Proceedings of Machine Learning Research vol 65:1–19, 2017

On Equivalence of Martingale Tail Bounds and Deterministic Regret Inequalities

Alexander Rakhlin RAKHLIN@WHARTON.UPENN.EDU
University of Pennsylvania

Karthik Sridharan SRIDHARAN@CS.CORNELL.EDU
Cornell University

Abstract

We study an equivalence of (i) deterministic pathwise statements appearing in the online learning literature (termed *regret bounds*), (ii) high-probability tail bounds for the supremum of a collection of martingales (of a specific form arising from uniform laws of large numbers), and (iii) in-expectation bounds for the supremum. By virtue of the equivalence, we prove exponential tail bounds for norms of Banach space valued martingales via deterministic regret bounds for the online mirror descent algorithm with an adaptive step size. We show that the phenomenon extends beyond the setting of online linear optimization and present the equivalence for the supervised online learning setting.

Keywords: martingale inequalities; online learning

1. Introduction

The paper investigates equivalence of regret inequalities that hold *for all sequences* and probabilistic inequalities for martingales. In recent years, it was shown that *existence* of regret-minimization strategies can be certified non-algorithmically by studying certain stochastic processes. In this paper, we make the connection in the opposite direction and show a certain equivalence. We present several new deviation inequalities that follow with surprising ease from pathwise regret inequalities, while it is far from clear how to prove them with other methods.

Arguably the simplest example of the equivalence between prediction of individual sequences and probabilistic inequalities can be found in the work of Cover (1965). Consider the task of predicting a binary sequence $\boldsymbol{y} = (y_1, \dots, y_n) \in \{\pm 1\}^n$ in an online manner. Let $\phi : \{\pm 1\}^n \to [0,1]$ be $1/n$-Lipschitz with respect to the Hamming distance. Then there exists a randomized strategy such that

$$\forall \boldsymbol{y}, \quad \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\left\{\widehat{\boldsymbol{y}}_t \neq y_t\right\}\right] \leq \phi(\boldsymbol{y}) \tag{1}$$

if and only if $\mathbb{E}\phi(\boldsymbol{\varepsilon}) \geq 1/2$. The expectation in (1) is with respect to the randomized predictions $\widehat{\boldsymbol{y}}_t = \widehat{\boldsymbol{y}}_t(y_1, \dots, y_{t-1}) \in \{\pm 1\}$ made by the algorithm, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$ is a sequence of independent Rademacher random variables, and $\mathbf{1}\{\}$ is the indicator loss function. While this result is not difficult to prove by backward induction (see e.g. (Rakhlin and Sridharan, 2016)), the message is rather intriguing: existence of a prediction strategy with a given

mistake bound ϕ is equivalent to a simple statement about the expected value of ϕ with respect to the uniform distribution. Furthermore, the Lipschitz condition on ϕ implies a high-probability bound for the deviation of ϕ from $\mathbb{E}\phi$ via McDiarmid's inequality.

Our second example of the equivalence is in the setting of online linear optimization. Consider the unit Euclidean ball $\mathcal{B}$ in $\mathbb{R}^d$. Let $z_1, \dots, z_n \in \mathcal{B}$ and define, recursively, the Euclidean projections

$$\widehat{\boldsymbol{y}}_{t+1} = \widehat{\boldsymbol{y}}_{t+1}(z_1, \dots, z_t) = \mathrm{Proj}_{\mathcal{B}}\left(\widehat{\boldsymbol{y}}_t - n^{-1/2} z_t\right) \tag{2}$$

for each $t = 1, \dots, n$, with the initial value $\widehat{\boldsymbol{y}}_1 = 0$. Elementary algebra[1] shows that for any $f \in \mathcal{B}$, the regret inequality $\sum_{t=1}^n \langle \widehat{\boldsymbol{y}}_t - f, z_t \rangle \le \sqrt{n}$ holds deterministically for any sequence $z_1, \dots, z_n \in \mathcal{B}$. By optimally choosing f in the direction of the sum, we re-write this statement equivalently as

$$\left\| \sum_{t=1}^n z_t \right\| - \sqrt{n} \le \sum_{t=1}^n \langle \widehat{\boldsymbol{y}}_t, -z_t \rangle \,. \tag{3}$$

