game (Rakhlin et al., 2011).

We also remark that the statements in this paper extend to the case when the learner is allowed to make decisions in a larger set $\mathcal{G}$, while the best-in-hindsight term in the regret definition is computed with respect to $\mathcal{F} \subseteq \mathcal{G}$. Such a setting—interesting especially with regard to computational concerns—is termed *improper learning*. For example, prediction with side information (or, the *supervised learning* problem) is one such case, where we choose $\mathcal{Y} \subset \mathbb{R}$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}} = \mathcal{G}$ and $\ell(f,(x,y)) = |f(x) - y|$. This setting will be studied later in the paper. Note that in the proper learning scenario, $\mathcal{V}_T(\mathcal{F},\mathcal{Z}) \geq 0$ (e.g., since all z_t's can be chosen to be the same), and thus the "lim sup" in Definition 2 can be simply replaced with the limit being equal to zero.

This paper is aimed at understanding the value of the game $\mathcal{V}_T(\mathcal{F},\mathcal{Z})$ for various function classes $\mathcal{F}$. Since our focus is on the complexity of $\mathcal{F}$, we shall often write $\mathcal{V}_T(\mathcal{F})$ keeping the dependence on $\mathcal{Z}$ (and ℓ) implicit. As we show, the sequential complexity notions—that were shown by Rakhlin et al. (2014) to characterize uniform martingale Laws of Large Numbers—also give us a handle on the value $\mathcal{V}_T(\mathcal{F})$. In the next section, we briefly define these sequential complexity notions and mention some of the key relations between them. A more detailed account of the relationships between sequential complexity measures along with complete proofs can be found in (Rakhlin et al., 2014).

3. Sequential Complexities

Unlike the well-studied statistical learning scenario with i.i.d. data, the online learning problem possesses a certain sequential dependence. Such dependence cannot be captured by classical notions of complexity that are based on a batch of data given as a *tuple* of T examples. A basic unit that does capture temporal dependence is a binary tree. Surprisingly, for the sequential prediction problems considered in this paper, one need not look further than binary trees to capture the relevant complexity.

A *$\mathcal{Z}$-valued tree* $\mathbf{z}$ *of depth* T is a complete rooted binary tree with nodes labeled by elements of $\mathcal{Z}$. Such a tree $\mathbf{z}$ is identified with the sequence $(\mathbf{z}_1,\dots,\mathbf{z}_T)$ of labeling functions $\mathbf{z}_i : \{\pm 1\}^{i-1} \to \mathcal{Z}$ which provide the labels for each node. Therefore, $\mathbf{z}_1 \in \mathcal{Z}$ is the label for the *root* of the tree, while $\mathbf{z}_i$ for $i > 1$ is the label of the node obtained by following the path of length $i-1$ from the root, with $+1$ indicating 'right' and -1 indicating 'left'. A *path* of length T is given by the sequence $\epsilon = (\epsilon_1,\dots,\epsilon_T) \in \{\pm 1\}^T$. For brevity, we shall often write $\mathbf{z}_t(\epsilon)$, where $\epsilon = (\epsilon_1,\dots,\epsilon_T)$, but it is understood that $\mathbf{z}_t$ depends only on the prefix $(\epsilon_1,\dots,\epsilon_{t-1})$.

Now, let $\epsilon_1,\dots,\epsilon_T$ be independent Rademacher random variables. Given a $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth T, we define the *sequential Rademacher complexity* of a function class $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ on a $\mathcal{Z}$-valued tree $\mathbf{z}$ as

$$\mathfrak{R}_T(\mathcal{G},\mathbf{z}) \triangleq \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T} \epsilon_t g(\mathbf{z}_t(\epsilon))\right],$$

and we denote by $\mathfrak{R}_T(\mathcal{G}) = \sup_{\mathbf{z}} \mathfrak{R}_T(\mathcal{G},\mathbf{z})$ its supremum over all $\mathcal{Z}$-valued trees of depth T. The importance of the introduced notion stems from the following result (Rakhlin et al.,

2014, Theorem 2): for any distribution over a sequence $(Z_1, \ldots, Z_T)$, we have

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[g(Z_t)|Z^{t-1}\right] - g(Z_t)\right)\right] \le 2\,\mathfrak{R}_T(\mathcal{G}) \ , \tag{3}$$

where $Z^{t-1} = (Z_1, \ldots, Z_{t-1})$. In other words, the martingale version of the uniform deviations of means from expectations is controlled by the worst-case sequential Rademacher complexity. A matching lower bound also holds for the supremum over distributions on sequences in $\mathcal{Z}^T$. It then follows that a uniform martingale Law of Large Numbers holds for $\mathcal{G}$ if and only if $\mathfrak{R}_T(\mathcal{G}) \to 0$. For i.i.d. random variables, a similar statement can be made in terms of the classical Rademacher complexity, and so one might hope that many other complexity notions from empirical process theory have martingale (or we may say, sequential) analogues. Luckily, this is indeed the case (Rakhlin et al., 2014). As we show in this paper, these generalizations of the classical notions also give a handle on (as well as necessary and sufficient conditions for) online learnability, thus painting a picture that completely parallels statistical learning theory. But before we present our main results, let us recall some key definitions and results in (Rakhlin et al., 2014).

In providing further upper bounds on sequential Rademacher complexity, the following definitions of an "effective size" of a function class generalize the classical notions of a covering number. A set V of $\mathbb{R}$-valued trees of depth T is *a (sequential) α-cover* (with respect to ℓ_p norm) of $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ on a tree $\mathbf{z}$ of depth T if

$$\forall g \in \mathcal{G},\ \forall \epsilon \in \{\pm 1\}^T,\ \exists \mathbf{v} \in V \text{ s.t.} \quad \left(\frac{1}{T}\sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - g(\mathbf{z}_t(\epsilon))|^p\right)^{1/p} \le \alpha.$$

The *(sequential) covering number* of a function class $\mathcal{G}$ on a given tree $\mathbf{z}$ is defined as

$$\mathcal{N}_p(\alpha, \mathcal{G}, \mathbf{z}) \ \triangleq\ \min\left\{|V| : V \text{ is an } \alpha\text{-cover w.r.t. } \ell_p \text{ norm of } \mathcal{G} \text{ on } \mathbf{z}\right\}.$$

It is straightforward to check that $\mathcal{N}_p(\alpha, \mathcal{G}, \mathbf{z}) \le \mathcal{N}_q(\alpha, \mathcal{G}, \mathbf{z})$ whenever $1 \le p \le q \le \infty$.

Further define $\mathcal{N}_p(\alpha, \mathcal{G}, T) = \sup_{\mathbf{z}} \mathcal{N}_p(\alpha, \mathcal{G}, \mathbf{z})$, the maximal ℓ_p covering number of $\mathcal{G}$ over depth T trees. For a class $\mathcal{G}$ of binary-valued functions, we also define a so-called 0-*cover* (or, cover at scale 0), denoted by $\mathcal{N}(0, \mathcal{G}, \mathbf{z})$, as equal to any $\mathcal{N}_p(0, \mathcal{G}, \mathbf{z})$. The definition of a 0-cover can be seen as the correct analogue of the *size of a projection* of $\mathcal{G}$ onto a tuple of points in the i.i.d. case. The size of this projection in the i.i.d. case was the starting point of the work of Vapnik and Chervonenkis.

When $\mathcal{G} \subseteq [-1, 1]^{\mathcal{Z}}$ is a *finite* class of bounded functions, one can show (Rakhlin et al., 2014, Lemma 1) that

$$\mathfrak{R}_T(\mathcal{G}, \mathbf{z}) \le \sqrt{\frac{2\log|\mathcal{G}|}{T}}, \tag{4}$$

a bound that should (correctly) remind the reader of the Exponential Weights regret bound. With the definition of an α-cover with respect to ℓ_1 norm, one can easily extend (4) beyond the finite case. Immediately from the definition of ℓ_1 covering number, it follows that for any $\mathcal{G} \subseteq [-1, 1]^{\mathcal{Z}}$, for any $\alpha > 0$,

$$\mathfrak{R}_T(\mathcal{G}, \mathbf{z}) \le \alpha + \sqrt{\frac{2\log\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})}{T}} \tag{5}$$

(Rakhlin et al., 2014, Eq. 9). A tighter control is obtained by integrating the covering numbers at different scales. To this end, consider the following analogue of the Dudley entropy integral bound. For $p \geq 1$, the *integrated complexity* of a function class $\mathcal{G} \subseteq [-1,1]^{\mathcal{Z}}$ on a $\mathcal{Z}$-valued tree of depth T is defined as

$$\mathfrak{D}_T^p(\mathcal{G}, \mathbf{z}) \triangleq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{T}} \int_\alpha^1 \sqrt{\log \mathcal{N}_p(\delta, \mathcal{G}, \mathbf{z})}\, d\delta \right\} \tag{6}$$

and $\mathfrak{D}_T^p(\mathcal{G}) = \sup_{\mathbf{z}} \mathfrak{D}_T^p(\mathcal{G}, \mathbf{z})$, with $\mathfrak{D}_T^2(\mathcal{G}, \mathbf{z})$ denoted simply by $\mathfrak{D}_T(\mathcal{G}, \mathbf{z})$. We have previously shown (Rakhlin et al., 2014, Theorem 3) that, for any function class $\mathcal{G} \subseteq [-1,1]^{\mathcal{Z}}$ and any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth T,

$$\mathfrak{R}_T(\mathcal{G}, \mathbf{z}) \leq \mathfrak{D}_T(\mathcal{G}, \mathbf{z}). \tag{7}$$

We next turn to the description of sequential combinatorial parameters. A $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth d is *shattered* by a function class $\mathcal{G} \subseteq \{\pm 1\}^{\mathcal{Z}}$ if for all $\epsilon \in \{\pm 1\}^d$, there exists $g \in \mathcal{G}$ such that $g(\mathbf{z}_t(\epsilon)) = \epsilon_t$ for all $t \in [d]$. The *Littlestone dimension* $\mathrm{Ldim}(\mathcal{G}, \mathcal{Z})$ is the largest positive integer d such that $\mathcal{G}$ shatters a $\mathcal{Z}$-valued tree of depth d (Littlestone, 1988; Ben-David et al., 2009). The scale-sensitive version of Littlestone dimension is defined as follows. A $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth d is *α-shattered* by a function class $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ if there exists an $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d such that

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

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

The notions introduced above can be viewed as sequential generalizations of the VC dimension and the fat-shattering dimension where tuples of points get replaced by complete binary trees. In fact, one recovers the classical notions if the tree $\mathbf{z}$ in the above definitions is restricted to have the same values within a level (hence, no temporal dependence). Crucially, the sequential combinatorial analogues provide control for the growth of sequential covering numbers, justifying the definitions.

First, let $\mathcal{G} \subseteq \{0, \dots, k\}^{\mathcal{Z}}$ be a class of functions with $\mathrm{fat}_2(\mathcal{G}) = d$. Then, it can be shown (Rakhlin et al., 2014, Theorem 4) that for any $T \geq 1$,

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

For the second result (Rakhlin et al., 2014, Corollary 1), suppose $\mathcal{G}$ is a class of $[-1,1]$-valued functions on $\mathcal{Z}$. Then, for any $\alpha > 0$, and any $T \geq 1$,

$$\mathcal{N}_\infty(\alpha, \mathcal{G}, T) \leq \left(\frac{2eT}{\alpha}\right)^{\mathrm{fat}_\alpha(\mathcal{G})} . \tag{8}$$

Finally, we recall a bound on the size of the 0-cover in terms of the fat_1 combinatorial parameter (Rakhlin et al., 2014, Theorem 5). For a class $\mathcal{G} \subseteq \{0, \dots, k\}^{\mathcal{Z}}$ with $\mathrm{fat}_1(\mathcal{G}) = d$, we have

$$\mathcal{N}(0, \mathcal{G}, T) \leq \sum_{i=0}^{d} \binom{T}{i} k^i \leq (ekT)^d \ . \tag{9}$$

In particular, for $k = 1$ (that is, binary classification) we have $\mathrm{fat}_1(\mathcal{G}) = \mathrm{Ldim}(\mathcal{G})$. The inequality (9) is therefore a sequential analogue of the celebrated Vapnik-Chervonenkis-Sauer-Shelah lemma.

