Chapter 15
On Martingale Extensions of Vapnik–Chervonenkis Theory with Applications to Online Learning

Alexander Rakhlin and Karthik Sridharan

Abstract We review recent advances on uniform martingale laws of large numbers and the associated sequential complexity measures. These results may be considered as forming a non-i.i.d. generalization of Vapnik–Chervonenkis theory. We discuss applications to online learning, provide a recipe for designing online learning algorithms, and illustrate the techniques on the problem of online node classification. We outline connections to statistical learning theory and discuss inductive principles of stochastic approximation and empirical risk minimization.

15.1 Introduction

Questions of uniform convergence of means to their expectations have been central to the development of statistical learning theory. In their seminal paper, Vapnik and Chervonenkis found necessary and sufficient conditions for such a uniform convergence for classes of binary-valued functions [40]. A decade later, this pioneering work was extended by Vapnik and Chervonenkis to classes of uniformly bounded real-valued functions [37]. These results now form an integral part of empirical process theory [15, 26, 16, 35, 14], with wide-ranging applications in statistical estimation and machine learning [9, 17].

In this review chapter, we summarize some recent advances on uniform *martingale* laws of large numbers [31], as well as their impact on both

Alexander Rakhlin
University of Pennsylvania
e-mail: rakhlin@wharton.upenn.edu

Karthik Sridharan
University of Pennsylvania
e-mail: skarthik@wharton.upenn.edu

theoretical and algorithmic understanding of online learning [29, 27]. The uniform martingale laws can be seen as natural extensions of the Vapnik–Chervonenkis theory beyond the i.i.d. scenario. In this chapter, we would like to highlight the striking similarity between the classical statements and the non-i.i.d. generalizations.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $Z_1, \ldots, Z_n, \ldots$ be a sequence of random variables taking values in a measurable space $(\mathcal{Z}, \mathcal{S})$. Suppose Z_t is $\mathcal{A}_t$-measurable, for all $t \geq 1$, where $(\mathcal{A}_t)_{t\geq 1}$ is a filtration. Let $\mathcal{F}$ be a class of measurable functions on $(\mathcal{Z}, \mathcal{S})$, with $|f| \leq 1$ for any $f \in \mathcal{F}$. Then $\{\mathbb{E}\left[f(Z_t)|\mathcal{A}_{t-1}\right] - f(Z_t) : t \geq 1\}$ is a martingale difference sequence for any $f \in \mathcal{F}$, and

$$M_n^f = \sum_{t=1}^{n} \mathbb{E}\left[f(Z_t)|\mathcal{A}_{t-1}\right] - f(Z_t)$$

is a martingale. For a fixed n, the collection $\{M_n^f : f \in \mathcal{F}\}$ defines a stochastic process over $\mathcal{F}$. It is then natural to ask whether the supremum of the process,[1]

$$\sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\left[f(Z_t)|\mathcal{A}_{t-1}\right] - f(Z_t), \tag{15.1}$$

converges to zero (as n tends to infinity) in probability or almost surely.[2] When (Z_t) is a sequence of i.i.d. random variables, the question reduces to precisely the one raised by Vapnik and Chervonenkis [40, 37], and the stochastic process over $\mathcal{F}$ becomes the well-studied *empirical process* (normalized by n).

For simplicity, we will focus on the expected value of the supremum in (15.1); convergence with probability one is shown along the same lines.

Example 15.1. Let $\mathcal{Z}$ be a unit ball in a Banach space $(\mathcal{B}, \|\cdot\|)$. Let $\mathcal{F}$ be a unit ball in a dual Banach space. By definition, the supremum in (15.1) can be written as $\frac{1}{n}\left\|\sum_{t=1}^{n} \mathbb{E}\left[Z_t|\mathcal{A}_{t-1}\right] - Z_t\right\|$ and can be interpreted as the (normalized) length of a random walk with bounded conditionally-zero-mean increments. The question of whether this quantity converges to zero is well-studied. It was shown in [25] that convergence occurs if and only if the Banach space $(\mathcal{B}, \|\cdot\|)$ is super-reflexive. Furthermore, in most "natural" Banach spaces, the expected length of such a walk is within a constant multiple of the expected length with bounded *i.i.d.* increments. In such situations, the uniform martingale convergence is equivalent to the i.i.d. case.

The example gives us hope that the martingale extension of uniform laws will be a simple enterprise. However, the situation is not so straightforward, as the next example shows.

Example 15.2. Let $\mathcal{F}$ be the class of indicators on the unit interval $\mathcal{Z} = [0, 1]$:

[1] We may also consider the absolute value of the average without any complications.

[2] Issues of measurability can be addressed with the techniques in [16].

$$\mathcal{F} = \{f_\theta(z) = \mathbf{I}\{z > \theta\} : \theta \in [0,1]\}\,. \tag{15.2}$$

Let $\epsilon_1, \dots, \epsilon_n, \dots$ be independent Rademacher random variables, $P(\epsilon_t = -1) = P(\epsilon_t = +1) = \frac{1}{2}$. Define a sequence of random variables (Z_t) as follows:

$$Z_t = \sum_{s=1}^{t} 2^{-s}\mathbf{I}\{\epsilon_s = 1\}$$

and notice that $Z_t \in [0,1)$. For the above definition, we have a dyadic filtration $\mathcal{A}_t = \sigma(\epsilon_1, \dots, \epsilon_t)$. Fix any $n \geq 1$ and observe that the sequence (Z_t) has the following two properties. For any $s \geq 1$, (a) conditionally on the event $\{\epsilon_s = -1\}$, $Z_t < Z_s + 2^{-s}$ for all $t \geq s$; (b) conditionally on the event $\{\epsilon_s = 1\}$, $Z_t \geq Z_s$ for all $t \geq s$. It then follows that for any sequence $\epsilon_1, \dots, \epsilon_n$ there exists $\theta^* \in [0,1]$ such that (15.1) is equal to the proportion of -1's in this sequence. Hence, the expected supremum is equal to 1/2. In summary, for the class $\mathcal{F}$ defined in (15.2), there exists a dyadic martingale ensuring that the expectation of the supremum in (15.1) is a constant.

The above example seems to significantly complicate the rosy picture painted by Example 15.1: the class of thresholds on a unit interval—the original question studied by Glivenko and Cantelli for i.i.d. data, and a flagship class with VC dimension one—does not satisfy the martingale analogue of the uniform law of large numbers.

The natural next question is whether there is a measure of capacity of $\mathcal{F}$ that characterizes whether the uniform martingale law of large numbers holds.

15.2 Random Averages, Combinatorial Dimensions, and Covering Numbers

A key argument in obtaining the necessary and sufficient conditions in the work of Vapnik and Chervonenkis [40, 37] relates the difference between an average and an expectation to the difference between averages on two independent samples. This step—now commonly termed *symmetrization*—allows one to reason conditionally on the data, and to study the geometry (and combinatorics) of the finite-dimensional projection

$$\mathcal{F}|_{z_1,\dots,z_n} \triangleq \{(f(z_1), \dots, f(z_n)) : f \in \mathcal{F}\}.$$

In the non-i.i.d. case, the symmetrization step is more involved due to the dependencies. Before stating the "sequential symmetrization" result, let us define the notion of a *tree*, the entity replacing "a tuple of n points" in the i.i.d. realm.