Since the inequality holds pathwise, by applying it to a $\mathcal{B}$-valued martingale difference sequence $-Z_1, \dots, -Z_n$, we conclude that

$$P\left(\left\| \sum_{t=1}^n Z_t \right\| - \sqrt{n} > u \right) \le P\left(\sum_{t=1}^n \langle \widehat{\boldsymbol{y}}_t, Z_t \rangle > u \right) \le \exp\left\{ -\frac{u^2}{2n} \right\}. \tag{4}$$

The latter upper bound is an application of the Azuma-Hoeffding's inequality. Indeed, the process $(\widehat{\boldsymbol{y}}_t)$ is predictable with respect to $\sigma(Z_1, \dots, Z_{t-1})$, and thus $(\langle \widehat{\boldsymbol{y}}_t, Z_t \rangle)$ is a $[-1, 1]$-valued martingale difference sequence. It is worth emphasizing the conclusion: *one-sided deviation tail bounds for a norm of a vector-valued martingale can be deduced from tail bounds for real-valued martingales with the help of a deterministic regret inequality.*

Next, integrating the tail bound in (4) yields a seemingly weaker in-expectation statement

$$\mathbb{E} \left\| \sum_{t=1}^n Z_t \right\| \le c\sqrt{n} \tag{5}$$

for an appropriate constant c. The twist in this uncomplicated story comes next: with the help of the minimax theorem, (Abernethy et al., 2009; Rakhlin et al., 2010) established *existence* of strategies $(\widehat{\boldsymbol{y}}_t)$ such that

$$\forall z_1, \dots, z_n, f \in \mathcal{B}, \qquad \sum_{t=1}^n \langle \widehat{\boldsymbol{y}}_t - f, z_t \rangle \le \sup \mathbb{E} \left\| \sum_{t=1}^n Z_t \right\|, \tag{6}$$

with the supremum taken over all $2\mathcal{B}$-valued martingale difference sequences. In view of (5), this bound is $c\sqrt{n}$.

What have we achieved? Let us summarize. The deterministic inequality (3), which holds for *all sequences*, implies a tail bound (4). The latter, in turn, implies an in-expectation bound (5), which implies (3) (with a worse constant) through a minimax argument, thus closing the loop. The equivalence—studied in depth in this paper—is informally stated below:

1. See the two-line proof in the Appendix, Lemma 12.

Informal: The following bounds imply each other: (a) an inequality that holds for all sequences; (b) a deviation tail probability for the size of a martingale; (c) an in-expectation bound on the size of a martingale.

The equivalence, in particular, allows us to *amplify* the in-expectation bounds to appropriate high-probability tail bounds.

While writing the paper, we learned of the *trajectorial approach*, extensively studied in recent years. In particular, it has been shown that Doob's maximal inequalities and Burkholder-Davis-Gundy inequalities have deterministic counterparts (Acciaio et al., 2013; Beiglböck and Nutz, 2014; Gushchin, 2014; Beiglböck and Siorpaes, 2015). The online learning literature contains a trove of pathwise inequalities, and further synthesis with the trajectorial approach (and the applications in mathematical finance) appears to be a promising direction.

This paper is organized as follows. In the next section, we extend the Euclidean result to martingales with values in Banach spaces and improve it by replacing $\sqrt{n}$ with square root of variation. In particular, we conclude a high probability self-normalized tail bound, a statement that appears to be difficult to obtain with other methods (see (Bercu et al., 2015; de la Peña et al., 2008) for a survey of techniques in this area). Section 3 is devoted to the analysis of equivalence for supervised learning. Finally, Section 4 shows that it is enough to consider dyadic martingales if one is interested in general martingale inequalities of a certain form.