4. Structural Properties

For the examples developed in this paper, it will be crucial to exploit a number of useful properties that $\mathfrak{R}_T(\mathcal{G})$ satisfies. These properties allow one to establish online learnability for complex function classes even if no explicit learning algorithms are available.

We first state some properties that are easily proved but are nevertheless very useful.

Lemma 3 *Let $\mathcal{F}, \mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ and let $\mathrm{conv}(\mathcal{G})$ denote the convex hull of $\mathcal{G}$. Let $\mathbf{z}$ be any $\mathcal{Z}$-valued tree of depth T. Then the following properties hold.*

1. *If $\mathcal{F} \subseteq \mathcal{G}$, then $\mathfrak{R}_T(\mathcal{F}, \mathbf{z}) \le \mathfrak{R}_T(\mathcal{G}, \mathbf{z})$.*
2. *$\mathfrak{R}_T(\mathrm{conv}(\mathcal{G}), \mathbf{z}) = \mathfrak{R}_T(\mathcal{G}, \mathbf{z})$*
3. *$\mathfrak{R}_T(c\mathcal{G}, \mathbf{z}) = |c|\mathfrak{R}_T(\mathcal{G}, \mathbf{z})$ for all $c \in \mathbb{R}$.*
4. *For any $h : \mathcal{Z} \to \mathbb{R}$, $\mathfrak{R}_T(\mathcal{G} + h, \mathbf{z}) = \mathfrak{R}_T(\mathcal{G}, \mathbf{z})$ where $\mathcal{G} + h = \{g + h : g \in \mathcal{G}\}$.*

These properties match those of the classical Rademacher complexity (Bartlett and Mendelson, 2003) and can be proved in essentially the same way (we therefore skip the straightforward proofs).

The next property is a key tool for many of the applications: it allows us to bound the sequential Rademacher complexity for the Cartesian product of function classes composed with a Lipschitz mapping in terms of complexities of the individual classes.

Lemma 4 *Let $\mathcal{G} = \mathcal{G}_1 \times \ldots \times \mathcal{G}_k$ where each $\mathcal{G}_j \subseteq [-1,1]^{\mathcal{Z}}$. Further, let $\phi : \mathbb{R}^k \times \mathcal{Z} \to \mathbb{R}$ be such that $\phi(\cdot, z)$ is L-Lipschitz with respect to $\|\cdot\|_\infty$ for all $z \in \mathcal{Z}$, and let*

$$\phi \circ \mathcal{G} = \{z \mapsto \phi((g_1(z), \ldots, g_k(z)), z) : g_j \in \mathcal{G}_j\}.$$

Then we have

$$\mathfrak{R}_T(\phi \circ \mathcal{G}) \le 8\, L \left(1 + 4\sqrt{2}\log^{3/2}(eT^2)\right) \textstyle\sum_{j=1}^k \mathfrak{R}_T(\mathcal{G}_j)$$

as long as $\mathfrak{R}_T(\mathcal{G}_j) \ge 1/T$ for each j.

Let us explicitly state the more familiar contraction property, an immediate corollary of the above result.

Corollary 5 *Fix a class $\mathcal{G} \subseteq [-1,1]^{\mathcal{Z}}$ with $\mathfrak{R}_T(\mathcal{G}) \ge 1/T$ and a function $\phi : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$. Assume $\phi(\cdot, z)$ is L-Lipschitz for all $z \in \mathcal{Z}$. Then*

$$\mathfrak{R}_T(\phi \circ \mathcal{G}) \le 8\, L \left(1 + 4\sqrt{2}\log^{3/2}(eT^2)\right) \cdot \mathfrak{R}_T(\mathcal{G}),$$

where $\phi \circ \mathcal{G} = \{z \mapsto \phi(g(z), z) : g \in \mathcal{G}\}$.

We state another useful corollary of Lemma 4.

Corollary 6 *For a fixed binary function $b : \{\pm 1\}^k \to \{\pm 1\}$ and classes $\mathcal{G}_1, \dots, \mathcal{G}_k$ of $\{\pm 1\}$-valued functions,*

$$\mathfrak{R}_T(b(\mathcal{G}_1, \dots, \mathcal{G}_k)) \le \mathcal{O}\left(\log^{3/2}(T)\right) \sum_{j=1}^k \mathfrak{R}_T(\mathcal{G}_j).$$

Note that, in the classical case, the Lipschitz contraction property holds without any extra poly-logarithmic factors in T (Ledoux and Talagrand, 1991). It is an open question whether the poly-logarithmic factors can be removed in the results above. It is worth pointing out ahead of time that Theorem 8 below—in the setting of supervised learning with convex Lipschitz loss—does allow us to avoid the extraneous factor that would otherwise appear from a combination of Theorem 7 and Corollary 5.

5. Main Results

We now relate the value of the game to the worst case expected value of the supremum of an empirical process. However, unlike empirical processes that involve i.i.d. sums, our process involves a sum of *martingale differences*. In view of (3), the expected supremum can be further upper-bounded by the sequential Rademacher complexity.

Theorem 7 *The minimax value is bounded as*

$$\frac{1}{T}\mathcal{V}_T(\mathcal{F}) \le \sup_{\mathbb{P}} \mathbb{E} \sup_{g \in \ell(\mathcal{F})} \left[\frac{1}{T} \sum_{t=1}^T \left(\mathbb{E}[g(Z_t)|Z_1, \dots, Z_{t-1}] - g(Z_t) \right) \right] \le 2\,\mathfrak{R}_T(\ell(\mathcal{F})),$$

where $\ell(\mathcal{F}) = \{\ell(f, \cdot) : f \in \mathcal{F}\}$ and the supremum is taken over all distributions $\mathbb{P}$ over $(Z_1, \dots, Z_T)$.

We can now employ the tools developed earlier in the paper to upper bound the value of the game. Interestingly, any non-trivial upper bound guarantees *existence* of a prediction strategy that has sublinear regret irrespective of the sequence of the moves of the adversary. This complexity-based approach of establishing learnability should be contrasted with the purely algorithm-based approaches found in the literature.

5.1 Supervised Learning

In this subsection we study the *supervised learning problem* mentioned earlier in the paper. In this improper learning scenario, the learner at time t picks a function $f_t : \mathcal{X} \to \mathbb{R}$ and the adversary provides the input target pair $z_t = (x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} \subset \mathbb{R}$. In particular, the *binary classification* problem corresponds to the case $\mathcal{Y} = \{\pm 1\}$. Let $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ and let us fix the absolute value loss function $\ell(\hat{y}, y) = |\hat{y} - y|$. While we focus on the absolute loss, it is easy to see that all the results hold (with modified rates) for any loss $\ell(\hat{y}, y)$ such that for all $\hat{y}$ and y, $\phi(\ell(\hat{y}, y)) \le |\hat{y} - y| \le \Phi(\ell(\hat{y}, y))$ where Φ and ϕ are monotonically increasing functions. For instance, the squared loss $(\hat{y} - y)^2$ is a classic example.

We now observe that the value of the improper supervised learning game can be equivalently written as

$$\mathcal{V}_T^{\mathrm{S}}(\mathcal{F}) = \sup_{x_1} \inf_{q_1 \in \tilde{\mathcal{Q}}} \sup_{y_1} \mathop{\mathbb{E}}_{\hat{y}_1 \sim q_1} \cdots \sup_{x_T} \inf_{q_T \in \tilde{\mathcal{Q}}} \sup_{y_T} \mathop{\mathbb{E}}_{\hat{y}_T \sim q_T} \left[\sum_{t=1}^T \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^T \ell(f(x_t), y_t) \right], \qquad (10)$$

where $\tilde{\mathcal{Q}}$ denotes the set of probability distributions over $\mathcal{Y}$ and $\hat{y}_t$ has distribution q_t. This equivalence is easy to verify: we may view the choice $f_t : \mathcal{X} \to \mathcal{Y}$ as pre-specifying predictions $f_t(x)$ for all the possible $x \in \mathcal{X}$, while alternatively we can simply make the choice $\hat{y}_t \in \mathcal{Y}$ having observed the particular move $x_t \in \mathcal{X}$. The advantage of rewriting the game in the form (10) is that the minimax theorem only needs to be applied to the pair $\hat{y}_t$ and y_t, given the fixed choice x_t. The minimax theorem then holds even if weak compactness cannot be shown for the set of distributions on the original space of functions of the type $\mathcal{X} \to \mathcal{Y}$.

An examination of the proof of Theorem 7 reveals that the value (10) is upper bounded in exactly the same way, and the side information simply appears as an additional tree $\mathbf{x}$ in sequential Rademacher complexity, giving us:

$$\frac{1}{T}\mathcal{V}_T^{\mathrm{S}}(\mathcal{F}) \le 2 \sup_{\mathbf{x},\mathbf{y}} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \epsilon_t \ell(f(\mathbf{x}_t(\epsilon)), \mathbf{y}_t(\epsilon))\right] . \tag{11}$$

However, for the supervised learning setting, we can strengthen Theorem 7. The following theorem allows us to remove any convex Lipschitz loss (including the absolute loss) before passing to the sequential Rademacher complexity.

Theorem 8 *Let $\mathcal{Y} = [-1, 1]$ and suppose, for any $y \in \mathcal{Y}$, $\ell(\cdot, y)$ is convex and L-Lipschitz. Then the minimax value of a supervised learning problem is upper bounded as*

$$\frac{1}{T}\mathcal{V}_T^{S}(\mathcal{F}) \le 2\, L\, \mathfrak{R}_T(\mathcal{F}).$$

We remark that the contraction property for sequential Rademacher complexity, stated in Section 4, yields an extraneous logarithmic factor when applied to (11); here, we achieve the desired bound by removing the Lipschitz function directly during the symmetrization step.

Armed with the theorem, we now prove the following result.

Proposition 9 *Consider the supervised learning problem with a function class $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$ and absolute loss $\ell(\hat{y}, y) = |\hat{y} - y|$. Then, for any $T \ge 1$, we have*

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

where $\mathrm{fat}_\alpha = \mathrm{fat}_\alpha(\mathcal{F})$.

The proposition above implies that finiteness of the fat-shattering dimension at all scales is *necessary and sufficient* for online learnability of the supervised learning problem. Further, all the complexity notions introduced so far are within a poly-logarithmic factor from each other whenever the problem is learnable. These results are summarized in the next theorem:

Theorem 10 *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 with absolute loss.*
2. *Sequential Rademacher complexity satisfies* $\lim_{T\to\infty} \mathfrak{R}_T(\mathcal{F}) = 0$.
3. *For any* $\alpha > 0$, *the scale-sensitive dimension* $\mathrm{fat}_\alpha(\mathcal{F})$ *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* $\mathfrak{R}_T(\mathcal{F})$, *and the integrated complexity* $\mathfrak{D}_T(\mathcal{F})$ *are within a multiplicative factor of* $\mathcal{O}(\log^{3/2} T)$ *of each other.*

Remark 11 *Additionally, the three statements of Theorem 10 are equivalent to $\mathcal{F}$ satisfying a martingale version of the uniform Law of Large Numbers. This property is termed* Sequential Uniform Convergence *by Rakhlin et al. (2014), and we refer to their paper for more details.*

For binary classification, we write $\mathcal{V}_T^{\mathrm{Binary}}$ for $\mathcal{V}_T^{\mathrm{S}}$. This case has been investigated thoroughly by Ben-David et al. (2009) and indeed served as a key motivation for this paper. As a consequence of Proposition 9 and (9), we have a tight control on the value of the game for the binary classification problem. Note that the absolute loss in the binary classification setting is simply the 0-1 loss $\ell(\hat{y}, y) = \mathbf{1}\{\hat{y} \neq y\}$, where $\mathbf{1}\{\mathcal{U}\}$ is 1 if $\mathcal{U}$ is true and 0 otherwise.

Corollary 12 *For the binary classification problem with function class $\mathcal{F}$ and the* 0-1 *loss, we have*

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

for some universal constants $K_1, K_2 > 0$.