All trees considered in this chapter are complete, binary, and rooted. A tree $\mathbf{z}$ with nodes labeled by elements of $\mathcal{Z}$ will be called *a $\mathcal{Z}$-valued tree*. Equivalently, a tree $\mathbf{z}$ of depth n is represented by n labeling functions $\mathbf{z}_t : \{\pm 1\}^{t-1} \to \mathcal{Z}$, with $\mathbf{z}_1$ being a constant, and the value $\mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1})$ indicating the label in $\mathcal{Z}$ obtained by following the path $(\epsilon_1, \ldots, \epsilon_{t-1}) \in \{\pm 1\}^{t-1}$ from the root to the node (we designate -1 as "left" and $+1$ as "right"). We write $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$. Henceforth, we will denote $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ as the full path down the tree $\mathbf{z}$, and for brevity we shall write $\mathbf{z}_t(\epsilon)$ instead of $\mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1})$. In a similar manner, we may define an $\mathbb{R}$-valued tree as a tree labeled by real numbers. For instance, $f \circ \mathbf{z} = (f \circ \mathbf{z}_1, \ldots, f \circ \mathbf{z}_n)$ is a real-valued tree for any $f : \mathcal{Z} \to \mathbb{R}$. If $\epsilon_1, \ldots, \epsilon_n$ are taken to be independent Rademacher random variables, a tree $\{\mathbf{z}_t\}$ is simply a *predictable process* with respect to the corresponding dyadic filtration.

Throughout the chapter, we will refer to a particular type of tree $\mathbf{z}$ where each labeling function is a constant within a level of the tree: there exist $z_1, \ldots, z_n \in \mathcal{Z}$ such that $\mathbf{z}_t(\epsilon) = z_t$ for all t. This tree will often witness the reduction from a sequential version of the question to the i.i.d. version involving a tuple of points. Let us term such a tree a *constant-level tree.*

With the introduced notation, we are ready to state the sequential symmetrization result. It is proved in [29, 31] that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\left[f(Z_t) | \mathcal{A}_{t-1}\right] - f(Z_t) \leq 2 \sup_{\mathbf{z}} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon)) \tag{15.3}$$

where the supremum is over all $\mathcal{Z}$-valued trees of depth n. The statement also holds for the absolute value of the average on both sides. The relation is, in fact, tight in the sense that there exists a sequence (Z_t) of random variables such that the term on the left-hand side of (15.3) is lower bounded by a multiple of the term on the right-hand side (modulo additional $\mathcal{O}(n^{-1/2})$ terms).

Given a tree $\mathbf{z}$, the expected supremum on the right-hand side of (15.3),

$$\mathcal{R}_n^{seq}(\mathcal{F}; \mathbf{z}) \;\triangleq\; \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon)), \tag{15.4}$$

is termed the *sequential Rademacher complexity* of $\mathcal{F}$ on $\mathbf{z}$. The key relation (15.3) allows us to study this complexity conditionally on $\mathbf{z}$, similarly to the way classical Rademacher averages can be studied conditionally on the tuple $(z_1, \ldots, z_n) \in \mathcal{Z}^n$.

We observe that the classical notion $\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(z_t)$ of a Rademacher average is recovered as a special case of sequential Rademacher complexity by taking a constant-level tree defined earlier: $\mathbf{z}_t(\epsilon) = z_t$ for all t. The tree structure becomes irrelevant, and we gracefully recover the complexity that arises in the study of i.i.d. data.

In the i.i.d. analysis, the supremum of the symmetrized processes can be written as

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(z_t) = \sup_{a \in \mathcal{F}|_{z_1,\dots,z_n}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t a_t. \tag{15.5}$$

For a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{Z}}$, the cardinality of the projection $\mathcal{F}|_{z_1,\dots,z_n}$ is finite and governed by the beautiful combinatorics discovered by Vapnik and Chervonenkis (and later independently by Sauer, Shelah). For the case of *sequential* Rademacher complexity, however, the size of the projection $\mathcal{F}|_{\mathbf{z}} = \{f \circ \mathbf{z} : f \in \mathcal{F}\}$ can be exponential in n for any interesting class $\mathcal{F}$. This can be seen by considering a tree $\mathbf{z}$ such that 2^n distinct functions in $\mathcal{F}$ take a value 1 on one of the 2^n leaves of the tree and zero everywhere else. The following crucial observation was made in [29, 31]: while the projection $\mathcal{F}|_{\mathbf{z}}$ may indeed be too large, it is enough to consider a potentially smaller set V of $\mathbb{R}$-valued trees of depth n with the property

$$\forall f \in \mathcal{F}, \forall \epsilon \in \{\pm 1\}^n, \exists \mathbf{v} \in V \quad \text{s.t.} \quad \forall t \in \{1,\dots,n\} \quad f(\mathbf{z}_t(\epsilon)) = \mathbf{v}_t(\epsilon). \tag{15.6}$$

In other words, a single $\mathbf{v}$ can match values of different f's on different paths. While the set V is potentially smaller, we still have, as in (15.5), for any $\epsilon \in \{\pm 1\}^n$,

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon)) = \max_{\mathbf{v} \in V} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon)$$

whenever V is finite (such as the case for a class of binary-valued functions). The set V with property (15.6) is termed a 0-*cover*, and its size, denoted by $\mathcal{N}(0, \mathcal{F}, \mathbf{z})$, fills the shoes of the growth function of Vapnik and Chervonenkis [40].

To check that all the pieces of our puzzle still fit correctly, we observe that for a constant-level tree $\mathbf{z}$, the set V satisfying (15.6) indeed reduces to the notion of projection $\mathcal{F}|_{z_1,\dots,z_n}$.

The natural next question is whether the 0-cover lends itself to the combinatorics of the flavor enjoyed by the growth function. Surprisingly, the answer is yes, and the relevant combinatorial dimension was introduced 25 years ago by Littlestone [22] within the context of online learning, and without any regard to the question of uniform martingale laws.

Definition 15.1 ([22]). A $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth d is shattered by a class $\mathcal{F} \in \{\pm 1\}^{\mathcal{Z}}$ of binary-valued functions if

$$\forall \epsilon \in \{\pm 1\}^d, \exists f \in \mathcal{F}, \quad \text{s.t.} \quad \forall t \in \{1,\dots,d\} \quad f(\mathbf{z}_t(\epsilon)) = \epsilon_t.$$

Following [10], the size of the largest $\mathcal{Z}$-valued tree shattered by $\mathcal{F}$ will be called the *Littlestone dimension* and denoted by $\operatorname{ldim}(\mathcal{F})$.

Once again, for a constant-level tree, the notion of shattering coincides with the definition of Vapnik and Chervonenkis [40]. In particular, it is clear that

$$\mathrm{vc}\,(\mathcal{F}) \le \mathrm{ldim}\,(\mathcal{F})\,.$$

The following analogue of the celebrated Vapnik–Chervonenkis lemma is proved in [29, 31] for a class of binary-valued functions:

$$\mathcal{N}(0, \mathcal{F}, \mathbf{z}) \le \sum_{i=0}^{d} \binom{n}{i}$$

where $d = \mathrm{ldim}\,(\mathcal{F})$ and $\mathbf{z}$ is any $\mathcal{Z}$-valued tree of depth n.