2. Adaptive Bounds and Probabilistic Inequalities in Banach Spaces

For the case of the Euclidean (or Hilbertian) norm, it is easy to see that the $\sqrt{n}$ bound of (5) can be improved to a *distribution-dependent* quantity $\left(\sum_{t=1}^{n} \mathbb{E}\left\|Z_t\right\|^2\right)^{1/2}$. Given the equivalence sketched earlier, one may wonder whether this upper bound is also equivalent to a gradient-descent-like online method with a *sequence-dependent* variation governing the rate of convergence. Below, we indeed present such an equivalence for 2-smooth Banach spaces. Furthermore, the probabilistic tail bounds obtained this way appear to be novel.

Suppose that we have a norm $\|\cdot\|$ on some vector space such that $\|\cdot\|^2$ is a smooth function:

$$\|x+y\|^2 \le \|x\|^2 + \left\langle \nabla \|x\|^2, y \right\rangle + C' \|y\|^2 \tag{7}$$

for some $C' > 0$. Repeatedly using smoothness of the norm, we conclude that

$$\mathbb{E}\left\|\sum_{t=1}^{n} Z_t\right\|^2 \le C' \sum_{t=1}^{n} \mathbb{E}\left\|Z_t\right\|^2 \tag{8}$$

for any martingale difference sequence taking values in that vector space, since the cross-terms vanish. Instead of (8), we will work with the tighter inequality

$$\mathbb{E}\left\|\sum_{t=1}^{n} Z_t\right\| \le C\mathbb{E}\sqrt{\sum_{t=1}^{n} \left\|Z_t\right\|^2}. \tag{9}$$

Let $(\mathfrak{B}, \|\cdot\|)$ be a reflexive Banach space with dual space $(\mathfrak{B}_*, \|\cdot\|_*)$. Assume that $(\mathfrak{B}, \|\cdot\|)$ is 2-smooth (that is, $\rho(\delta) \triangleq \sup\left\{\frac{1}{2}(\|x+y\| + \|x-y\|) - 1 : \|x\| = 1, \|y\| = \delta\right\}$, the modulus of

smoothness, behaves as $c\delta^2$). Then there exists an equivalent norm $\|\cdot\|_{\mathfrak{B}}$ (in the sense that $c_1\|\cdot\|_{\mathfrak{B}} \le \|\cdot\| \le c_2\|\cdot\|_{\mathfrak{B}}$ for some possibly dimension-dependent c_1, c_2) that is smooth. In this case, we can expect that (9) holds for martingale difference sequences taking values in $\mathfrak{B}$.

Let us now argue this more formally, and also show equivalence to the existence of deterministic prediction strategies.

2.1. From regret inequality to expected value and back

Lemma 1 *Existence of a (deterministic) prediction strategy $(\widehat{\boldsymbol{y}}_t)_{t=1}^n$, with values $\widehat{\boldsymbol{y}}_t(z_1,\dots,z_{t-1})$ in the unit ball $\mathcal{B}_*$ of $\mathfrak{B}_*$ such that*

$$\forall z_1,\dots,z_n \in \mathfrak{B}, f \in \mathcal{B}_*, \qquad \sum_{t=1}^n \langle \widehat{\boldsymbol{y}}_t - f, z_t\rangle \le C\sqrt{\sum_{t=1}^n \|z_t\|^2} \tag{10}$$

for some C is equivalent to (9) (with a possibly different constant C) holding for all martingale difference sequences with values in $\mathfrak{B}$.