Both the upper and the lower bound in the above result were originally derived in Ben-David et al. (2009). Notably, we achieved the same bounds non-constructively through purely combinatorial and covering number arguments.

It is natural to ask whether being able to learn in the online model is different from learning in the i.i.d. model (in the distribution-free supervised setting). The standard example that exhibits a gap between the two frameworks (e.g., Littlestone, 1988; Ben-David et al., 2009) is binary classification using the class of step functions

$$\mathcal{F} = \{f_\theta(x) = \mathbf{1}\{x \le \theta\} : \theta \in [0,1]\}$$

on $[0,1]$. This class has VC dimension 1, but is *not* learnable in the online setting. Indeed, it is possible to verify that the Littlestone dimension is infinite. Interestingly, the closely-related class of "ramp" functions with slope $L > 0$

$$\mathcal{F}_L = \left\{f_\theta(x) = \mathbf{1}\{x \le \theta\} + (1 - L(x-\theta))\mathbf{1}\{\theta < x \le \theta + 1/L\} : \theta \in [0,1]\right\}$$

is learnable (say for supervised learning using absolute loss) in the online setting (and hence also in the i.i.d. case). Furthermore, the larger class of all bounded *L-Lipschitz* functions on a bounded interval is also online learnable (see Eq. 14 and proof of Proposition 18). Once again, we are able to make these statements from purely complexity-based considerations, without exhibiting an algorithm. Further examples where we can demonstrate online learnability are explored in Section 6.

5.2 Online Convex Optimization

Over the past decade, Online Convex Optimization (OCO) has emerged as a unified online learning framework (Zinkevich, 2003; Shalev-Shwartz, 2011). Various methods, such as Exponential Weights, can be viewed as instances of online mirror descent, solving the associated OCO problem. Much research effort has been devoted to understanding this abstract and simplified setting. It is tempting to say that any problem of online learning, as defined in the Introduction, can be viewed as OCO (in fact, online *linear* optimization) over the set of probability distributions; however, one should also recognize that by linearizing the problem, any interesting structure is lost and one instead suffers from the possibly unnecessary dependence on the number of functions in the class $\mathcal{F}$. Nevertheless, OCO is a central part of the recent literature, and we will study this scenario using techniques developed in this paper.

The standard setting of online convex optimization is as follows. The set of moves of the learner $\mathcal{F}$ is a bounded closed convex subset of a Banach space $(\mathcal{B}, \|\cdot\|)$ with $\|f\| \le D$ for all $f \in \mathcal{F}$ (the reader can think of $\mathbb{R}^d$ equipped with an ℓ_p norm for simplicity). Let $\|\cdot\|_\star$ be the dual norm. The adversary's set $\mathcal{Z}$ consists of convex G-Lipschitz (with respect to $\|\cdot\|_\star$) functions over $\mathcal{F}$:

$$\mathcal{Z} = \mathcal{Z}_{\mathrm{cvx}} = \{g : \mathcal{F} \to \mathbb{R} \,:\, g \text{ convex and } G\text{-Lipschitz w.r.t. } \|\cdot\|_\star\} \ .$$

Let the loss function be $\ell(f, g) = g(f)$, the evaluation of the adversarially chosen function at f. For the particular case of online *linear* optimization, we instead take

$$\mathcal{Z} = \mathcal{Z}_{\mathrm{lin}} = \{f \mapsto \langle f, z\rangle \,:\, \|z\|_\star \le G\}$$

with $\mathcal{Z}$ now a subset of the dual space. It is well-known (e.g., Abernethy et al., 2008) that the online convex optimization problem (without further assumptions on the functions in $\mathcal{Z}_{\mathrm{cvx}}$) is as hard as the corresponding linear optimization problem with $\mathcal{Z}_{\mathrm{lin}}$ if one considers deterministic algorithms. The same trivially extends to randomized methods:

Lemma 13 *Suppose $\mathcal{F}, \mathcal{Z}_{\mathrm{cvx}}, \mathcal{Z}_{\mathrm{lin}}$ be defined as above. Then we have*

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

We will now show how to use the above result to derive minimax regret guarantees for OCO. The reader may wonder why we do not directly try to bound the value $\mathcal{V}_T(\mathcal{F}, \mathcal{Z}_{\mathrm{cvx}})$ by $\mathfrak{R}_T(\mathcal{F}, \mathcal{Z}_{\mathrm{cvx}})$. In fact, this proof strategy cannot give a non-trivial bound if $\mathcal{F}$ is a subset of a high-dimensional (or infinite-dimensional) space (Shalev-Shwartz et al., 2009, Sec. 4.1). Instead, we use the lemma above to bound the value of the game where adversary plays convex functions with that of the game where adversary plays linear functions.

A function $\Psi : \mathcal{F} \to \mathbb{R}$ is (σ, q)-uniformly convex (for $q \in [2, \infty)$) on $\mathcal{F}$ with respect to a norm $\|\cdot\|$ if, for all $\theta \in [0, 1]$ and $f_1, f_2 \in \mathcal{F}$,

$$\Psi(\theta f_1 + (1-\theta) f_2) \le \theta \Psi(f_1) + (1-\theta)\Psi(f_2) - \frac{\sigma\, \theta\, (1-\theta)}{q} \|f_1 - f_2\|^q \ .$$

A $(\sigma, 2)$-uniformly convex function will be called σ-strongly convex.

We will give examples shortly but we first state a proposition that is useful to bound the sequential Rademacher complexity of linear function classes. The crucial duality fact exploited in its proof is that Ψ is (σ, q)-uniformly convex with respect to $\|\cdot\|$ if and only if $\Psi^\star$ is $(1/\sigma, p)$-uniformly smooth with respect to $\|\cdot\|_\star$ where $1/p + 1/q = 1$.

Proposition 14 (Rakhlin et al., 2014) *Let $\mathcal{F}$ be a subset of some Banach space $\mathcal{B}$ with norm $\|\cdot\|$ and let $\mathcal{Z}$ be a subset of the dual space $\mathcal{B}^\star$ equipped with norm $\|\cdot\|_\star$. Suppose that $\Psi : \mathcal{F} \to \mathbb{R}$ is (σ, q)-uniformly convex with respect to $\|\cdot\|$ and $0 \le \Psi(f) \le \Psi_{\max}$ for all $f \in \mathcal{F}$. Then we have*

$$\mathfrak{R}_T(\mathcal{F}) \le C_p \|\mathcal{Z}\|_\star \left(\frac{\Psi_{\max}^{p-1}}{\sigma T^{p-1}} \right)^{1/p},$$

where $\|\mathcal{Z}\|_\star = \sup_{z \in \mathcal{Z}} \|z\|_\star$, p is such that $1/p + 1/q = 1$, and $C_p = (p/(p-1))^{\frac{p-1}{p}}$.

Using the above Proposition in conjunction with Lemma 13 and Theorem 7, we can immediately conclude that

$$\mathcal{V}_T(\mathcal{F}, \mathcal{Z}_{\text{cvx}}) \le 2\,T\,\mathfrak{R}_T(\mathcal{F}) \le 2G\sqrt{\frac{2\,\Psi_{\max}\,T}{\sigma}}$$

for any non-negative function $\Psi : \mathcal{F} \to \mathbb{R}$ that is σ-strongly convex 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 bound becomes $2GD\sqrt{T}$ and recovers the guarantee for the online gradient descent algorithm. In general, for $\|\cdot\| = \|\cdot\|_p$ and $\|\cdot\|_\star = \|\cdot\|_q$, we can use $\Psi(u) = \frac{1}{2}\|u\|_p^2$ to get a bound of $2GD\sqrt{T/(p-1)}$ since Ψ is $(p-1)$-strongly convex w.r.t. $\|\cdot\|_p$. These $\mathcal{O}(\sqrt{T})$ regret rates are not new but we re-derive them to illustrate the usefulness of the tools we developed.

6. Further Examples

Now we present some further applications of the tools we have developed in this paper for some specific learning problems. To begin, we show how to bound the sequential Rademacher complexity of functions computed by neural networks. Then, we derive margin based regret bounds in a fairly general setting. The classical analogues of these margin bounds have played a big role in the modern theory of supervised learning where they help explain the success of linear classifiers in high dimensional spaces (e.g., Schapire et al., 1997; Koltchinskii and Panchenko, 2002). We then study the complexity of classes formed by decision trees, analyze the setting of transductive learning, and consider an online version of the Isotron problem. Finally, we make a connection to the seminal work of Cesa-Bianchi and Lugosi (1999) by re-deriving their bound on the minimax regret in a static experts game in terms of the classical Rademacher averages.

6.1 Neural Networks

We provide below a bound on the sequential Rademacher complexity for classic multi-layer neural networks thus showing they are learnable in the online setting. The model of neural

networks we consider below and the bounds we provide are analogous to the ones considered in the i.i.d. setting by Bartlett and Mendelson (2003).

Consider a k-layer 1-norm neural network, defined by a base function class $\mathcal{F}_1$ and, recursively, for each $2 \le i \le k$,

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

where σ is a Lipschitz transfer function, such as the sigmoid function.

Proposition 15 *Suppose $\sigma : \mathbb{R} \to [-1,1]$ is L-Lipschitz with $\sigma(0) = 0$. Then it holds that*

$$\mathfrak{R}_T(\mathcal{F}_k) \le \left(\prod_{i=2}^{k} 16 B_i \right) L^{k-1} \left(1 + 4\sqrt{2}\log^{3/2}(eT^2)\right)^k \mathfrak{R}_T(\mathcal{F}_1).$$

In particular, for the case of

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

and $\mathcal{X} \subset \mathbb{R}^d$ we have the bound

$$\mathfrak{R}_T(\mathcal{F}_k) \le \left(\prod_{i=1}^{k} 16 B_i \right) L^{k-1} \left(1 + 4\sqrt{2}\log^{3/2}(eT^2)\right)^k X_\infty \sqrt{\frac{2\log d}{T}}$$

where X_∞ is such that $\forall x \in \mathcal{X}$, $\|x\|_\infty \le X_\infty$.

Our result is a non-constructive guarantee, and, to the best of our knowledge, no algorithms for learning neural networks within the online learning model exist. It is not clear if the above bounds could be obtained via computationally efficient methods.

6.2 Margin Based Regret

In the classical statistical setting, margin bounds provide guarantees on the 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 Schapire et al., 1997; Koltchinskii and Panchenko, 2002). Recently, in the online setting, bounds of a similar flavor have been shown through the concept of margin via the Littlestone dimension (Ben-David et al., 2009). We show that our machinery can easily lead to margin bounds for binary classification problems for general function classes $\mathcal{F}$ based on their sequential Rademacher complexity. We use ideas from (Koltchinskii and Panchenko, 2002) to do this.

Proposition 16 *For any function class $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$, there exists a randomized prediction strategy given by τ such that for any sequence $z_1, \ldots, z_T$ where each $z_t = (x_t, y_t) \in \mathcal{X} \times \{\pm 1\}$,*

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{y}_t \sim \tau_t(z_{1:t-1})} \left[\mathbf{1}\left\{ \hat{y}_t y_t < 0 \right\} \right]$$
$$\le \inf_{\gamma > 0} \left\{ \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathbf{1}\left\{ f(x_t) y_t < 2\gamma \right\} + \frac{16}{\gamma} \left(1 + 4\sqrt{2}\log^{3/2}(eT^2)\right) T \mathfrak{R}_T(\mathcal{F}) + 2\sqrt{T}\left(1 + \log\log\left(\frac{1}{\gamma}\right)\right) \right\}.$$

To interpret the above bound, suppose that the sequence of y_t's is predicted with a margin 2γ by some function $f \in \mathcal{F}$. The upper bound guarantees that there exists a strategy (that does not need to know the value of γ) with cumulative loss given by the sequential Rademacher complexity of $\mathcal{F}$ divided by the margin, up to poly-logarithmic factors. Crucially, the bound does not directly depend on the dimensionality of the input space $\mathcal{X}$.