When $\mathrm{ldim}\,(\mathcal{F})$ is infinite (as is the case with the class of thresholds discussed in Example 15.2), it is possible to show that there exists a sequence of trees of increasing size such that $\mathcal{R}_n^{seq}(\mathcal{F}; \mathbf{z})$ does not converge to zero with increasing n. Similarly, the uniform martingale deviations in (15.1) do not converge to zero for an appropriately chosen sequence of distributions. Thus, finiteness of the Littlestone dimension is *necessary and sufficient* for the uniform martingale law of large numbers to hold universally for all distributions. In other words, this dimension plays the role of the Vapnik–Chervonenkis dimension for the non-i.i.d. extension studied here.

We now review developments for classes of real-valued functions. Here, the property (15.6) is extended to the notion of a sequential cover as follows (see [29]). A set V of $\mathbb{R}$-valued trees forms a *sequential α-cover* (w.r.t. ℓ_2) of $\mathcal{F}$ on a given $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n if

$$\forall f \in \mathcal{F}, \forall \epsilon \in \{\pm 1\}^n, \exists \mathbf{v} \in V \quad \text{s.t.} \quad \frac{1}{n}\sum_{t=1}^{n} (f(\mathbf{z}_t(\epsilon)) - \mathbf{v}_t(\epsilon))^2 \le \alpha^2.$$

The size of the smallest α-cover is denoted by $\mathcal{N}_2(\alpha, \mathcal{F}, \mathbf{z})$, and the above definition naturally extends to the ℓ_p case, $p \in [1, \infty]$. As with the 0-cover, the order of quantifiers in the above definition is crucial: an element $\mathbf{v} \in V$ can be chosen given the path ϵ.

Let us define a scale-sensitive version of Littlestone dimension as follows. We say that a $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth d is α-shattered by $\mathcal{F}$ if there exists an $\mathbb{R}$-valued witness tree $\mathbf{s}$ such that

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

The size of the largest α-shattered tree is called the *sequential fat-shattering dimension* and denoted by $\mathrm{fat}_\alpha(\mathcal{F})$.

One can see that the definition of sequential cover readily reduces to the classical notion of an α-net of $\mathcal{F}|_{z_1,\ldots,z_n}$ when $\mathbf{z}$ is a constant-level tree. By the same token, the definition of α-shattering reduces to the corresponding notion in the i.i.d. case [19, 8, 6]. The following estimate for the sequential covering number is proved in [29] for a class of uniformly bounded functions $\mathcal{F} \subset [-1,1]^{\mathcal{Z}}$:

$$\mathcal{N}_\infty(\alpha, \mathcal{F}, \mathbf{z}) \leq \left(\frac{2en}{\alpha}\right)^{\mathrm{fat}_\alpha(\mathcal{F})} .$$

The corresponding result for the classical case involves extra logarithmic factors that appear to be difficult to remove [33]. As for obtaining an n-independent upper bound on ℓ_2 sequential covering numbers (an analogue to [24]) — the question is still open.

We close this section with the main result of [31]: the almost sure convergence of uniform martingale deviations in (15.1) to zero for all distributions is equivalent to both finiteness of $\mathrm{fat}_\alpha(\mathcal{F})$ for all $\alpha > 0$ and to convergence of $\sup_{\mathbf{z}} \mathcal{R}_n^{seq}(\mathcal{F}; \mathbf{z})$ to zero. The characterization in terms of the scale-sensitive dimension is analogous to the celebrated result of Alon et al [6] in the i.i.d. case. We refer to [31] for the details of these statements, as well as for more tools, such as the extension of the Dudley chaining technique to sequential covering numbers.

15.3 Online Learning: Theory

The study of uniform laws of large numbers by Vapnik and Chervonenkis was motivated by interest in the theoretical analysis of "learning machines" and the inductive principle of empirical risk minimization. In a similar vein, the study of uniform martingale analogues is motivated by questions of sequential prediction (or, online learning).

The hallmark of statistical learning theory is that it provides distribution-free learning guarantees. The prior knowledge about the problem at hand is placed not on the data-generating mechanism, but rather implicitly encapsulated in the benchmark against which we compare the performance. The objective takes the form

$$\mathbb{E}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}(X), Y) - \inf_{f \in \mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f(X), Y) \tag{15.7}$$

where $\widehat{\boldsymbol{y}}$ is a hypothesis $\mathcal{X} \to \mathcal{Y}$ produced by the learner upon observing i.i.d. data $\{(X_i, Y_i)\}_{i=1}^n$, $\mathcal{F}$ is some class of functions $\mathcal{X} \to \mathcal{Y}$ that captures the inductive bias of the practitioner, and $\boldsymbol{\ell}$ is a loss function that measures the quality of $\widehat{\boldsymbol{y}}$.

The online learning scenario goes a step further: the i.i.d. assumption is removed and the learning process is assumed to be sequential [13]. In fact, we assume nothing about the evolution of the sequence. Such a scenario is also known by the name of *individual sequence prediction.* Since the only available sequence is the one which we attempt to predict, the measure of performance is based purely on this very sequence.

While the statistical learning paradigm has proved to be successful in many applications (such as face detection, character recognition, etc.), some modern problems are inherently sequential, and the i.i.d. assumption on data — dubious at best. One such problem is described in Sect. 15.4.3. Thankfully, the martingale extensions described earlier allow us to analyze online problems of this flavor.

Let us describe the online learning scenario within the supervised setting (that is, data are pairs of predictor-response variables). In round t, the forecaster observes $x_t \in \mathcal{X}$, chooses a prediction $\widehat{\boldsymbol{y}}_t \in \mathcal{Y}$, and then observes the outcome $y_t \in \mathcal{Y}$. The quality of the prediction is evaluated by the loss function $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t)$, an "out-of-sample" performance measure. The new data point (x_t, y_t) is then incorporated into the growing dataset. In contrast to the statistical learning scenario, we make no assumptions[3] about the evolution of the sequence $(x_1, y_1), \ldots, (x_n, y_n), \ldots$. The problem becomes well-posed by considering a goal that is called *regret*:

$$\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t) - \inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \tag{15.8}$$

for some class $\mathcal{F}$ of functions $\mathcal{X} \to \mathcal{Y}$. The term subtracted is a benchmark that encodes the inductive bias in a way similar to (15.7). The fact that the loss $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t)$ is an out-of-sample performance measure facilitates a deeper connection between (15.8) and (15.7).

We remark that if the loss function is not convex in $\widehat{\boldsymbol{y}}_t$ (or if $\mathcal{Y}$ is not convex), then the forecaster commits to a randomized strategy q_t and draws $\widehat{\boldsymbol{y}}_t \sim q_t$ after observing y_t.