Proof By rearranging (10) as in (3), choosing a unit vector f, and taking an expectation on both sides implies (9) with the same constant C as in (10). We now argue the reverse direction: (9) implies existence of a strategy with a regret bound (10). First, consider an arbitrary collection $(X_1,\dots,X_n)$ of random variables taking values in an R-radius centered ball of $\mathfrak{B}$ and define the conditional expectations $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot \mid X_1,\dots,X_{t-1}]$. Observe that the collection $(X_t - \mathbb{E}_{t-1}X_t)$, $t = 1,\dots,n$, is a martingale difference sequence. Hence, by triangle inequality and our assumption,

$$\mathbb{E}\left\|\sum_{t=1}^n X_t\right\| - \mathbb{E}\sum_{t=1}^n \|\mathbb{E}_{t-1}X_t\| \le \mathbb{E}\left\|\sum_{t=1}^n (X_t - \mathbb{E}_{t-1}X_t)\right\| \le C\mathbb{E}\sqrt{\sum_{t=1}^n \|X_t - \mathbb{E}_{t-1}X_t\|^2}. \tag{11}$$

The right-most expression in (11) can be upper bounded by

$$\sqrt{2}C\mathbb{E}\left(\sqrt{\sum_{t=1}^n \|X_t\|^2} + \sqrt{\sum_{t=1}^n \|\mathbb{E}_{t-1}X_t\|^2}\right) \le \sqrt{8}C\mathbb{E}\sqrt{\sum_{t=1}^n \|X_t\|^2}. \tag{12}$$

To justify the last inequality, first observe that $\|\mathbb{E}_{t-1}X_t\| \le \mathbb{E}_{t-1}\|X_t\|$. Second, the function $x \mapsto \sqrt{A + x^2}$ is convex, and (12) follows by Jensen's inequality.

Combining (11) and (12), for any finite R and any collection $(X_1,\dots,X_n)$ with values in $R\cdot\mathcal{B}$,

$$\mathbb{E}\left\{\left\|\sum_{t=1}^n X_t\right\| - \sum_{t=1}^n \|\mathbb{E}_{t-1}X_t\| - C'\sqrt{\sum_{t=1}^n \|X_t\|^2}\right\} \le 0 \tag{13}$$

with $C' = \sqrt{8}C$. Writing

$$-\|\mathbb{E}_{t-1}X_t\| = \inf_{\|\widehat{\boldsymbol{y}}_t\|_* \le 1} \langle \widehat{\boldsymbol{y}}_t, \mathbb{E}_{t-1}X_t\rangle,$$

we conclude that

$$\sup \mathbb{E}\left\{\sum_{t=1}^n \inf_{\|\widehat{\boldsymbol{y}}_t\|_* \le 1} \langle \widehat{\boldsymbol{y}}_t, \mathbb{E}_{t-1}X_t\rangle - \inf_{\|f\|_* \le 1}\left\langle f, \sum_{t=1}^n X_t\right\rangle - C'\sqrt{\sum_{t=1}^n \|X_t\|^2}\right\} \le 0 \tag{14}$$

where the supremum is over the distributions of $(X_1, \ldots, X_n)$ with values in $R \cdot \mathcal{B}$. The rest of the argument can be seen as running the proof of (Abernethy et al., 2009) backwards. The minimax theorem holds because of finiteness of R, the radius of the support of the X_t's, via arguments in (Rakhlin et al., 2014, Appendix A). ■

Thankfully, the strategy that guarantees (10) is already known: it is an adaptive version of Mirror Descent. For completeness, the proof is provided in the Appendix. To define the strategy, we need the following fact: if $\mathfrak{B}$ is a 2-smooth Banach space, then there is a function $\mathcal{R}$ on the dual space $\mathfrak{B}_*$ which is strongly convex with respect to the norm $\|\cdot\|_*$. In fact, one can take the squared dual norm corresponding to the smooth equivalent norm on $\mathfrak{B}$ (Borwein et al., 2009). To avoid extra constants, let us simply assume that $\mathcal{R}$ is 1-strongly convex on the unit ball $\mathcal{B}_*$ of $\mathfrak{B}_*$. The function $\mathcal{R}$ induces the Bregman divergence $D_{\mathcal{R}} : \mathfrak{B}_* \times \mathfrak{B}_* \to \mathbb{R}$, defined as $D_{\mathcal{R}}(f, g) = \mathcal{R}(f) - \mathcal{R}(g) - \langle \nabla \mathcal{R}(g), f - g \rangle$.