6.3 Decision Trees

We consider here the binary classification problem where the learner competes with a set of decision trees of depth no more than d. The function class $\mathcal{F}$ for this problem is defined as follows. Each $f \in \mathcal{F}$ is defined by choosing a rooted binary tree of depth no more than d and associating to each node a binary valued decision function from a set $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$. A binary value for a given x can be obtained by traversing the tree from the root according to the value of the decision function at each node and then reading off the label of the leaf. Importantly, x "reaches" only one leaf of the tree. Alternatively, for any leaf l, the membership of x is given by the conjunction

$$\prod_i \mathbf{1}\left\{h_{l,i}(x) = 1\right\}$$

where $h_{l,i}$ is either the decision function at node i along the path to the leaf l, or its negation. To complete the definition of f, we choose weights $w_l > 0$, $\sum_l w_l = 1$, along with the value $\sigma_l \in \{\pm 1\}$ of the function on each leaf l. The resulting function f can be written as

$$f(x) = \sum_l w_l \sigma_l \prod_i \mathbf{1}\left\{h_{l,i}(x) = 1\right\}$$

where the sum runs over all the leaves of the tree.

The following proposition is the online analogue of a result about decision tree learning that Bartlett and Mendelson (2003) proved in the i.i.d. setting.

Proposition 17 *Denote by $\mathcal{F}$ the class of decision trees of depth at most d with decision functions in $\mathcal{H}$. There exists a randomized strategy τ for the learner such that for any sequence of instances $z_1, \ldots, z_T$, with $z_t = (x_t, y_t) \in \mathcal{X} \times \{\pm 1\}$,*

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{y}_t \sim \tau_t(z_{1:t-1})}\left[\mathbf{1}\left\{\hat{y}_t \neq y_t\right\}\right] \leq \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathbf{1}\left\{f(x_t) \neq y_t\right\} + \mathcal{O}\left(\sum_l \min\left(C(l), d\log^3(T)\ T\ \mathfrak{R}(\mathcal{H})\right) + \sqrt{T}\log(N)\right),$$

where $C(l)$ denotes the number of instances that reach the leaf l and are correctly classified in the decision tree f that minimizes $\sum_{t=1}^{T} \mathbf{1}\{y_t f(x_t) \leq 0\}$, with $N > 2$ being the number of leaves in this tree.

It is not clear whether computationally feasible online methods exist for learning decision trees, and this represents an interesting avenue of further research.

6.4 Transductive Learning

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

$$\widehat{\mathcal{N}}_\infty(\alpha, \mathcal{F}) = \min\left\{|G| : G \subseteq \mathbb{R}^{\mathcal{X}} \text{ s.t. } \forall f \in \mathcal{F} \;\; \exists g \in G \;\; \text{satisfying} \;\; \|f - g\|_\infty \le \alpha\right\} \tag{13}$$

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

$$\forall T, \quad \mathcal{N}_\infty(\alpha, \mathcal{F}, T) \le \widehat{\mathcal{N}}_\infty(\alpha, \mathcal{F}) \; . \tag{14}$$

Indeed, let G be a minimal cover of $\mathcal{F}$ at scale α. We claim that for any $\mathcal{X}$-valued tree of depth T, the set $V = \{\mathbf{v}^g = g \circ \mathbf{x} : g \in G\}$ 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 $\|f - g\|_\infty \le \alpha$. Clearly $|\mathbf{v}^g_t(\epsilon) - f(\mathbf{x}_t(\epsilon))| \le \alpha$ for any $1 \le t \le T$, concluding the proof.

This simple observation can be applied in several situations. First, consider the problem of *transductive learning*, where the set $\mathcal{X} = \{x_1, \ldots, x_n\}$ 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 $\widehat{\mathcal{N}}_\infty(\alpha, \mathcal{F}) \le (c/\alpha)^d$ 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 (6) and choosing $\alpha = 0$, we observe that the value of the supervised game is upper bounded by $2\mathfrak{D}_T(\mathcal{F}) \le 48\sqrt{dT \log c}$ by Proposition 9. It is easy to see that if n is fixed and the problem is learnable in the batch (i.e., i.i.d.) setting, then the problem is learnable in the online transductive model.

In the transductive setting considered by Kakade and Kalai (2006), it is assumed that $n \le T$ and $\mathcal{F}$ consists of binary-valued functions. If $\mathcal{F}$ is a class with VC dimension d, the Sauer-Shelah lemma ensures that the ℓ_∞ cover is smaller than $(en/d)^d \le (eT/d)^d$. Using the previous argument with $c = eT$, we obtain a bound of $4\sqrt{dT \log(eT)}$ for the value of the game, matching the bound of Kakade and Kalai (2006) up to a constant factor.

6.5 Isotron

Kalai and Sastry (2009) 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 considered by the authors. In its "batch" version, we assume that the data are revealed at once as a set $\{(x_t, y_t)\}_{t=1}^T \in \mathbb{R}^d \times \mathbb{R}$ where $y_t = u(\langle w, x_t \rangle)$ for some unknown $w \in \mathbb{R}^d$ of bounded norm and an unknown non-decreasing $u : \mathbb{R} \to \mathbb{R}$ 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 by Kalai and Sastry (2009) 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 of

$$\mathcal{H} = \left\{ f(x,y) = (y - u(\langle w, x\rangle))^2 \mid u : [-1,1] \to [-1,1] \text{ 1-Lipschitz }, \ \|w\|_2 \le 1 \right\} \tag{15}$$

in the supervised setting, over $\mathcal{X} = B_2$ (the unit Euclidean ball in $\mathbb{R}^d$) and $\mathcal{Y} = [-1,1]$. 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 shows that the covering number of the class does not increase much under these compositions.

Proposition 18 *The class $\mathcal{H}$ defined in* (15) *is online learnable in the (improper) supervised learning setting. Moreover, the minimax regret is*

$$\mathcal{O}(\sqrt{T}\log^{3/2}(T)).$$

Once again, it is not clear whether a computationally efficient method attaining the above guarantee exists.

6.6 Prediction of Individual Sequences with Static Experts

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 (1999) proved upper bounds on the minimax value in terms of the (classical) Rademacher complexity and the (classical) Dudley integral. One of the assumptions made by Cesa-Bianchi and Lugosi (1999) 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 vectors $\bar{f} = (f_1, \dots, f_T) \in [0,1]^T$, and let $\mathcal{F}$ denote a class of such experts. Let $\mathcal{Y} = \{0,1\}$, putting us in the scenario of binary classification with no side information. Then regret on a particular sequence $y_1, \dots, y_T$ can be written as

$$\sum_{t=1}^{T} \ell_t(\bar{f}_t, y_t) - \inf_{\bar{f}\in\mathcal{F}} \sum_{t=1} \ell_t(\bar{f}, y_t),$$

where $\bar{f}_t$ is the expert chosen by the learning algorithm at time t. Observe that the proof of Theorem 7 does not require the loss to be time independent. In the case of absolute loss, the Rademacher complexity appearing on the right hand side in Theorem 7 becomes

$$\sup_{\mathbf{y}} \mathbb{E}_\epsilon \left[\sup_{\bar{f}\in\mathcal{F}} \sum_{t=1}^{T} \epsilon_t \ell_t(\bar{f}, \mathbf{y}_t(\epsilon)) \right] = \sup_{\mathbf{y}} \mathbb{E}_\epsilon \left[\sup_{\bar{f}\in\mathcal{F}} \sum_{t=1}^{T} \epsilon_t |f_t - \mathbf{y}_t(\epsilon)| \right] .$$

where the supremum is over all $\mathcal{Y}$-valued trees of depth T. Noting that for $f \in [0,1], y \in \{0,1\}$, $|f-y|$ can be written as $(1-2y)f + y$, the above equals

$$\sup_{\mathbf{y}} \mathbb{E}_\epsilon \left[\left(\sup_{\bar{f}\in\mathcal{F}} \sum_{t=1}^{T} \epsilon_t (1 - 2\mathbf{y}_t(\epsilon)) f_t \right) + \sum_{t=1}^{T} \epsilon_t \mathbf{y}_t(\epsilon) \right] = \sup_{\mathbf{y}} \mathbb{E}_\epsilon \left[\sup_{\bar{f}\in\mathcal{F}} \sum_{t=1}^{T} \epsilon_t (1 - 2\mathbf{y}_t(\epsilon)) f_t \right].$$

It can be easily verified that the joint distribution of $\{\epsilon_t(1-2\mathbf{y}_t(\epsilon))\}_{t=1}^T$ is still i.i.d. Rademacher and hence the value of the game is upper bounded by

$$2\mathbb{E}_\epsilon\left[\sup_{\bar{f}\in\mathcal{F}}\sum_{t=1}^T \epsilon_t f_t\right],$$

recovering the upper bound of Theorem 3 in (Cesa-Bianchi and Lugosi, 1999). We note that for this particular scenario, the factor of 2 (that appears because of symmetrization) is not needed. This factor is the price we pay for deducing the result from the general statement of Theorem 7.

7. Discussion

The tools provided in this paper allow us to establish existence of regret minimization algorithms by working directly with the minimax value. The non-constructive nature of our results is due to the application of the minimax theorem: the dual strategy does not give a handle on the primal strategy. Furthermore, by passing to upper bounds on the dual formulation (2) of the value of the game, we remove the dependence on the dual strategy altogether. After the original paper (Rakhlin et al., 2010) appeared, the algorithmic approach has been developed by Rakhlin et al. (2012) who showed that the prediction for round t can be obtained by appealing to the minimax theorem for rounds $t+1$ to T, yet keeping the minimax expression for round t as is. The notion of a relaxation (in the spirit of approximate dynamic programming) then allowed the authors to develop a general recipe for deriving computationally feasible prediction methods. The techniques of the present paper form the basis for the algorithmic developments of Rakhlin et al. (2012). We refer the reader to (Rakhlin and Sridharan, 2014; Rakhlin et al., 2012) for details.

Acknowledgments

We would like to thank J. Michael Steele and Dean Foster for helpful discussions. We gratefully acknowledge the support of NSF under grants CAREER DMS-0954737 and CCF-1116928.

Appendix A. A Minimax Theorem

The minimax theorem is one of this paper's main workhorses. For completeness, we state a general version of this theorem — the von Neumann-Fan minimax theorem — due to Borwein (2014) (see also Borwein and Zhuang, 1986).

Theorem 19 (Borwein, 2014) *Let $\mathcal{A}$ and $\mathcal{B}$ be Banach spaces. Let $A \subset \mathcal{A}$ be nonempty, weakly compact, and convex, and let $B \subset \mathcal{B}$ be nonempty and convex. Let $g : \mathcal{A}\times\mathcal{B} \to \mathbb{R}$ be concave with respect to $b \in B$ and convex and lower-semicontinuous with respect to $a \in A$, and weakly continuous in a when restricted to A. Then*

$$\sup_{b\in B}\inf_{a\in A} g(a,b) = \inf_{a\in A}\sup_{b\in B} g(a,b). \tag{16}$$

In the proof of Theorem 1, the minimax theorem is invoked to assure that

$$\inf_{q_t \in \mathcal{Q}} \sup_{p_t \in \mathcal{P}} \mathbb{E}\left[\ell(f_t, z_t) + \xi(z_t)\right] = \sup_{p_t \in \mathcal{P}} \inf_{q_t \in \mathcal{Q}} \mathbb{E}\left[\ell(f_t, z_t) + \xi(z_t)\right], \tag{17}$$

where $\xi(z_t)$ is a rather complicated function that includes the repeated infima and suprema from steps $t+1$ to T of regret expression that includes the variable z_t (but not f_t). The expectation in (17) is with respect to $f_t \sim q_t$ and $z_t \sim p_t$. To apply (16), we take g to be the bilinear form in q_t and p_t, with $A = \mathcal{Q}$ and $B = \mathcal{P}$. Equipped with the total variation distance, $\mathcal{Q}$ and $\mathcal{P}$ can be seen as subsets of a Banach space of measures on $\mathcal{F}$ and $\mathcal{Z}$, respectively. In terms of conditions, it is enough to check weak compactness of $\mathcal{Q}$ and assume continuity of the loss function (lower semi-continuity can be used as well).