Littlestone [22] studied the online learning problem under the so-called realizability assumption: there exists $f^* \in \mathcal{F}$ such that the presented sequence satisfies $y_t = f^*(x_t)$ for all t. For the indicator loss and a binary sequence of outcomes, Littlestone presented a method (essentially, a variant of "halving") that makes at most $\mathrm{ldim}\,(\mathcal{F})$ mistakes; moreover, Littlestone showed that there exists a strategy of Nature ensuring that at least $\mathrm{ldim}\,(\mathcal{F})$ mistakes are incurred by any prediction method. This result has been extended by Ben-David, Pál, and Shalev–Shwartz [10] to the "agnostic" setting that lifts the realizability assumption on the sequence. For the case of indicator loss and binary outcomes, the authors exhibited a method that guarantees an $\mathcal{O}\left(n^{-1/2}\sqrt{\mathrm{ldim}\,(\mathcal{F}) \log n}\right)$ upper bound on regret and also provided a nearly matching lower bound of $\Omega\left(n^{-1/2}\sqrt{\mathrm{ldim}\,(\mathcal{F})}\right)$. The upper bounds were derived—like the vast majority of results in online learning—by exhibiting an algorithm (in this case, a clever modification of the Exponential Weights algorithm) and proving a bound on its regret. The work of [22]

[3] It is also possible to study an intermediate setting, where some knowledge about the sequence is available (see, e.g., [30]).

and [10] were the first indications that one may find characterizations of learnability for the sequential prediction setting.

In contrast to the algorithmic approach, an emerging body of literature aimed to study online learning by working directly with the minimax value of a multistage prediction problem [1, 34, 29, 2]. Since a prediction method is required to do well on all sequences, it is instructive to think of the online learning problem as a game between the Learner and Nature. The minimax regret (or, the value of the game) is then defined as

$$\mathcal{V}_n(\mathcal{F}) \triangleq \left\langle\!\!\left\langle \sup_{x_t} \inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t} \mathop{\mathbb{E}}_{\widehat{\boldsymbol{y}}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t) - \inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) \right\}. \qquad (15.9)$$

where $\Delta(\mathcal{Y})$ is the set of distributions on $\mathcal{Y}$ and the $\langle\!\langle \cdots \rangle\!\rangle_{t=1}^{n}$ notation is the shorthand for the repeated application of operators, from $t = 1$ to n. For instance, we would write the minimax value of an abstract two-stage game in our notation as

$$\min_{a_1} \max_{b_1} \min_{a_2} \max_{b_2} \phi(a_1, b_1, a_2, b_2) = \left\langle\!\!\left\langle \min_{a_t} \max_{b_t} \right\rangle\!\!\right\rangle_{t=1}^{2} \phi(a_1, b_1, a_2, b_2).$$

Given any upper bound on minimax regret $\mathcal{V}_n(\mathcal{F})$, there exists a prediction method that guarantees such a bound; any lower bound on $\mathcal{V}_n(\mathcal{F})$ ensures the existence of a strategy for Nature that inflicts at least that much regret for any prediction method.

The link to the uniform martingale laws of large numbers comes from the following theorem proved in [29]: for the case of absolute loss $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t) = |\widehat{\boldsymbol{y}}_t - y_t|$, $\mathcal{Y} = [-1, 1]$, and $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$, it holds that

$$\mathcal{R}_n^{seq}(\mathcal{F}) \le \mathcal{V}_n(\mathcal{F}) \le 2\mathcal{R}_n^{seq}(\mathcal{F}) \qquad (15.10)$$

where $\mathcal{R}_n^{seq}(\mathcal{F}) = \sup_{\mathbf{x}} \mathcal{R}_n^{seq}(\mathcal{F}; \mathbf{x})$ as defined in (15.4). A similar statement holds for more general loss functions. It is also possible to prove the upper bound for a more general non-supervised scenario (such as online convex optimization) in terms of the sequential Rademacher complexity of the loss class $\boldsymbol{\ell} \circ \mathcal{F} = \{\boldsymbol{\ell} \circ f : f \in \mathcal{F}\}$.

Together with the results of Sect. 15.2, one obtains a characterization of the *existence* of an algorithm with diminishing regret. The sequential complexities discussed earlier also provide rates of convergence of minimax regret to zero. As in the case of statistical learning, it is possible to establish control of sequential covering numbers, combinatorial parameters, or—directly—sequential Rademacher complexity for the particular questions at hand. Without much work, this approach yields rates of convergence of minimax regret for such classes as neural networks, decision trees, and so forth (see [29]). For many of these, a computationally feasible algorithm is unknown; nevertheless the minimax approach is able to discern the relevant complexity of the class in a non-constructive manner.

We remark that in many cases of interest sequential Rademacher complexity is of the same order as classical Rademacher complexity. In such cases, one obtains the same rates of convergence in online learning as in statistical learning with i.i.d. data. An analogous statement also holds for "curved losses," such as the square loss. Of course, the class of thresholds—as well as many other natural VC classes of binary-valued functions—is an exception to this equivalence, as the value $\mathcal{V}_n(\mathcal{F})$ does not decrease to zero and uniform martingale convergence does not hold. In some sense, the difficulty of learning thresholds in the online manner comes from the infinite precision in the arbitrary choices x_t of Nature, coupled with the lack of information coming from a binary-valued response y_t. The situation is conveniently remedied by considering Lipschitz functions, such as a "ramp" version of a threshold. Banach spaces, in particular, are a rich source of examples where the rates of online learning and statistical learning match.

15.4 Online Learning: Algorithms

As mentioned earlier, the upper bounds of Sect. 15.3 are non-algorithmic since the starting point is Eq. (15.10) — an upper bound that contains no prescription for how an algorithm should form predictions. While it is attractive to be able to understand the inherent complexity of online learning without the need to search for a prediction strategy, it is still desirable to find an algorithm that achieves the promised bounds. In this section, we recover the algorithms that were "lost" through the non-constructive derivations. In fact, we will see how to come up with prediction methods through a rather general recipe. As a bonus, the algorithms can also be used in the statistical learning scenario with i.i.d. data: an algorithm with a bound on regret (15.8) can also guarantee a bound on (15.7), under some conditions.

15.4.1 How to Relax

Let us examine (15.9) for a step $t \geq 1$. Since the choices $\widehat{\boldsymbol{y}}_1, y_1, \ldots, \widehat{\boldsymbol{y}}_{t-1}, y_{t-1}$ have been made, the sum $\sum_{s=1}^{t-1} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_s, y_s)$ does not enter into the optimization objective for x_t, q_t, y_t. Recall that, according to the protocol, x_t is observed before the mixed strategy q_t is chosen. Given x_t, the optimization problem for q_t, y_t now becomes

$$\inf_{q_t}\sup_{y_t}\mathbb{E}_{\widehat{\boldsymbol{y}}_t\sim q_t}\left\{\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t,y_t)+\left\langle\!\!\left\langle\sup_{x_s}\inf_{q_s\in\Delta(\mathcal{Y})}\sup_{y_s}\mathbb{E}_{\widehat{\boldsymbol{y}}_s\sim q_s}\right\rangle\!\!\right\rangle_{t+1}^{n}\left[\sum_{s=t+1}^{n}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_s,y_s)-\inf_{f\in\mathcal{F}}\sum_{s=1}^{n}\boldsymbol{\ell}(f(x_s),y_s)\right]\right\}$$

where we omit the normalization by n throughout. Let us denote the inner term, the optimization over variables from step $t+1$ onwards, by $\mathcal{V}_n\left(x_{1:t},y_{1:t}\right)$. Henceforth, we use the notation $x_{1:t}=(x_1,\ldots,x_t)$. We may think of $\mathcal{V}_n\left(x_{1:t},y_{1:t}\right)$ as a conditional minimax value, given the prefix of data up to time t. With this notation, the optimization objective at time t, given x_t, is

$$\inf_{q_t}\sup_{y_t}\mathbb{E}_{\widehat{\boldsymbol{y}}_t\sim q_t}\left\{\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t,y_t)+\mathcal{V}_n\left(x_{1:t},y_{1:t}\right)\right\}$$

and the recursive definition of the conditional value is

$$\mathcal{V}_n\left(x_{1:t-1},y_{1:t-1}\right)=\sup_{x_t}\inf_{q_t}\sup_{y_t}\mathbb{E}_{\widehat{\boldsymbol{y}}_t\sim q_t}\left\{\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t,y_t)+\mathcal{V}_n\left(x_{1:t},y_{1:t}\right)\right\}.$$