Lemma 2 *Let $\mathcal{F} \subset \mathfrak{B}_*$ be a convex set. Define, recursively,*

$$\widehat{\boldsymbol{y}}_{t+1} = \widehat{\boldsymbol{y}}_{t+1}(z_1, \ldots, z_t) = \operatorname*{argmin}_{f \in \mathcal{F}} \ \eta_t \langle f, z_t \rangle + D_{\mathcal{R}}(f, \widehat{\boldsymbol{y}}_t) \tag{15}$$

with $\widehat{\boldsymbol{y}}_1 = 0$, $\eta_t \triangleq R_{\max} \left(\sum_{s=1}^t \|z_s\|^2 \right)^{-1/2}$, and $R_{\max}^2 \triangleq \sup_{f,g \in \mathcal{F}} D_{\mathcal{R}}(f, g)$. Then for any $f \in \mathcal{F}$ and any $z_1, \ldots, z_n \in \mathfrak{B}$,

$$\textstyle\sum_{t=1}^n \langle \widehat{\boldsymbol{y}}_t - f, z_t \rangle \le 2 R_{\max} \sqrt{\sum_{t=1}^n \|z_t\|^2}.$$

Lemma 2 is complementary to Lemma 1, as it gives the algorithm whose existence was guaranteed by Lemma 1. In Section 3, we will not have the luxury of producing an explicit algorithm, yet the equivalence will still be established.

2.2. From regret inequalities to tail bounds and back

We now start from a regret-minimization strategy and deduce a new probabilistic inequality for martingales. We then conclude the in-expectation bound and use the equivalence of Lemma 1 to close the loop.

The adaptive Mirror Descent algorithm of the previous section implies the following theorem:

Theorem 3 *Let $Z_1, \ldots, Z_n$ be a $\mathfrak{B}$-valued martingale difference sequence, and let $\mathbb{E}_t$ stand for the conditional expectation given $Z_1, \ldots, Z_t$. Define*

$$V_n = \sum_{t=1}^n 2 \|Z_t\|^2 \qquad \text{and} \qquad W_n = 2 \sum_{t=1}^n \mathbb{E}_{t-1} \|Z_t\|^2 , \tag{16}$$

which are assumed to have a finite expected value. For any $u > 0$, it holds that

$$P\left(\frac{\left\| \sum_{t=1}^n Z_t \right\| - 2 R_{\max} \sqrt{V_n}}{\sqrt{V_n + W_n + \left(\mathbb{E} \sqrt{V_n + W_n} \right)^2}} > u \right) \le \sqrt{2} \exp\left\{ -u^2/16 \right\}, \tag{17}$$

and for any $u \geq \sqrt{2}$*, it holds that*

$$P\left(\frac{\|\sum_{t=1}^n Z_t\| - 2R_{\max}\sqrt{V_n}}{\sqrt{(V_n + W_n + 1)\left(1 + \frac{1}{2}\log\left(V_n + W_n + 1\right)\right)}} \geq u\right) \leq \exp\left\{-u^2/2\right\}. \tag{18}$$

Furthermore, both bounds also hold with $W_n \equiv 0$ *and* $V_n = \sum_{t=1}^n \|Z_t\|^2$ *if the martingale differences are conditionally symmetric.*[2]

In addition to extending the Euclidean result of the previous section to Banach spaces, (17) and (18) offer several advantages. First, the bounds are n-independent. The deviations in (17) and (18) are *self-normalized* (that is, scaled by root-variation terms) and all the terms are either distribution-dependent or data-dependent, as in the case of the Student's t-statistic (de la Peña et al., 2008). The advantage of (18), especially in the case of conditional symmetry, is that all the terms, modulo the additive constants 1, are data-dependent. We are not aware of similar bounds for norms of random vectors in the literature, and we wish to stress that the proof of the result is almost immediate, given the regret inequality. We would also like to stress that Theorem 3 holds without any assumption on the martingale difference sequence beyond square integrability.