Weak compactness of the set of probability measures on a complete separable metric space is equivalent to uniform tightness by the fundamental result of Prohorov (see, e.g., Bogachev 2007, Theorem 8.6.2., and van der Vaart and Wellner 1996). If $\mathcal{F}$ itself is compact, then the set $\Delta(\mathcal{F})$ of probability measures on $\mathcal{F}$ is tight, and hence (under the continuity of the loss) the minimax theorem holds. If $\mathcal{F}$ is not compact, tightness can be established under the following general condition. According to Example 8.6.5 (ii) in Bogachev (2007), a family $\Delta(\mathcal{F})$ of Borel probability measures on a separable *reflexive* Banach space E is uniformly tight (under the weak topology) precisely when there exists a function $V : E \to [0, \infty)$ continuous in the norm topology such that

$$\lim_{\|f\| \to \infty} V(f) = \infty \quad \text{and} \quad \sup_{q \in \Delta(\mathcal{F})} \mathbb{E}_{f \sim q} V(f) < \infty.$$

As an example, if $\mathcal{F}$ is a subset of a ball in E, it is enough to take $V(f) = \|f\|$.

Finally, we remark that in the supervised learning case by considering the improper learning scenario we allow x_t to be observed before the choice $\hat{y}_t$ is made. Therefore, we do not need to invoke the minimax theorem on the space of functions $\mathcal{F}$, but rather (see the proof of Theorem 8) for two real-valued decisions in a bounded interval. This makes the application of the minimax theorem straightforward.

Appendix B. Proofs

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

$$\begin{aligned}
\mathcal{V}_T(\mathcal{F}) &= \inf_{q_1} \sup_{p_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ z_1 \sim p_1}} \dots \inf_{q_T} \sup_{p_T} \mathbb{E}_{\substack{f_T \sim q_T \\ z_T \sim p_T}} \left\{ \sum_{t=1}^{T} \ell(f_t, z_t) - \psi(z_{1:T}) \right\} \\
&= \sup_{p_1} \inf_{q_1} \mathbb{E}_{\substack{f_1 \sim q_1 \\ z_1 \sim p_1}} \dots \sup_{p_T} \inf_{q_T} \mathbb{E}_{\substack{f_T \sim q_T \\ z_T \sim p_T}} \left\{ \sum_{t=1}^{T} \ell(f_t, z_t) - \psi(z_{1:T}) \right\} \\
&= \sup_{p_1} \inf_{f_1} \mathbb{E}_{z_1 \sim p_1} \dots \sup_{p_T} \inf_{f_T} \mathbb{E}_{z_T \sim p_T} \left\{ \sum_{t=1}^{T} \ell(f_t, z_t) - \psi(z_{1:T}) \right\},
\end{aligned}$$

where q_t and p_t range over $\mathcal{Q}$ and $\mathcal{P}$, the sets of distributions on $\mathcal{F}$ and $\mathcal{Z}$, respectively. From now on, it will be understood that z_t has distribution p_t. By moving the expectation

with respect to z_T and then the infimum with respect to f_T inside the expression, we arrive at

$$\sup_{p_1}\inf_{f_1}\mathbb{E}_{z_1}\ldots\sup_{p_{T-1}}\inf_{f_{T-1}}\mathbb{E}_{z_{T-1}}\sup_{p_T}\left\{\sum_{t=1}^{T-1}\ell(f_t,z_t)+\left[\inf_{f_T}\mathbb{E}_{z_T}\ell(f_T,z_T)\right]-\mathbb{E}_{z_T}\psi(z_{1:T})\right\}$$
$$=\sup_{p_1}\inf_{f_1}\mathbb{E}_{z_1}\ldots\sup_{p_{T-1}}\inf_{f_{T-1}}\mathbb{E}_{z_{T-1}}\sup_{p_T}\mathbb{E}_{z_T}\left\{\sum_{t=1}^{T-1}\ell(f_t,z_t)+\left[\inf_{f_T}\mathbb{E}_{z_T}\ell(f_T,z_T)\right]-\psi(z_{1:T})\right\}. \tag{18}$$

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

$$\sup_{p_1}\inf_{f_1}\mathbb{E}_{z_1}\ldots\sup_{p_{T-1}}\inf_{f_{T-1}}\mathbb{E}_{z_{T-1}}\left\{\sum_{t=1}^{T-1}\ell(f_t,z_t)+\sup_{p_T}\mathbb{E}_{z_T}\left[\inf_{f_T}\mathbb{E}_{z_T}\ell(f_T,z_T)-\psi(z_{1:T})\right]\right\}$$

which, in turn, is equal to

$$\sup_{p_1}\inf_{f_1}\mathbb{E}_{z_1}\ldots\sup_{p_{T-1}}\left\{\sum_{t=1}^{T-2}\ell(f_t,z_t)+\left[\inf_{f_{T-1}}\mathbb{E}_{z_{T-1}}\ell(f_{T-1},z_{T-1})\right]\right.$$
$$\left.+\mathbb{E}_{z_{T-1}}\sup_{p_T}\mathbb{E}_{z_T}\left[\inf_{f_T}\mathbb{E}_{z_T}\ell(f_T,z_T)-\psi(z_{1:T})\right]\right\}$$
$$=\sup_{p_1}\inf_{f_1}\mathbb{E}_{z_1}\ldots\sup_{p_{T-1}}\mathbb{E}_{z_{T-1}}\sup_{p_T}\mathbb{E}_{z_T}\left\{\sum_{t=1}^{T-2}\ell(f_t,z_t)+\left[\inf_{f_{T-1}}\mathbb{E}_{z_{T-1}}\ell(f_{T-1},z_{T-1})\right]\right.$$
$$\left.+\left[\inf_{f_T}\mathbb{E}_{z_T}\ell(f_T,z_T)\right]-\psi(z_{1:T})\right\}.$$

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

Proof [**of Lemma 4**] Without loss of generality assume that the Lipschitz constant $L=1$, as the general case follows by scaling ϕ. Fix a $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth T. We first claim that

$$\log\,\mathcal{N}_2(\beta,\phi\circ\mathcal{G},\mathbf{z})\le\sum_{j=1}^{k}\log\,\mathcal{N}_\infty(\beta,\mathcal{G}_j,\mathbf{z})\ .$$

Suppose $V_1,\ldots,V_k$ are minimal β-covers with respect to ℓ_∞ for $\mathcal{G}_1,\ldots,\mathcal{G}_k$ on the tree $\mathbf{z}$. Consider the set

$$V^\phi=\{\mathbf{v}^\phi\,:\,\mathbf{v}\in V_1\times\ldots\times V_k\},$$

where $\mathbf{v}^\phi$ is the tree such that $\mathbf{v}_t^\phi(\epsilon)=\phi(\mathbf{v}_t(\epsilon),\mathbf{z}_t(\epsilon))$. Then, for any $g=(g_1,\ldots,g_k)\in\mathcal{G}$ and any $\epsilon\in\{\pm1\}^T$, with representatives $(\mathbf{v}^1,\ldots,\mathbf{v}^k)\in V_1\times\ldots\times V_k$, we have,

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\phi(g(\mathbf{z}_t(\epsilon)),\mathbf{z}_t(\epsilon))-\mathbf{v}_t^\phi(\epsilon)\right)^2}\le\max_{t\in[T]}\left|\phi(g(\mathbf{z}_t(\epsilon)),\mathbf{z}_t(\epsilon))-\mathbf{v}_t^\phi(\epsilon)\right|$$
$$=\max_{t\in[T]}|\phi(g(\mathbf{z}_t(\epsilon)),\mathbf{z}_t(\epsilon))-\phi(\mathbf{v}_t(\epsilon),\mathbf{z}_t(\epsilon))|\le\max_{j\in[k]}\max_{t\in[T]}|g_j(\mathbf{z}_t(\epsilon)))-\mathbf{v}_t^j(\epsilon)|\le\beta.$$

Thus we see that V^ϕ is an β-cover with respect to ℓ_∞ for $\phi \circ \mathcal{G}$ on $\mathbf{z}$. Hence

$$\log\ \mathcal{N}_2(\beta, \phi \circ \mathcal{G}, \mathbf{z}) \leq \log(|V^\phi|) = \sum_{j=1}^{k} \log(|V_j|) = \sum_{j=1}^{k} \log\ \mathcal{N}_\infty(\beta, \mathcal{G}_j, \mathbf{z}). \tag{19}$$

For any $g \in \mathcal{G}$ and $z \in \mathcal{Z}$, the value $\phi(g(z), z)$ is contained in the interval $[-1 + \phi(\mathbf{0}, z), +1 + \phi(\mathbf{0}, z)]$ by the Lipschitz property. Consider the $\mathbb{R}$-valued tree $\phi(\mathbf{0}, \cdot) \circ \mathbf{z}$. We now center by this tree and consider the set of trees

$$\{\phi(g(\cdot), \cdot) \circ \mathbf{z} - \phi(\mathbf{0}, \cdot) \circ \mathbf{z} : g \in \mathcal{G}\}.$$

The centering does not change the size of the cover calculated in (19), but allows us to invoke (7) since the function values are now in $[-1, 1]$:

$$\begin{aligned}\mathfrak{R}_T(\phi \circ \mathcal{G}, \mathbf{z}) &\leq \inf_\alpha \left\{ 4\alpha + \frac{12}{\sqrt{T}} \int_\alpha^1 \sqrt{\sum_{j=1}^{k} \log\ \mathcal{N}_\infty(\beta, \mathcal{G}_j, \mathbf{z})}\ d\beta \right\} \\ &\leq \inf_\alpha \left\{ 4\alpha + \frac{12}{\sqrt{T}} \sum_{j=1}^{k} \int_\alpha^1 \sqrt{\log\ \mathcal{N}_\infty(\beta, \mathcal{G}_j, \mathbf{z})}\, d\beta \right\}. \end{aligned} \tag{20}$$

We substitute the upper bound on covering numbers in (8) for each $\mathcal{G}_j$ and arrive at an upper bound of

$$\inf_\alpha \left\{ 4\alpha + \frac{12}{\sqrt{T}} \sum_{j=1}^{k} \int_\alpha^1 \sqrt{\mathrm{fat}_\beta(\mathcal{G}_j) \log(2eT/\beta)} d\beta \right\}. \tag{21}$$

Lemma 2 of Rakhlin et al. (2014) implies that for any $\beta > 2\mathfrak{R}_T(\mathcal{G}_j)$,

$$\mathrm{fat}_\beta(\mathcal{G}_j) \leq \frac{32T\ \mathfrak{R}_T(\mathcal{G}_j)^2}{\beta^2}\ .$$

Let $j^* = \operatorname{argmax}_j \mathfrak{R}_T(\mathcal{G}_j)$. Substituting this together with the value of $\alpha = 2\mathfrak{R}_T(\mathcal{G}_{j^*})$ into (21) yields an upper bound

$$8\ \mathfrak{R}_T(\mathcal{G}_{j^*}) + 48\sqrt{2}\ \sum_{j=1}^{k} \mathfrak{R}_T(\mathcal{G}_j) \int_{2\mathfrak{R}_T(\mathcal{G}_{j^*})}^1 \frac{1}{\beta} \sqrt{\log(2eT/\beta)} d\beta.$$

Using the fact that for any $b > 1$ and $\alpha \in (0, 1)$

$$\int_\alpha^1 \frac{1}{\beta} \sqrt{\log(b/\beta)} d\beta = \int_b^{b/\alpha} \frac{1}{x} \sqrt{\log x} dx = \frac{2}{3} \log^{3/2}(x) \Big|_b^{b/\alpha} \leq \frac{2}{3} \log^{3/2}(b/\alpha) \tag{22}$$

we obtain a further upper bound of

$$8\ \mathfrak{R}_T(\mathcal{G}_{j^*}) + 32\sqrt{2}\ \sum_{j=1}^{k} \mathfrak{R}_T(\mathcal{G}_j)\ \log^{3/2}\left(\frac{eT}{\mathfrak{R}_T(\mathcal{G}_{j^*})}\right)\ .$$