To make the recursive definition complete, the starting point is taken to be

$$\mathcal{V}_n\left(x_{1:n},y_{1:n}\right)=-\inf_{f\in\mathcal{F}}\sum_{s=1}^{n}\boldsymbol{\ell}(f(x_s),y_s).$$

It is then easy to see that the terminal point of the recursive definition yields $\mathcal{V}_n(\mathcal{F})=\frac{1}{n}\mathcal{V}_n\left(\emptyset\right)$. An optimal regret minimization algorithm is given by any choice

$$q_t^*\;\in\;\operatorname*{argmin}_{q_t}\;\sup_{y_t}\mathbb{E}_{\widehat{\boldsymbol{y}}_t\sim q_t}\left\{\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t,y_t)+\mathcal{V}_n\left(x_{1:t},y_{1:t}\right)\right\}$$

yet this choice is likely to be computationally infeasible. While written as a dynamic programming problem, the function $\mathcal{V}_n\left(x_{1:t},y_{1:t}\right)$ needs to be computed for all possible sequences—an insurmountable task for any interesting scenario. The idea put forth in [27] is to derive upper bounds on the conditional value. To this end, let $\mathbf{Rel}_n\left(x_{1:t},y_{1:t}\right)$ be a function $\cup_{t=1,\ldots,n}(\mathcal{X}\times\mathcal{Y})^t\to\mathbb{R}$ such that

$$\mathbf{Rel}_n\left(x_{1:n},y_{1:n}\right)\geq-\inf_{f\in\mathcal{F}}\sum_{s=1}^{n}\boldsymbol{\ell}(f(x_s),y_s)\qquad(15.11)$$

and

$$\mathbf{Rel}_n\left(x_{1:t-1},y_{1:t-1}\right)\geq\sup_{x_t}\inf_{q_t}\sup_{y_t}\mathbb{E}_{\widehat{\boldsymbol{y}}_t\sim q_t}\left\{\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t,y_t)+\mathbf{Rel}_n\left(x_{1:t},y_{1:t}\right)\right\}.\qquad(15.12)$$

The function $\mathbf{Rel}_n\left(\right)$ is called a *relaxation.* One of the main tools for verifying the *admissibility condition* (15.12) is the minimax theorem, as the maximin dual objective is often easier to analyze. Once admissibility is established, the algorithm

$$q_t^* \in \underset{q_t}{\operatorname{argmin}} \sup_{y_t} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \left\{ \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t) + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right) \right\} \tag{15.13}$$

automatically comes with a regret guarantee of $\frac{1}{n}\mathbf{Rel}_n\left(\emptyset\right)$ (see [27]). The search for computationally feasible regret minimization algorithms is thus reduced to finding an appropriate relaxation that is not too much larger than the conditional value. This is where the techniques from Sect. 15.2 come in.

Suppose $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, y)$ is 1-Lipschitz in the first coordinate. By sequential symmetrization, it is possible to show that the conditional sequential Rademacher complexity

$$\mathcal{R}(x_{1:t}, y_{1:t}) \triangleq \sup_{\mathbf{x}} \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ 2 \sum_{s=t+1}^{n} \epsilon_s f(\mathbf{x}_s(\epsilon)) - \sum_{s=1}^{t} \boldsymbol{\ell}(f(x_s), y_s) \right\} \tag{15.14}$$

is an admissible relaxation, where expectation is over $\epsilon_{t+1:n}$, the supremum is taken over trees $\mathbf{x}$ of depth $n-t$, and indexing of the tree starts at $t+1$ for simplicity. Observe that (15.14) reduces to the sequential Rademacher complexity when $t=0$. At the other extreme, it satisfies (15.11) with equality for $t=n$. In view of (15.10), an algorithm that uses this relaxation is nearly optimal. However, the supremum over $\mathbf{x}$ is not computationally tractable in general. We thus take (15.14) as a starting point for finding relaxations, and the main goal is to remove the supremum over $\mathbf{x}$ via further inequalities or via sampling, as illustrated in the next two paragraphs.

Let us illustrate the idea of "removing a tree" in (15.14) on the example of a finite class $\mathcal{F}$ of functions $\mathcal{X} \to [-1,1]$. Defining $L_t(f) = \sum_{s=1}^{t} \boldsymbol{\ell}(f(x_s), y_s)$ to be the cumulative loss of f at time t, we have, given any $\mathbf{x}$ and for any $\lambda > 0$,

$$\begin{aligned} &\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ 2 \sum_{s=t+1}^{n} \epsilon_s f(\mathbf{x}_s(\epsilon)) - L_t(f) \right\} \\ &\qquad \le \frac{1}{\lambda} \log \sum_{f \in \mathcal{F}} \mathbb{E} \exp \left\{ 2\lambda \sum_{s=t+1}^{n} \epsilon_s f(\mathbf{x}_s(\epsilon)) - \lambda L_t(f) \right\} \\ &\qquad \le \frac{1}{\lambda} \log \sum_{f \in \mathcal{F}} \exp\left\{-\lambda L_t(f)\right\} + 2\lambda(n-t). \end{aligned}$$

Using the last upper bound as a relaxation, we immediately obtain a parameter-free version of the celebrated Weighted Majority (or, the Aggregating Algorithm) [41, 23]. We refer for the details to [27], where a number of known and novel methods are derived more or less mechanically by following the above line of reasoning.

As emphasized throughout this chapter, the rates for online learning and for statistical learning often match. In these cases, there is hope that the supremum over a tree $\mathbf{x}$ in (15.14) can be replaced by an expectation over an i.i.d. draw $x_{t+1}, \dots, x_n$ from a fixed distribution $\mathcal{D}$. Under appropriate conditions (see [27]), one can then obtain a randomized method of the "random rollout" style. At time t, we first draw $x_{t+1}, \dots, x_n \sim \mathcal{D}$ and Rademacher random variables $\epsilon_{t+1}, \dots, \epsilon_n$. The randomized prediction is given by

$$q_t^* = \underset{q_t}{\operatorname{argmin}} \sup_{y_t} \left\{ \mathbb{E}\left[\ell(\hat{y}_t, y_t)\right] + \sup_{f \in \mathcal{F}} \left\{ 2 \sum_{s=t+1}^{n} \epsilon_s f(x_s) - \sum_{s=1}^{t} \ell(f(x_s), y_s) \right\} \right\}.$$

In some sense, the "future" is simulated through an i.i.d. draw rather than a worst-case tree $\mathbf{x}$. This technique leads to a host of efficient randomized prediction methods, and an example will be presented in Sect. 15.4.3.