Proof [**Theorem 3**] We take $\mathcal{F}$ in Lemma 2 to be the unit ball in $\mathfrak{B}_*$, ensuring $\|\widehat{\boldsymbol{y}}_t\|_* \leq 1$. For any martingale difference sequence (Z_t) with values in $\mathfrak{B}$, the above lemma implies, by the definition of the norm,

$$\|\textstyle\sum_{t=1}^n Z_t\| - 2R_{\max}\sqrt{V_n} \leq \sum_{t=1}^n \langle \widehat{\boldsymbol{y}}_t, Z_t\rangle \tag{19}$$

deterministically for all sample paths. Dividing both sides by $\sqrt{V_n + W_n + \left(\mathbb{E}\sqrt{V_n + W_n}\right)^2}$, we conclude that the left-hand side in (17) is upper bounded by

$$P\left(\frac{\sum_{t=1}^n \langle \widehat{\boldsymbol{y}}_t, Z_t\rangle}{\sqrt{V_n + W_n + \left(\mathbb{E}\sqrt{V_n + W_n}\right)^2}} > u\right). \tag{20}$$

To control this probability, we recall the following results (de la Peña et al., 2008, Theorem 12.4, Corollary 12.5):

Theorem 4 ((de la Peña et al., 2008)) *For a pair of random variables* A, B*, with* $B > 0$*, such that*

$$\mathbb{E}\exp\left\{\lambda A - \lambda^2 B^2/2\right\} \leq 1 \qquad \forall \lambda \in \mathbb{R}, \tag{21}$$

it holds that for any $u > 0$*,*

$$P\left(\frac{|A|}{\sqrt{B^2 + (\mathbb{E}B)^2}} > u\right) \leq \sqrt{2}\exp\left\{-u^2/4\right\}$$

2. A martingale difference sequence $Z_1, \ldots, Z_n$ is conditionally symmetric if the law $\mathcal{L}(Z_t \mid Z_1, \ldots, Z_{t-1}) = \mathcal{L}(-Z_t \mid Z_1, \ldots, Z_{t-1})$.

and for any $y > 0$ *and* $u \geq \sqrt{2}$,

$$P\left(\frac{|A|}{\sqrt{(B^2+y)\left(1+\frac{1}{2}\log\left(B^2/y+1\right)\right)}} \geq u\right) \leq \exp\left\{-u^2/2\right\}.$$

To apply this theorem, we verify assumption (21):

Lemma 5 *The random variables* $A = \sum_{t=1}^n \langle \widehat{\boldsymbol{y}}_t, Z_t \rangle$ *and* $B^2 = 2\sum_{t=1}^n (\|Z_t\|^2 + \mathbb{E}_{t-1}\|Z_t\|^2)$ *satisfy* (21). *Furthermore, if* Z_t*'s are conditionally symmetric, then* $A = \sum_{t=1}^n \langle \widehat{\boldsymbol{y}}_t, Z_t \rangle$ *and* $B^2 = \sum_{t=1}^n \|Z_t\|^2$ *satisfy* (21).

The simple proof of the Lemma is postponed to the Appendix. Putting together (20) with Lemma 5 and Theorem 4 concludes the proof of Theorem 3. ■

To close the loop of equivalences, we need to deduce (9) from the tail bound inequality. Let us use the first part of Theorem 3. Denote the random variable in the numerator of the fraction in (17) as Y and the denominator as a random variable U. Then (17) implies that (Y/U) is a subgaussian random variable. Hence, its second moment is bounded by a constant: $\mathbb{E}(Y/U)^2 \leq c$. However, by Cauchy-Schwartz inequality,

$$\mathbb{E}Y = \mathbb{E}\left(U \cdot \frac{Y}{U}\right) \leq (\mathbb{E}U^2)^{1/2}\left(\mathbb{E}\frac{Y^2}{U^2}\right)^{1/2} \leq \sqrt{c\mathbb{E}U^2},$$

implying

$$\mathbb{E}\left\|\sum_{t=1}^n Z_t\right\| \leq 2R_{\max}\mathbb{E}\sqrt{V_n} + 2\sqrt{c\mathbb{E}V_n}. \tag{22}$$

This almost closes the loop, except the last term in (22) has the expectation inside the square root rather than outside, and thus presents a weaker upper bound (in the sense of (8) rather than (9)). We conjecture that there is a way to prove the upper bound with the expectation outside the square root. Nonetheless, to keep the promise of closing the loop, we observe that the upper bound of (8) implies that the Banach space has martingale type 2, which implies, via (Srebro et al., 2011), existence of a strongly convex function on the dual space, and, hence, existence of a strategy that guarantees (10) with a constant C that may depend at most logarithmically on n.