Replacing the first term by $8\sum_j \mathfrak{R}_T(\mathcal{G}_j)$, we conclude that

$$\mathfrak{R}_T(\phi\circ\mathcal{G},\mathbf{z}) \le 8\left(1+4\sqrt{2}\log^{3/2}(eT^2)\right)\sum_{j=1}^{k}\mathfrak{R}_T(\mathcal{G}_j)$$

as long as $\mathfrak{R}_T(\mathcal{G}_j)\ge 1/T$ for each j. The statement is concluded by observing that $\mathbf{z}$ was chosen arbitrarily. ■

Proof [**of Corollary 6**] We first extend the binary function b to a function $\bar{b}$ to any $x\in\mathbb{R}^k$ as follows :

$$\bar{b}(x) = \begin{cases} (1-\|x-a\|_\infty)b(a) & \text{if } \|x-a\|_\infty < 1 \text{ for some } a\in\{\pm1\}^k \\ 0 & \text{otherwise} \end{cases}$$

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 4 we conclude the statement of the corollary. ■

Proof [**of Theorem 7**] Let $\mathbb{E}_{t-1}[\cdot]=\mathbb{E}[\cdot|Z_1,\dots,Z_{t-1}]$ denote the conditional expectation. Using Theorem 1 we have,

$$\begin{aligned}
\mathcal{V}_T(\mathcal{F}) &= \sup_{p_1}\mathop{\mathbb{E}}_{Z_1\sim p_1}\dots\sup_{p_T}\mathop{\mathbb{E}}_{Z_T\sim p_T}\left[\sum_{t=1}^{T}\inf_{f_t\in\mathcal{F}}\mathbb{E}_{t-1}\ell(f_t,\cdot)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f,Z_t)\right] \\
&= \sup_{p_1}\mathop{\mathbb{E}}_{Z_1\sim p_1}\dots\sup_{p_T}\mathop{\mathbb{E}}_{Z_T\sim p_T}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{T}\inf_{f_t\in\mathcal{F}}\mathbb{E}_{t-1}\ell(f_t,\cdot)-\sum_{t=1}^{T}\ell(f,Z_t)\right\}\right] \\
&\le \sup_{p_1}\mathop{\mathbb{E}}_{Z_1\sim p_1}\dots\sup_{p_T}\mathop{\mathbb{E}}_{Z_T\sim p_T}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{T}\mathbb{E}_{t-1}\ell(f,\cdot)-\sum_{t=1}^{T}\ell(f,Z_t)\right\}\right]. \qquad (23)
\end{aligned}$$

The upper bound is obtained by replacing each infimum by a particular choice f. This step also holds if the choice f_t of the learner comes from a larger set $\mathcal{G}$, as long as $\mathcal{F}\subseteq\mathcal{G}$. The proof is concluded by appealing to (3).

■

Proof [**of Theorem 8**]

Let $\tilde{Q}$ denote the set of distributions on $\mathcal{Y}=[-1,1]$. By convexity,

$$\sum_{t=1}^{T}\ell(\hat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t),y_t)\le\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell'(\hat{y}_t,y_t)\left(\hat{y}_t-f(x_t)\right),$$

where $\ell'(\hat{y}_t,y_t)$ is a subgradient of the function $y\mapsto\ell(\cdot,y_t)$ at $\hat{y}_t$. Then the minimax value (10) can be upper bounded as

$$\mathcal{V}_T^S(\mathcal{F})\le\sup_{x_1}\inf_{q_1\in\tilde{Q}}\sup_{y_1}\mathop{\mathbb{E}}_{\hat{y}_1\sim q_1}\dots\sup_{x_T}\inf_{q_T\in\tilde{Q}}\sup_{y_T}\mathbb{E}_{\hat{y}_T\sim q_T}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell'(\hat{y}_t,y_t)\left(\hat{y}_t-f(x_t)\right)\right].$$

By the Lipschitz property of ℓ, we can replace each subgradient $\ell'(\hat{y}_t, y_t)$ with a number $s_t \in [-L, L]$ to obtain the upper bound

$$\sup_{x_1} \inf_{q_1 \in \tilde{Q}} \sup_{y_1} \mathop{\mathbb{E}}_{\hat{y}_1 \sim q_1} \sup_{s_1 \in [-L,L]} \dots \sup_{x_T} \inf_{q_T \in \tilde{Q}} \sup_{y_T} \mathop{\mathbb{E}}_{\hat{y}_T \sim q_T} \sup_{s_T \in [-L,L]} \left\{ \sup_{f \in \mathcal{F}} \sum_{t=1}^{T} s_t \left(\hat{y}_t - f(x_t) \right) \right\}.$$

Since y_t's no longer appear in the optimization objective, we can simply write the above as

$$\begin{aligned} &\sup_{x_1} \inf_{q_1 \in \tilde{Q}} \mathop{\mathbb{E}}_{\hat{y}_1 \sim q_1} \sup_{s_1 \in [-L,L]} \dots \sup_{x_T} \inf_{q_T \in \tilde{Q}} \mathop{\mathbb{E}}_{\hat{y}_T \sim q_T} \sup_{s_T \in [-L,L]} \left\{ \sup_{f \in \mathcal{F}} \sum_{t=1}^{T} s_t \left(\hat{y}_t - f(x_t) \right) \right\} \\ &= \sup_{x_1} \inf_{\hat{y}_1 \in [-1,1]} \sup_{s_1 \in [-L,L]} \dots \sup_{x_T} \inf_{\hat{y}_T \in [-1,1]} \sup_{s_T \in [-L,L]} \left\{ \sup_{f \in \mathcal{F}} \sum_{t=1}^{T} s_t \left(\hat{y}_t - f(x_t) \right) \right\}, \end{aligned}$$

where the equality follows because infima are obtained at point distributions. By the same reasoning, we now pass to distributions over s_t's:

$$\sup_{x_1} \inf_{\hat{y}_1 \in [-1,1]} \sup_{p_1} \mathop{\mathbb{E}}_{s_1 \sim p_1} \dots \sup_{x_T} \inf_{\hat{y}_T \in [-1,1]} \sup_{p_T} \mathbb{E}_{s_T \sim p_T} \left[\sum_{t=1}^{T} s_t \cdot \hat{y}_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} s_t f(x_t) \right]. \tag{24}$$

From now on, it will be understood that the supremum over p_t ranges over all distributions supported on $[-L, L]$, for any t, and s_t has distribution p_t. Now note that

$$\mathbb{E}_{s_T} \left[\sum_{t=1}^{T} s_t \cdot \hat{y}_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} s_t \cdot f(x_t) \right]$$

is concave (linear) in p_T and is convex in $\hat{y}_T$ and hence by the minimax theorem,

$$\begin{aligned} &\inf_{\hat{y}_T \in [-1,1]} \sup_{p_T} \mathbb{E}_{s_T} \left[\sum_{t=1}^{T} s_t \cdot \hat{y}_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} s_t f(x_t) \right] = \sup_{p_T} \inf_{\hat{y}_T \in [-1,1]} \mathbb{E}_{s_T} \left[\sum_{t=1}^{T} s_t \cdot \hat{y}_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} s_t f(x_t) \right] \\ &= \sum_{t=1}^{T-1} s_t \cdot \hat{y}_t + \sup_{p_T} \mathbb{E}_{s_T} \left[\inf_{\hat{y}_T \in [-1,1]} \mathbb{E}_{s_T} [s_T] \cdot \hat{y}_T - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} s_t f(x_t) \right], \end{aligned}$$

where the last step is similar to the one in the proof of Theorem 1, specifically (18). Similarly note that the term

$$\mathbb{E}_{s_{T-1}} \left[\sum_{t=1}^{T-1} s_t \cdot \hat{y}_t + \sup_{p_T, x_T} \mathbb{E}_{s_T} \left[\inf_{\hat{y}_T \in [-1,1]} \mathbb{E}_{s_T} [s_T] \cdot \hat{y}_T - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} s_t f(x_t) \right] \right]$$

is concave (linear) in p_{T-1} and is convex in $\hat{y}_{T-1}$ and hence again by the minimax theorem,

$$\begin{aligned} &\inf_{\hat{y}_{T-1} \in [-1,1]} \sup_{p_{T-1}} \mathbb{E}_{s_{T-1}} \left[\sum_{t=1}^{T-1} s_t \cdot \hat{y}_t + \sup_{p_T, x_T} \mathbb{E}_{s_T} \left[\inf_{\hat{y}_T \in [-1,1]} \mathbb{E}_{s_T} [s_T] \cdot \hat{y}_T - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} s_t f(x_t) \right] \right] \\ &= \sup_{p_{T-1}} \inf_{\hat{y}_{T-1} \in [-1,1]} \mathbb{E}_{s_{T-1}} \left[\sum_{t=1}^{T-1} s_t \cdot \hat{y}_t + \sup_{p_T, x_T} \mathbb{E}_{s_T} \left[\inf_{\hat{y}_T \in [-1,1]} \mathbb{E}_{s_T} [s_T] \cdot \hat{y}_T - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} s_t f(x_t) \right] \right] \\ &= \sum_{t=1}^{T-2} s_t \cdot \hat{y}_t + \sup_{p_{T-1}} \mathop{\mathbb{E}}_{s_{T-1}} \sup_{p_T, x_T} \mathbb{E}_{s_T} \left[\sum_{t=T-1}^{T} \inf_{\hat{y}_t \in [-1,1]} \mathbb{E}_{s_t} [s_t] \cdot \hat{y}_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} s_t f(x_t) \right]. \end{aligned}$$

Proceeding in similar fashion and using this in (24) we conclude that,

$$\begin{aligned}
\mathcal{V}_T^S(\mathcal{F}) &\le \sup_{x_1} \inf_{\hat{y}_1 \in [-1,1]} \sup_{p_1} \mathop{\mathbb{E}}_{s_1 \sim p_1} \dots \sup_{x_T} \inf_{\hat{y}_T \in [-1,1]} \sup_{p_T} \mathbb{E}_{s_T \sim p_T} \left[\sum_{t=1}^T s_t \cdot \hat{y}_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^T s_t f(x_t) \right] \\
&= \sup_{x_1} \sup_{p_1} \mathop{\mathbb{E}}_{s_1 \sim p_1} \dots \sup_{x_T} \sup_{p_T} \mathop{\mathbb{E}}_{s_T \sim p_T} \left[\sum_{t=1}^T \inf_{\hat{y}_t \in [-1,1]} \mathbb{E}_{s_t \sim p_t}[s_t] \cdot \hat{y}_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^T s_t f(x_t) \right] \\
&\le \sup_{x_1} \sup_{p_1} \mathop{\mathbb{E}}_{s_1 \sim p_1} \dots \sup_{x_T} \sup_{p_T} \mathbb{E}_{s_T \sim p_T} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \left(\mathbb{E}_{s_t \sim p_t}[s_t] - s_t \right) f(x_t) \right],
\end{aligned}$$

where we replaced each $\hat{y}_t$ with a potentially suboptimal choice $f(x_t)$. Passing the expectation past the suprema we obtain an upper bound

$$\sup_{x_1} \sup_{p_1} \mathop{\mathbb{E}}_{s_1, s_1' \sim p_1} \dots \sup_{x_T} \sup_{p_T} \mathbb{E}_{s_T, s_T' \sim p_T} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \left(s_t' - s_t \right) f(x_t) \right] \tag{25}$$

$$= \sup_{x_1} \sup_{p_1} \mathop{\mathbb{E}}_{s_1, s_1' \sim p_1} \mathop{\mathbb{E}}_{\epsilon_1} \dots \sup_{x_T} \sup_{p_T} \mathop{\mathbb{E}}_{s_T, s_T' \sim p_T} \mathbb{E}_{\epsilon_T} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \epsilon_t \left(s_t' - s_t \right) f(x_t) \right]$$

$$\le \sup_{x_1} \sup_{s_1 \in [-2L, 2L]} \mathop{\mathbb{E}}_{\epsilon_1} \dots \sup_{x_T} \sup_{s_T \in [-2L, 2L]} \mathbb{E}_{\epsilon_T} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \epsilon_t s_t f(x_t) \right]$$

$$= \sup_{x_1} \sup_{s_1 \in \{-2L, 2L\}} \mathop{\mathbb{E}}_{\epsilon_1} \dots \sup_{x_T} \sup_{s_T \in \{-2L, 2L\}} \mathbb{E}_{\epsilon_T} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \epsilon_t s_t f(x_t) \right] \tag{26}$$

$$= 2L \sup_{x_1} \sup_{s_1 \in \{-1,1\}} \mathop{\mathbb{E}}_{\epsilon_1} \dots \sup_{x_T} \sup_{s_T \in \{-1,1\}} \mathbb{E}_{\epsilon_T} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \epsilon_t s_t f(x_t) \right], \tag{27}$$

where the last inequality is because, for every $t \in [T]$, we have convexity in s_t and so supremum is achieved at either $-2L$ or $2L$. Notice that after using convexity to go to gradients, the proof technique above basically mimics the proofs of Theorems 1 and 7 to get to a symmetrized term as we did in those theorems. Now consider any arbitrary function $\psi : \{\pm 1\} \mapsto \mathbb{R}$, we have that

$$\sup_{s \in \{\pm 1\}} \mathbb{E}_\epsilon [\psi(s \cdot \epsilon)] = \sup_{s \in \{\pm 1\}} \frac{1}{2} \left(\psi(+s) + \psi(-s) \right) = \frac{1}{2} \left(\psi(+1) + \psi(-1) \right) = \mathbb{E}_\epsilon [\psi(\epsilon)] .$$