In summary, the relaxation techniques give a principled way of deriving computationally feasible online learning algorithms. The uniform martingale laws of large numbers and the sequential complexity measures described earlier become a toolbox for such derivations.

15.4.2 From Online to Statistical Learning: The Improper Way

To describe the next idea, let us for simplicity fix $\mathcal{Y} = \{0,1\}$, $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, y) = \mathbf{I}\{\widehat{\boldsymbol{y}} \neq y\}$, and $\mathcal{F}$ a class of functions $\mathcal{X} \to \mathcal{Y}$. Then $\mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t) = |q_t - y_t|$ and (15.13) becomes

$$q_t^* = \underset{q_t}{\operatorname{argmin}} \max_{y_t \in \{0,1\}} \left\{ |q_t - y_t| + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right) \right\}$$

which is equal to

$$q_t^* = \frac{1}{2}\left(1 + \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, 1)\right) - \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, 0)\right)\right). \tag{15.15}$$

Since q_t^* is calculated based on x_t, we may write the solution as a function

$$f_t(x) = \frac{1}{2}\left(1 + \mathbf{Rel}_n\left((x_{1:t-1}, x), (y_{1:t-1}, 1)\right) - \mathbf{Rel}_n\left((x_{1:t-1}, x), (y_{1:t-1}, 0)\right)\right).$$

Then the guarantee given by the relaxation algorithm can be written as

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} |f_t(x_t) - y_t| - \inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \mathbf{I}\{f(x_t) \neq y_t\}\right] \leq \frac{1}{n}\mathbf{Rel}_n(\emptyset).$$

The so-called online-to-batch conversion [12, 11] method defines

$$\widehat{f} = \frac{1}{n}\sum_{t=1}^{n} f_t,$$

the average of the trajectory output by the online learning algorithm. If data $(X_1, Y_1), \ldots, (X_n, Y_n)$ presented to the online learning algorithm are i.i.d. with a common distribution $P_{X \times Y}$, an easy calculation shows that $\widehat{f}$ enjoys a statistical learning guarantee

$$\mathbb{E}|\widehat{f}(X) - Y| - \inf_{f \in \mathcal{F}} \mathbb{E}|f(X) - Y| \leq \frac{1}{n}\mathbf{Rel}_n(\emptyset).$$

Note that $\widehat{f}$ is $[0,1]$-valued, and the first term can be interpreted as the expected indicator loss of a randomized method. Randomized prediction methods are not commonplace in statistical learning, but here they arise naturally because of the non-convexity of the indicator loss. Whenever the loss function is convex and $\mathcal{Y}$ is a convex set (e.g., square loss for regression) there is no need for randomized predictions.

The important point we would like to discuss is about the nature of the function $\widehat{f}$. Observe that $\widehat{f}$ is not necessarily in $\mathcal{F}$, as it is obtained purely as an average of point-wise solutions in (15.15). Such a learning method is called "improper." In contrast, the most-studied method in statistical learning, the empirical risk minimization (ERM) algorithm, is "proper," as it selects a member of $\mathcal{F}$. Recently, however, it was shown that ERM—and any selector method for that matter—is suboptimal for the problem of learning with square loss and a finite set of functions [21, 18]. Several "improper" algorithms were suggested, mixing the finite set of functions via a convex combination [7, 20]. These methods, however, appear somewhat ad hoc, as there is presently no language in statistical learning theory for talking about improper learning methods. This is where online learning appears to come in and give us such a language.

The power of the above reasoning is demonstrated in [28], where optimal rates are exhibited for the problem of online regression with square loss through a direct analysis of the value in (15.8). Because of the curvature of the loss, a slightly different quantity (an offset sequential Rademacher complexity) governs the rates of convergence. An algorithm guaranteeing these optimal rates automatically follows from a recipe similar to that for the absolute loss. Coupled with online-to-batch conversion, the method yields optimal rates for *statistical learning* with square loss whenever sequential complexities are of the same order of magnitude as the i.i.d. ones. Remarkably, this method is very different from that obtained in [32] for the same problem.

15.4.3 Application: Online Node Classification

We now consider the problem of online node classification and illustrate how Rademacher relaxation and the idea of randomized methods can be used to develop an efficient prediction algorithm with low regret. The motivation for the setting comes, for instance, from the need by advertising agencies to predict in a sequential manner whether, say, an individual in a social network is of a particular type, given information about her friends. The problem cannot be modeled as i.i.d., and the online learning framework provides an attractive alternative.

More precisely, assume we are provided with a weighted graph $G = (V, E, W)$ where V is the set of vertices, E the set of edges, and $W : E \to [-1, 1]$ is the weight of each edge, indicating similarity/dissimilarity between vertices. In a sequential manner, the vertices of G are presented without repetition. In round $t \le n \le |V|$, after a node v_t is presented, the learner predicts the label of the node $\widehat{\boldsymbol{y}}_t \in \{\pm 1\}$ and the true label $y_t \in \{\pm 1\}$ is subsequently revealed. Informally, the structure of the problem is such that similar nodes (as measured by W) should have similar labels, and dissimilar nodes should have different labels. For a class $\mathcal{F}^G \subseteq \{\pm 1\}^V$ of labelings of vertices, the (unnormalized) regret is defined as

$$\sum_{t=1}^{n} \mathbf{I}\{\widehat{\boldsymbol{y}}_t \neq y_t\} - \inf_{f \in \mathcal{F}^G} \sum_{t=1}^{n} \mathbf{I}\{f(v_t) \neq y_t\}$$
$$= \frac{1}{2}\left(\sum_{t=1}^{n} (-\widehat{\boldsymbol{y}}_t y_t) - \inf_{f \in \mathcal{F}^G} \sum_{t=1}^{n} (-f(v_t) y_t) \right)$$

where the last step is because $\mathbf{I}\{a \neq b\} = \frac{1 - a \cdot b}{2}$ for $a, b \in \{\pm 1\}$. The conditional sequential Rademacher complexity (15.14) is then given by

$$\mathcal{R}(v_{1:t}, y_{1:t}) = \sup_{\mathbf{v}} \mathbb{E} \sup_{f \in \mathcal{F}^G} \left\{ \sum_{s=t+1}^{|V|} \epsilon_s f(\mathbf{v}_s(\epsilon)) + \frac{1}{2} \sum_{s=1}^{t} y_s f(v_s) \right\}$$

where $\mathbf{v}$ is a $V \setminus \{v_1, \ldots, v_t\}$-valued tree that has sequences of non-repeating nodes on each path. Since $\sum_{s=t+1}^{n} \epsilon_s f(\mathbf{v}_s(\epsilon))$ is invariant w.r.t. the order in which the nodes appear in the tree, we can take these nodes without repetition in any order we please:

$$\mathcal{R}(v_{1:t}, y_{1:t}) = \mathbb{E} \sup_{f \in \mathcal{F}^G} \left\{ \sum_{s=t+1}^{|V|} \epsilon_s f(v_s) + \frac{1}{2} \sum_{s=1}^{t} y_s f(v_s) \right\} \tag{15.16}$$