2.3. Remarks

We compare our result to that of Pinelis (1994). Let $Z_1, \ldots, Z_n$ be a martingale difference sequence taking values in a separable $(2, D)$-smooth Banach space $(\mathfrak{B}, \|\cdot\|)$. Pinelis (1994) proved, through a significantly more difficult analysis, that for any $u > 0$,

$$P\left(\sup_{n\geq 1}\left\|\sum_{t=1}^n Z_t\right\| \geq \sigma u\right) \leq 2\exp\left\{-\frac{u^2}{2D^2}\right\}, \tag{23}$$

where σ is a constant satisfying $\sum_{t=1}^\infty \|Z_t\|_\infty^2 \leq \sigma^2$. In comparison to Theorem 3, this result involves a distribution-independent variation σ as a worst-case pointwise upper bound.

The reader will notice that the pathwise inequality (19) does not depend on n and the construction of $\widehat{\boldsymbol{y}}_t$ is also oblivious to this value. A simple argument then allows us to lift the real-valued Burkholder-Davis-Gundy inequality (with the constant from (Burkholder, 2002)) to the Banach space valued martingales:

Lemma 6 *With the notation of Theorem 3,*

$$\mathbb{E} \max_{s=1,\dots,n} \left\| \sum_{t=1}^{s} Z_t \right\| \le \left(2R_{\max} + \sqrt{3}\right) \mathbb{E}\sqrt{V_n} \ .$$

Remarkably, the constant in the resulting BDG inequality is, up to an additive constant, proportional to $R_{\max}$. Once again, we have not seen such results in the literature, yet they follow with ease from regret inequalities.

We also remark that Theorem 3 can be naturally extended to p-smooth Banach spaces $\mathfrak{B}$. This is accomplished in a straightforward manner by extending Lemma 2.

3. Probabilistic Inequalities and Supervised Learning

We now look beyond linear prediction and analyze supervised learning problems with side information. Here again we establish a strong connection between existence of prediction strategies, the in-expectation inequalities for martingales, and high-probability tail bounds. In contrast to Section 2, *we will not present any algorithms.* Note that the simplest example of the equivalence (for binary prediction and in the absence of side information) was already stated in the very beginning of this paper.

3.1. Supervised learning with side information

We let $y_1, \dots, y_n \in \{\pm 1\}$ and $x_1, \dots, x_n \in \mathcal{X}$ for some abstract measurable set $\mathcal{X}$. Let $\mathcal{F}$ be a class of $[-1,1]$-valued functions on $\mathcal{X}$. Fix a cost function $\boldsymbol{\ell} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, convex in the first argument. For a given function $B : \mathcal{F} \times \mathcal{X}^n \to \mathbb{R}$, we aim to construct $\widehat{\boldsymbol{y}}_t = \widehat{\boldsymbol{y}}_t(x_1, \dots, x_t, y_1, \dots, y_{t-1}) \in [-1,1]$ such that the following *adaptive bound* holds:

$$\forall \ (x_t, y_t)_{t=1}^n, \quad \sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t) \le \inf_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \boldsymbol{\ell}(f(x_t), y_t) + B(f; x_1, \dots, x_n) \right\}. \tag{24}$$