Since in (27), for each t, s_t and ϵ_t appear together as $\epsilon_t \cdot s_t$ using the above equation repeatedly, we conclude that

$$\begin{aligned}
\mathcal{V}_T^S(\mathcal{F}) &\le 2L \sup_{x_1} \sup_{s_1 \in \{-1,1\}} \mathbb{E}_{\epsilon_1} \dots \sup_{x_T} \sup_{s_T \in \{-1,1\}} \mathbb{E}_{\epsilon_T} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \epsilon_t s_t f(x_t) \right] \\
&= 2L \sup_{x_1} \mathbb{E}_{\epsilon_1} \dots \sup_{x_T} \mathbb{E}_{\epsilon_T} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \epsilon_t f(x_t) \right].
\end{aligned} \tag{28}$$

We now claim that the above supremum can be written in terms of an $\mathcal{X}$-valued tree. Briefly, the solution for x_1 in (28) is attained (for simplicity, assume the supremum is attained) at

an optimal value x_1^*. The optimal value x_2^* can be calculated for $\epsilon_1 = 1$ and $\epsilon_1 = -1$. Arguing in this manner leads to a tree $\mathbf{x}$. We conclude

$$\mathcal{V}_T^S(\mathcal{F}) \le 2L \sup_{\mathbf{x}} \mathbb{E}_{\epsilon_{1:T}}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon))\right] = 2\,L\,T\,\mathfrak{R}_T(\mathcal{F}).$$

■

Proof [**of Proposition 9**] For the upper bound, we start by using Theorem 8 for absolute loss, which has a Lipschitz constant of 1, to bound the value of the game by sequential Rademacher complexity,

$$\frac{1}{T}\mathcal{V}_T^{\mathrm{S}}(\mathcal{F}) \le 2\,\mathfrak{R}_T(\mathcal{F})\ .$$

We combine the above inequality with (7) and (8) to obtain the upper bound.

Observe that a lower bound on the value can be obtained by choosing any particular joint distribution on sequences $(x_1, y_1), \ldots, (x_t, y_t)$ in (2):

$$\mathcal{V}_T^{\mathrm{S}}(\mathcal{F}) \ge \mathbb{E}\left[\sum_{t=1}^{T} \inf_{f_t\in\mathcal{F}} \mathbb{E}_{(x_t,y_t)}\left[|y_t - f_t(x_t)| \;\middle|\; (x,y)_{1:t-1}\right] - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} |y_t - f(x_t)|\right].$$

To this end, choose any $\mathcal{X}$-valued tree $\mathbf{x}$ of depth T. Let $y_1, \ldots, y_T$ be i.i.d. Rademacher random variables and define $x_t = \mathbf{x}(y_{1:t-1})$ deterministically (that is, the conditional distribution of x_t is a point distribution on $\mathbf{x}(y_{1:t-1})$). It is easy to see that this distribution makes the choice f_t irrelevant, yielding

$$\mathcal{V}_T^{\mathrm{S}}(\mathcal{F}) \ge \mathbb{E}\left[\sum_{t=1}^{T} 1 - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} |y_t - f(x_t)|\right] = \mathbb{E}_{y_1,\ldots,y_T} \sup_{f\in\mathcal{F}} \sum_{t=1}^{T} y_t f(x_t).$$

Since this holds for any tree $\mathbf{x}$, we obtain the desired lower bound $\mathcal{V}_T^{\mathrm{S}}(\mathcal{F}) \ge \mathfrak{R}_T(\mathcal{F})$. The final lower bound on $\mathfrak{R}_T(\mathcal{F})$ (in terms of the fat-shattering dimensions) is proved by Rakhlin et al. (2014, Lemma 2).

■

Proof [**of Theorem 10**] The equivalence of *1* and *2* follows directly from Proposition 9. First, suppose that fat_α is infinite for some $\alpha > 0$. Then, the lower bound says that $\mathcal{V}_T^{\mathrm{S}}(\mathcal{F}) \ge \alpha T/(4\sqrt{2})$ and hence $\limsup_{T\to\infty} \mathcal{V}_T^{\mathrm{S}}(\mathcal{F})/T \ge \alpha/(4\sqrt{2})$. 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, $\limsup_{T\to\infty} \mathcal{V}_T^{\mathrm{S}}(\mathcal{F})/T \le \epsilon$. Since $\epsilon > 0$ was arbitrary, this proves that $\mathcal{F}$ is online learnable in the supervised setting.

The statement that $\mathcal{V}_T^{\mathrm{S}}(\mathcal{F})$, $\mathfrak{R}_T(\mathcal{F})$, and $\mathfrak{D}_T(\mathcal{F})$ are within a multiplicative factor of $\mathcal{O}(\log^{3/2} T)$ of each other whenever the problem is online learnable follows immediately from Eq. (10) in (Rakhlin et al., 2014) and Proposition 9.

■

Proof [**of Lemma 13**] Consider the game $(\mathcal{F}, \mathcal{Z}_{\mathrm{cvx}})$ and fix a randomized strategy π of the player. Then, the expected regret of a randomized strategy π against any adversary playing $g_1, \ldots, g_T$ can be lower-bounded via Jensen's inequality as

$$\sum_{t=1}^{T} \mathbb{E}_{u_t \sim \pi_t(g_{1:t-1})}\left[g_t(u_t)\right] - \inf_{u\in\mathcal{F}} \sum_{t=1}^{T} g_t(u) \ge \sum_{t=1}^{T} g_t\left(\mathbb{E}_{u_t \sim \pi_t(g_{1:t-1})}\left[u_t\right]\right) - \inf_{u\in\mathcal{F}} \sum_{t=1}^{T} g_t(u),$$

which is simply regret of a *deterministic* strategy obtained from π by playing $\mathbb{E}_{u_t \sim \pi_t(g_{1:t-1})}\left[u_t\right]$ on round t. Thus, to any randomized strategy corresponds a deterministic one that is no worse. On the other hand, the set of randomized strategies contains the set of deterministic ones. Hence, $\mathcal{V}_T(\mathcal{F}, \mathcal{Z}_{\mathrm{cvx}}) = \mathcal{V}_T^{\mathrm{det}}(\mathcal{F}, \mathcal{Z}_{\mathrm{cvx}})$ where $\mathcal{V}_T^{\mathrm{det}}$ is defined as the minimax regret obtainable only using deterministic player strategies. Now, we appeal to Theorem 14 of Abernethy et al. (2008) that says $\mathcal{V}_T^{\mathrm{det}}(\mathcal{F}, \mathcal{Z}_{\mathrm{cvx}}) = \mathcal{V}_T^{\mathrm{det}}(\mathcal{F}, \mathcal{Z}_{\mathrm{lin}})$. Note that Abernethy et al. (2008) deal with convex sets in finite dimensional spaces only. However, their proof relies on fundamental properties of convex functions that are true in any general vector space (such as the fact that the first order Taylor expansion of a convex function globally lower bounds the convex function). Since $\mathcal{Z}_{\mathrm{lin}}$ also consists of convex (in fact, linear) functions, the above argument again gives $\mathcal{V}_T^{\mathrm{det}}(\mathcal{F}, \mathcal{Z}_{\mathrm{lin}}) = \mathcal{V}_T(\mathcal{F}, \mathcal{Z}_{\mathrm{lin}})$. This finishes the proof of the lemma. ■

Proof [**of Proposition 15**] We shall prove that for any $i \in \{2, \ldots, k\}$,

$$\mathfrak{R}_T(\mathcal{F}_i) \le 16 L B_i \left(1 + 4\sqrt{2}\log^{3/2}(eT^2)\right) \mathfrak{R}_T(\mathcal{F}_{i-1}).$$

To see this note that for any $\mathbf{x}$, $\mathfrak{R}_T(\mathcal{F}_i, \mathbf{x})$ is equal to

$$\mathbb{E}_\epsilon\left[\sup_{\substack{w^i:\|w^i\|_1\le B_i \\ \forall j\ f_j\in\mathcal{F}_{i-1}}} \sum_{t=1}^{T} \epsilon_t \left(\sum_j w_j^i \sigma\left(f_j(\mathbf{x}_t(\epsilon))\right)\right)\right] \le \mathbb{E}_\epsilon\left[\sup_{\substack{w^i:\|w^i\|_1\le B_i \\ \forall j\ f_j\in\mathcal{F}_{i-1}}} \|w^i\|_1 \max_j \left|\sum_{t=1}^{T} \epsilon_t \sigma\left(f_j(\mathbf{x}_t(\epsilon))\right)\right|\right]$$

by Hölder's inequality. Then $\mathfrak{R}_T(\mathcal{F}_i)$ is upper bounded as

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

Since $0 \in \mathcal{F}_i$ together with the assumption of $\sigma(0) = 0$, both terms are non-negative, and thus the maximum above can be upper bounded by the sum

$$\sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[B_i \sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^{T} \epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right) \right] + \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[B_i \sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^{T} -\epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right) \right] .$$

We now claim that the two terms are equal. Indeed, let $\mathbf{x}^*$ be the tree achieving the supremum in the first term (a modified analysis can be carried out if the supremum is not achieved). Then the mirror tree $\mathbf{x}$ defined via $\mathbf{x}_t(\epsilon) = \mathbf{x}_t^*(-\epsilon)$ yields the same value for the second term. Since the argument can be carried out in the reverse direction, the two terms are equal, and the upper bound of

$$2B_i \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^{T} \epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right) \right]$$

follows. In view of contraction in Corollary 5, we obtain a further upper bound of

$$16 B_i L \left(1 + 4\sqrt{2} \log^{3/2}(eT^2)\right) \mathfrak{R}_T(\mathcal{F}_{i-1}). \tag{29}$$

To finish the proof we note that for the base case of $i = 1$, $\mathfrak{R}_T(\mathcal{F}_1)$ is equal to

$$\sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{w \in \mathbb{R}^d : \|w\|_1 \le B_1} \sum_{t=1}^{T} \epsilon_t w^\top \mathbf{x}_t(\epsilon) \right]$$

which is upper bounded by

$$\sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{w \in \mathbb{R}^d : \|w\|_1 \le B_1} \|w\|_1 \left\| \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t(\epsilon) \right\|_\infty \right] \le B_1 \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\max_{i \in [d]} \left\{ \sum_{t=1}^{T} \epsilon_t \mathbf{x}_t(\epsilon)[i] \right\} \right].$$

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. By (4),

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

Using the above and (29) repeatedly we conclude the proof. ■

Proof [**of Proposition 16**] Fix a $\gamma > 0$ and use loss

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

Since this loss is $1/\gamma$-Lipschitz, we can use (11) and the Rademacher contraction Corollary 5 to show that for each $\gamma > 0$ there exists a randomized strategy τ^γ such that for any data sequence