where $v_{t+1}, \ldots v_n$ is any fixed order in which future nodes could appear (e.g., ascending order). Now, depending on G, the supremum over $\mathcal{F}^G$ might still be difficult to compute. We therefore might want to further relax the problem by using a larger set $\overline{\mathcal{F}^G} \supseteq \mathcal{F}^G$ for the supremum in (15.16). The randomized algorithm we derive from such a relaxation is given by first drawing $\epsilon_1, \ldots, \epsilon_n$ Rademacher random variables and then predicting $\widehat{\boldsymbol{y}}_t \in \{\pm 1\}$ by picking $+1$ with probability

$$q_t^* = \frac{1}{2} + \frac{1}{2}\text{Clip}\left(\sup_{f \in \overline{\mathcal{F}^G}} \left\{ \sum_{s=t+1}^{|V|} \epsilon_s f(v_s) + \frac{1}{2}\sum_{s=1}^{t-1} y_s f(v_s) + \frac{1}{2} f(v_t) \right\} \right. \tag{15.17}$$
$$\left. - \sup_{f \in \overline{\mathcal{F}^G}} \left\{ \sum_{s=t+1}^{|V|} \epsilon_s f(v_s) + \frac{1}{2}\sum_{s=1}^{t-1} y_s f(v_s) - \frac{1}{2} f(v_t) \right\} \right)$$

where $\text{Clip}(\alpha) = \alpha$ if $|\alpha| \le 1$ and $\text{Clip}(\alpha) = \text{sign}(\alpha)$ otherwise. The expected regret for the above algorithm is bounded by

$$\frac{1}{n}\mathbf{Rel}_n(\emptyset) = \frac{1}{n}\mathbb{E}\left[\sup_{f \in \overline{\mathcal{F}^G}} \sum_{t=1}^{|V|} \epsilon_t f(v_t)\right]. \tag{15.18}$$

Since the algorithm in (15.17) is independent of n, the regret guarantee in fact holds for any $n \le |V|$.

Let us discuss the computational requirements of the proposed method. From Eq. (15.17), our randomized prediction q_t^* can be obtained by solving two optimization problems per round (say round t) as:

$$\begin{aligned} \text{Val}_t^+ = \text{Maximize} \quad & f^\top X_t^+ & \qquad \text{Val}_t^- = \text{Maximize} \quad & f^\top X_t^- \\ \text{subject to} \quad & f \in \overline{\mathcal{F}^G} & \text{subject to} \quad & f \in \overline{\mathcal{F}^G} \end{aligned} \tag{15.19}$$

where X_t^+ is the vector such that $X_t^+[v_t] = +1/2$, $X_t^+[v_s] = \frac{1}{2}y_s$ for any $s \le t-1$, and $X_t^+[v_s] = \epsilon_s$ when $s > t$. Similarly X_t^- is the vector such that $X_t^-[v_t] = -1/2$, $X_t^-[v_s] = \frac{1}{2}y_s$ for any $s \le t-1$, and $X_t^-[v_s] = \epsilon_s$ when $s > t$. The randomized predictor is given by $q_t^* = 0.5 + 0.5\ \text{Clip}(\text{Val}_t^+ - \text{Val}_t^-)$. To further detail the computational requirements, consider the following example.

Example 15.3 (Laplacian Node Classification). Assume that W is some matrix measuring similarities/dissimilarities between nodes and let L denote the Laplacian matrix of the graph. A natural choice for a class $\mathcal{F}^G$ is then

$$\begin{aligned} \mathcal{F}^G &= \left\{ f \in \{\pm 1\}^{|V|} : \sum_{e_{u,v} \in E} |W(e_{u,v})| \left(1 - \text{sgn}(W(e_{u,v})) f(u) f(v)\right) \le K \right\} \\ &= \left\{ f \in \{\pm 1\}^{|V|} : f^\top L f \le K \right\} \subseteq \left\{ f \in [-1,1]^{|V|} : f^\top L f \le K \right\} \triangleq \overline{\mathcal{F}^G} \end{aligned}$$

for some $K > 0$. The optimization problem (15.19) with the above set is computationally feasible. We observe that the bound in (15.18) only increases if $\overline{\mathcal{F}^G}$ is replaced with a superset obtained as follows. Since $[-1,1]^{|V|} \subset \left\{\|f\|_2 \leq \sqrt{|V|}\right\}$, it holds that $\overline{\mathcal{F}^G} \subseteq \left\{f \in \mathbb{R}^{|V|} : f^\top M f \leq 1\right\}$ where $M = \frac{1}{2K}L + \frac{1}{2|V|}I_{|V|}$. Hence, the bound on expected regret is

$$\frac{1}{n}\mathbb{E}\left[\epsilon^\top M^{-1}\epsilon\right] \leq \frac{1}{n}\sqrt{\sum_{j=1}^{|V|}\frac{1}{\lambda_j\left(M\right)}}$$

where for any square matrix M, $\lambda_j\left(M\right)$ denotes the j^{th} eigenvalue of M.

15.5 Completing the Circle: Vapnik and Chervonenkis 1968

Some of the key statements of the celebrated paper of Vapnik and Chervonenkis [40] already appeared in the three-page abstract [39] in 1968. Perhaps the less-known paper of Vapnik and Chervonenkis that also appeared in 1968 (submitted in 1966) is the manuscript "Algorithms with Complete Memory and Recurrent Algorithms in Pattern Recognition Learning" [38]. We would like to finish our chapter with a short discussion of this work, as it can be viewed through the prism of online methods.

In the early sixties, the interest in studying "learning machines" was fueled by the introduction of the Perceptron algorithm by Rosenblatt in 1957 and by the mistake-bound proof of Novikoff in 1962. According to [36, p. 33], two inductive approaches were discussed at the seminars of the Moscow Institute of Control Sciences starting in 1963: (a) the principle of stochastic approximation, and (b) the principle of empirical risk minimization. The first approach to minimizing the risk functional is *recurrent*, or online, and finds its roots in the work of Robbins and Monroe in the early 1950s. In a string of influential papers, Aizerman, Braverman, and Roeznoer [4, 3, 5] introduced a generalization of the Perceptron using the idea of potential functions (presently termed *kernels*). Both the Perceptron and the potential-based updates were shown to be instances of the stochastic approximation approach, minimizing an appropriate functional. In parallel to these developments, Vapnik and Chervonenkis were working on the second approach to learning – direct minimization of the empirical risk.

It is especially interesting to read [38] against this backdrop. The authors find the length of the training sequence that suffices for ERM to achieve the desired accuracy for linear classifiers in d dimensions, under the assumption of realizability. This sample complexity is computed based on the fact that the growth function behaves polynomially with exponent d, a precursor to the general combinatorial result. The authors compare this sample complexity to the one obtained from the Perceptron mistake bound (with a conversion to the i.i.d. guarantee). The latter sample complexity depends inversely on the square of the margin. Vapnik and Chervonenkis discuss the fact that for a Perceptron-based approach one cannot obtain distribution-free statements, as the required sample complexity becomes infinite when the margin is taken to zero. In contrast, sample complexity of ERM, irrespective of the distribution, can be upper bounded in terms of the dimension and independently of the margin. The authors also note that the margin may enter into the computation time of ERM – a statement that already foreshadows the focus on computational complexity by Valiant in 1984.