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

where $\rho_T = 16\left(1 + 4\sqrt{2}\log^{3/2}(eT^2)\right)$ throughout the proof. Further, observe that the loss function is lower bounded by the zero-one loss $\mathbf{1}\{\hat{y}y < 0\}$ and is upper bounded by the margin zero-one loss $\mathbf{1}\{\hat{y}y < \gamma\}$. Hence,

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{y}_t \sim \tau_t^\gamma(z_{1:t-1})}\left[\mathbf{1}\{\hat{y}_t y_t < 0\}\right] \le \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \mathbf{1}\{y_t f(x_t) < \gamma\} + \gamma^{-1}\rho_T T \mathfrak{R}_T(\mathcal{F}). \tag{30}$$

The above bound holds for randomized each strategy given by τ^γ, for any given γ. Now we discretize the set of γ's as $\gamma_i = 1/2^i$ and use the output of the randomized strategies $\tau^{\gamma_1}, \tau^{\gamma_2}, \ldots$, that attain the regret bounds given in (30), as experts. We then run a countable experts algorithm (Algorithm 1) with initial weight for expert i as $p_i = \frac{6}{\pi^2 i^2}$. Such an algorithm achieves $\mathcal{O}(\sqrt{T}\log(1/p_i))$ regret w.r.t. expert i. In view of Proposition 20, for this randomized strategy τ, for any i,

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

For any $\gamma > 0$, let $i_\gamma \in 0, 1, \ldots,$ be such that $2^{-(i_\gamma+1)} < \gamma \le 2^{-i_\gamma}$. Then above right-hand side is upper bounded by

$$\inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \mathbf{1}\{y_t f(x_t) < 2\gamma\} + \gamma^{-1}\rho_T T \mathfrak{R}_T(\mathcal{F}) + \sqrt{T}\left(1 + 2\log\left(\frac{i_\gamma\pi}{\sqrt{6}}\right)\right).$$

The proof is concluded using the inequality $i_\gamma \le \log(1/\gamma)$ and upper bounding constants. ■

Proof [**of Proposition 17**] Fix some $L > 0$. 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 7 and Corollary 5 we have that for every $L > 0$, there exists a randomized strategy τ^L for the player, such that for any sequence $z_1 = (x_1, y_1), \ldots, z_T = (x_T, y_T)$,

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{y}_t \sim \tau_t^L(z_{1:t-1})}\left[\phi_L(y_t\hat{y}_t)\right] \le \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \phi_L(y_t f(x_t)) + L\rho_T T \mathfrak{R}_T(\mathcal{F}), \tag{31}$$

where $\rho_T = 16\left(1 + 4\sqrt{2}\log^{3/2}(eT^2)\right)$ throughout this proof. Since ϕ_L dominates the step function, the left hand side of (31) also upper-bounds the expected indicator loss

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{y}_t \sim \tau_t^L(z_{1:t-1})}\left[\mathbf{1}\{\hat{y}_t \ne y_t\}\right].$$

For any $f \in \mathcal{F}$, we can relate the ϕ_L-loss to the indicator loss by

$$\sum_{t=1}^{T} \phi_L(y_t f(x_t)) = \sum_{t=1}^{T} \mathbf{1}\{y_t f(x_t) \le 0\} + \sum_{l} C(l)\phi_L(w_l).$$

Let us now use the above decomposition in (31). Crucially, the sign of $f(x)$ does not depend on w_l, but only on the label σ_l of the unique leaf l reached by x. Thus, the infimum in (31) can be split into two infima:

$$\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\phi_L(y_t f(x_t)) = \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\mathbf{1}\left\{y_t f(x_t)\le 0\right\} + \inf_{w_l}\sum_{l} C(l)\phi_L(w_l),$$

where it is understood that the $C(l)$ term on the right hand side is computed using the function f minimizing the first sum on the right hand side. We can further write

$$\sum_l C(l)\phi_L(w_l) \le \sum_l C(l)\max(0, 1 - Lw_l) = \sum_l \max\left(0, (1-Lw_l)C(l)\right).$$

So far, we have derived a regret bound for a given L. Let us now remove the requirement to know L a priori by running the experts Algorithm 1 with $\tau^1, \tau^2, \ldots$ as a countable set of experts corresponding to the values $L \in \mathbb{N}$. The prior on expert L is taken to be $p_L = \frac{6}{\pi^2}L^{-2}$ so that $\sum p_L = 1$. For the randomized strategy τ obtained in this manner, from Proposition 20, for any sequence of instances and any $L \in \mathbb{N}$,

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_{\hat{y}_t\sim\tau_t(z_{1:t-1})}\left[\mathbf{1}\left\{\hat{y}\neq y_t\right\}\right] \le \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\mathbf{1}\left\{y_t f(x_t)\le 0\right\} + \inf_{f\in\mathcal{F}}\sum_l \max\left(0,(1-Lw_l)C(l)\right)\\ + L\rho_T T\mathfrak{R}_T(\mathcal{F}) + \sqrt{T} + 2\sqrt{T}\log(L\pi/\sqrt{6}).\end{aligned}$$

Now we pick $L = |\{l : C(l) > \rho_T T\mathfrak{R}_T(\mathcal{F})\}| \le N$ and upper bound the second infimum by choosing $w_l = 0$ if $C(l) \le \rho_T T\mathfrak{R}_T(\mathcal{F})$ and $w_l = 1/L$ otherwise:

$$\begin{aligned}\inf_{w_l}\sum_l \max\left(0,(1-Lw_l)C(l)\right) + L\rho_T T\mathfrak{R}_T(\mathcal{F}) \le \sum_l C(l)\mathbf{1}\left\{C(l)\le\rho_T T\mathfrak{R}_T(\mathcal{F})\right\}\\ + \rho_T T\mathfrak{R}_T(\mathcal{F})\sum_l \mathbf{1}\left\{C(l) > \rho_T T\mathfrak{R}_T(\mathcal{F})\right\}\end{aligned}$$

which can be written succinctly as

$$\sum_l \min\{C(l), \rho_T T\mathfrak{R}_T(\mathcal{F})\}.$$

We conclude that

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_{\hat{y}_t\sim\tau_t(z_{1:t-1})}\left[\mathbf{1}\left\{\hat{y}_t\neq y_t\right\}\right] \le \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\mathbf{1}\left\{y_t f(x_t)\le 0\right\}\\ + \sum_l \min(C(l), \rho_T T\mathfrak{R}_T(\mathcal{F})) + \sqrt{T}\left(1 + 2\log(N\pi/\sqrt{6})\right).\end{aligned}$$

Finally, we apply Corollary 6 and Lemma 3(2) to bound $\mathfrak{R}_T(\mathcal{F}) \le d\mathcal{O}(\log^{3/2} T)\ \mathfrak{R}_T(\mathcal{H})$ and thus conclude the proof. ■

Proof [**of Proposition 18**] First, by the classical result of Kolmogorov and Tikhomirov (1959), the class $\mathcal{G}$ of all bounded Lipschitz functions on a bounded interval 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 $2/\alpha$. Considering all 1-Lipschitz functions increases this to c_0/α for some universal constant c_0.

Next, consider the class $\mathcal{F} = \{\langle w, x\rangle \mid \|w\|_2 \le 1\}$ over the Euclidean ball. By Proposition 14, $\mathfrak{R}_T(\mathcal{F}) \le 1/\sqrt{T}$. Using the lower bound of Proposition 9, $\mathrm{fat}_\alpha \le 32/\alpha^2$ whenever $\alpha > 4\sqrt{2}/\sqrt{T}$. This implies that $\mathcal{N}_\infty(\alpha, \mathcal{F}, T) \le (2eT/\alpha)^{32/\alpha^2}$ whenever $\alpha > 4\sqrt{2}/\sqrt{T}$. 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 $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ be an ℓ_∞ cover of $\mathcal{F}$ on $\mathbf{x}$ at scale α. Analogously, let $W = \{g_1, \ldots, g_M\}$ 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 $\max_{t \in [T]} |f(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon)| \le \alpha$, and let $g' \in W$ be such that $\|g - g'\|_\infty \le \alpha$. Then, using the fact that functions in $\mathcal{G}$ are 1-Lipschitz, for any $t \in [T]$,

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

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

Finally, we put all the pieces together. By Theorem 8, the minimax value is bounded by $8T$ times the sequential Rademacher complexity of the class $\mathcal{G} \circ \mathcal{F} = \{u(\langle w, x\rangle) \mid u : [-1,1] \to [-1,1] \text{ is 1-Lipschitz }, \|w\|_2 \le 1\}$ since the squared loss is 4-Lipschitz on the space of possible values. The latter complexity is then bounded by

$$\begin{aligned} T\mathfrak{D}_T(\mathcal{G} \circ \mathcal{F}) &\le 32\sqrt{T} + 12 \int_{8/\sqrt{T}}^1 \sqrt{T \ \log \ \mathcal{N}(\delta, \mathcal{G} \circ \mathcal{F}, T)} \, d\delta \\ &\le 32\sqrt{T} + 12\sqrt{T} \int_{8/\sqrt{T}}^1 \sqrt{\frac{4c_0}{\delta} + \frac{128}{\delta^2} \log(2eT)} d\delta \ . \end{aligned}$$

We therefore conclude that the value of the game for the supervised learning problem is bounded by $\mathcal{O}(\sqrt{T} \log^{3/2}(T))$. ■

Appendix C. Exponentially Weighted Average (EWA) Algorithm on Countable Experts

We consider here a version of the exponentially weighted experts algorithm for a 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., Cesa-Bianchi and Lugosi, 2006, Theorem 2.2). This result is well known and we include it here for completeness, as it is needed in the proofs of Proposition 16 and Proposition 17.

Suppose 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 we also assume that $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$. 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 $\sum_i p_i = 1$.

Algorithm 1 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

Proposition 20 *The exponentially weighted average forecaster (Algorithm 1) with* $\eta = T^{-1/2}$ *enjoys the regret bound*

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

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

References

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.

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

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

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

D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956a.

D. Blackwell. Controlled random walks. In *Proceedings of the International Congress of Mathematicians, 1954*, volume 3, pages 336–338. North Holland, 1956b.

V.I. Bogachev. *Measure Theory*, volume 2. Springer, 2007. ISBN 3540345132.

J.M. Borwein. A very complicated proof of the minimax theorem. *Minimax Theory and Its Applications*, 1(1), 2014.

J.M. Borwein and D Zhuang. On Fan's minimax theorem. *Mathematical programming*, 34 (2):232–234, 1986.

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

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

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

T. Cover. Behavior of sequential predictors of binary sequences. In *Transactions of the Fourth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, 1965*, pages 263–272. Publishing House of the Czechoslovak Academy of Sciences, 1967.

T. M. Cover and A. Shenhar. Compound Bayes predictors for sequences with apparent Markov structure. *IEEE Transactions on Systems, Man and Cybernetics*, 7(6):421–424, 1977.

L. Davisson. Universal noiseless coding. *Information Theory, IEEE Transactions on*, 19 (6):783–795, 1973.

M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *Information Theory, IEEE Transactions on*, 38(4):1258–1270, 1992.

D. P. Foster and R. V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1):40–55, 1997.

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

S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

S. M. Kakade and A. T. Kalai. From batch to transductive online learning. In Y. Weiss, B. Schölkopf, and J.C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 611–618. MIT Press, 2006.

A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

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

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.

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

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

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

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

A. Rakhlin and K. Sridharan. Statistical learning and sequential prediction, 2014. Available at `http://stat.wharton.upenn.edu/~rakhlin/courses/stat928/stat928_notes.pdf`.

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

A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Beyond regret. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *JMLR Workshop and Conference Proceedings*, pages 559–594, 2011.

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

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

J. Rissanen. Universal coding, information, prediction, and estimation. *Information Theory, IEEE Transactions on*, 30(4):629–636, 1984.

H. Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 131–149. University of California Press, 1950.

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.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

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

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York, 1996.

V. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 928–936, 2003.

J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *Information Theory, IEEE Transactions on*, 23(3):337–343, 1977.