The discussion by Vapnik and Chervonenkis on sample complexity of recurrent vs. full-memory algorithms can be seen through the lens of results in the present chapter. The Perceptron is an online learning method (we can write down a relaxation that yields the corresponding update), and its convergence is governed by uniform martingale laws of large numbers. Such a distribution-free convergence is not possible for thresholds, as shown in Example 15.2. More generally, an attempt to pass to a statistical learning guarantee through an online statement can only be successful if there is no gap between the classical and sequential complexities. On the positive side, there are many examples where uniform martingale convergence is equivalent to i.i.d. convergence, in which case the world of recurrent algorithms meets the world of empirical risk minimization. It is rather remarkable that the topic of online vs. batch learning algorithms—a current topic of interest in the learning community—was already explored by Vapnik and Chervonenkis in the 1960s.

Acknowledgements We gratefully acknowledge the support of NSF under grants CAREER DMS-0954737 and CCF-1116928.

References

1. Abernethy, J., Agarwal, A., Bartlett, P., Rakhlin, A.: A stochastic view of optimal regret through minimax duality. In: Proceedings of the 22th Annual Conference on Learning Theory (2009)
2. Abernethy, J., Bartlett, P.L., Rakhlin, A., Tewari, A.: Optimal strategies and minimax lower bounds for online convex games. In: Proceedings of the 21st Annual Conference on Learning Theory, pp. 414–424. Omnipress (2008)
3. Aizerman, M.A., Braverman, E.M., Roeznoer, L.I.: The probability problem of pattern recognition learning and the method of potential functions. Avtomatika i Telemekhanika **25**, 1175–1193 (1964)
4. Aizerman, M.A., Braverman, E.M., Roeznoer, L.I.: Theoretical foundations of the potential function method in pattern recognition learning. Avtomatika i Telemekhanika **25**, 821–837 (1964)
5. Aizerman, M.A., Braverman, E.M., Rozonoer, L.I.: The Method of Potential Functions in the Theory of Machine Learning. Nauka, Moscow (1970)

6. Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D.: Scale-sensitive dimensions, uniform convergence, and learnability. Journal of the ACM **44**(4), 615–631 (1997)
7. Audibert, J.: Progressive mixture rules are deviation suboptimal. Advances in Neural Information Processing Systems **20**(2), 41–48
8. Bartlett, P.L., Long, P.M., Williamson, R.C.: Fat-shattering and the learnability of real-valued functions. Journal of Computer and System Sciences **52**(3), 434–452 (1996)
9. Bartlett, P.L., Mendelson, S.: Rademacher and Gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research **3**, 463–482 (2002)
10. Ben-David, S., Pál, D., Shalev-Shwartz, S.: Agnostic online learning. In: Proceedings of the 22th Annual Conference on Learning Theory (2009)
11. Cesa-Bianchi, N., Conconi, A., Gentile, C.: On the generalization ability of on-line learning algorithms. IEEE Transactions on Information Theory **50**(9), 2050–2057 (2004)
12. Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D.P., Schapire, R.E., Warmuth, M.K.: How to use expert advice. Journal of the ACM **44**(3), 427–485 (1997)
13. Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press (2006)
14. Van Der Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, New York (1996)
15. Dudley, R.M.: A course on empirical processes. In: P.L. Hennequin (ed.) École d'Été de Probabilités de Saint-Flour XII—1982, *Lecture Notes in Mathematics*, vol. 1097, pp. 2–142. Springer, Berlin (1984)
16. Dudley, R.M.: Uniform Central Limit Theorems. Cambridge University Press (1999)
17. Van de Geer, S.A.: Empirical Processes in M-Estimation. Cambridge University Press (2000)
18. Juditsky, A., Rigollet, P., Tsybakov, A.: Learning by mirror averaging. Annals of Statistics **36**(5), 2183–2206 (2008)
19. Kearns, M.J., Schapire, R.E.: Efficient distribution-free learning of probabilistic concepts. Journal of Computer and System Sciences **48**(3), 464–497 (1994)
20. Lecué, G., Mendelson, S.: Aggregation via empirical risk minimization. Probability Theory and Related Fields **145**(3), 591–613 (2009)
21. Lee, W.S., Bartlett, P.L., Williamson, R.C.: The importance of convexity in learning with squared loss. Information Theory, IEEE Transactions on **44**(5), 1974–1980 (1998)
22. Littlestone, N.: Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. Machine Learning **2**(4), 285–318 (1988)
23. Littlestone, N., Warmuth, M.K.: The weighted majority algorithm. Information and Computation **108**(2), 212–261 (1994)
24. Mendelson, S., Vershynin, R.: Entropy and the combinatorial dimension. Inventiones mathematicae **152**(1), 37–55 (2003)
25. Pisier, G.: Martingales with values in uniformly convex spaces. Israel Journal of Mathematics **20**, 326–350 (1975)
26. Pollard, D.: Convergence of stochastic processes. Springer, Berlin (1984)
27. Rakhlin, A., Shamir, O., Sridharan, K.: Relax and randomize: From value to algorithms. In: Advances in Neural Information Processing Systems 25, pp. 2150–2158 (2012)
28. Rakhlin, A., Sridharan, K.: Online Nonparametric Regression. In: The 27th Annual Conference on Learning Theory (2014)
29. Rakhlin, A., Sridharan, K., Tewari, A.: Online learning: Random averages, combinatorial parameters, and learnability. Advances in Neural Information Processing Systems 23 pp. 1984–1992 (2010)
30. Rakhlin, A., Sridharan, K., Tewari, A.: Online learning: Stochastic, constrained, and smoothed adversaries. In: Advances in Neural Information Processing Systems (2011)
31. Rakhlin, A., Sridharan, K., Tewari, A.: Sequential complexities and uniform martingale laws of large numbers. Probability Theory and Related Fields (2014)

32. Rakhlin, A., Sridharan, K., Tsybakov, A.: Empirical entropy, minimax regret and minimax risk. Bernoulli Journal (2015). Forthcoming
33. Rudelson, M., Vershynin, R.: Combinatorics of random processes and sections of convex bodies. Annals of Mathematics **164**(2), 603–648 (2006)
34. Sridharan, K., Tewari, A.: Convex games in Banach spaces. In: Proceedings of the 23nd Annual Conference on Learning Theory (2010)
35. Steele, J.M.: Empirical discrepancies and subadditive processes. The Annals of Probability **6**(1), 118–127 (1978)
36. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
37. Vapnik, V., Chervonenkis, A.Y.: The necessary and sufficient conditions for the uniform convergence of averages to their expected values. Teoriya Veroyatnostei i Ee Primeneniya **26**(3), 543–564 (1981)
38. Vapnik, V.N., Chervonenkis, A.J.: Algorithms with complete memory and recurrent algorithms in pattern recognition learning. Avtomatika i Telemekhanika **4**, 95–106 (1968)
39. Vapnik, V.N., Chervonenkis, A.Y.: Uniform convergence of frequencies of occurrence of events to their probabilities. Doklady Akademii Nauk SSSR **181**, 915–918 (1968)
40. Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications **16**(2), 264–280 (1971). This volume, Chap. 3
41. Vovk, V.: Aggregating strategies. In: Proceedings of the Third Annual Workshop on Computational Learning Theory, pp. 371–386. Morgan Kaufmann, San Mateo (1990)