We may view $\widehat{\boldsymbol{y}}_t$ as a prediction of the next value y_t having observed x_t and all the data thus far. In this paper, we focus on the linear loss $\boldsymbol{\ell}(a,b) = -ab$ (equivalently, absolute loss $|a-b| = 1 - ab$ when $a \in [-1,1]$ and $b \in \{\pm 1\}$) and the square loss $\boldsymbol{\ell}(a,b) = (a-b)^2$. We write (24) for the linear cost function as

$$\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} y_t f(x_t) - B(f; x_1, \dots, x_n) \right\} \le \sum_{t=1}^{n} y_t \widehat{\boldsymbol{y}}_t \tag{25}$$

while for the square loss it becomes

$$\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} 2y_t f(x_t) - f(x_t)^2 - B(f; x_1, \dots, x_n) \right\} \le \sum_{t=1}^{n} 2y_t \widehat{\boldsymbol{y}}_t - \widehat{\boldsymbol{y}}_t^2. \tag{26}$$

Given a function B and a class $\mathcal{F}$, there are two goals we may consider: (a) certify the existence of $(\widehat{\boldsymbol{y}}_t) \triangleq (\widehat{\boldsymbol{y}}_1, \dots, \widehat{\boldsymbol{y}}_n)$ satisfying the pathwise inequality (24) for all sequences $(x_t, y_t)_{t=1}^n$; or (b) give an explicit construction of $(\widehat{\boldsymbol{y}}_t)$. Both questions have been studied in the online learning literature, but the non-constructive approach will play an especially important role. Indeed, explicit constructions—such as the simple gradient descent update (2) — might not be available in more complex situations, yet it is the *existence* of $(\widehat{\boldsymbol{y}}_t)$ that yields the sought-after tail bounds.

To certify the existence of a strategy $(\widehat{\boldsymbol{y}}_t)$, consider the following object:

$$\mathcal{A}(\mathcal{F}, B) = \left\langle\!\!\left\langle \sup_{x_t} \inf_{\hat{y}_t} \max_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \ell(f(x_t), y_t) + B(f; x_1, \dots, x_n) \right\} \right\} \tag{27}$$

where the notation $\langle\!\langle \cdots \rangle\!\rangle_{t=1}^n$ stands for the repeated application of the operators (the outer operators corresponding to $t = 1$). The variable x_t ranges over $\mathcal{X}$, y_t is in the set $\{\pm 1\}$, and $\hat{y}_t$ ranges in $[-1, 1]$. It follows that

> $\mathcal{A}(\mathcal{F}, B) \le 0$ is a *necessary and sufficient condition* for the existence of $(\widehat{\boldsymbol{y}}_t)$ such that (24) holds.

Indeed, the optimal choice for $\hat{y}_1$ is made given x_1; the optimal choice for $\hat{y}_2$ is made given x_1, y_1, x_2, and so on. This choice defines the optimal strategy $(\widehat{\boldsymbol{y}}_t)$.[3] The other direction is immediate.

Suppose we can find an upper bound on $\mathcal{A}(\mathcal{F}, B)$ and then prove that this upper bound is non-positive. This would serve as a *sufficient* condition for the existence of $(\widehat{\boldsymbol{y}}_t)$. Next, we present such an upper bound for the case when the cost function is linear. More general results for convex Lipschitz cost functions can be found in (Foster et al., 2015).

3.2. Linear loss

As in the introduction, let $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$ be a sequence of independent Rademacher random variables. Let $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ and $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ be predictable processes with respect to the dyadic filtration $(\sigma(\varepsilon_1, \dots, \varepsilon_t))_{t=0}^n$, with values in $\mathcal{X}$ and $\{\pm 1\}$, respectively. In other words, $\mathbf{x}_t = \mathbf{x}_t(\varepsilon_1, \dots, \varepsilon_{t-1}) \in \mathcal{X}$ and $\mathbf{y}_t = \mathbf{y}_t(\varepsilon_1, \dots, \varepsilon_{t-1}) \in \{\pm 1\}$ for each $t = 1, \dots, n$. One can think of the collections $(\mathbf{x}_t)$ and $(\mathbf{y}_t)$ as trees labeled, respectively, by elements of $\mathcal{X}$ and $\{\pm 1\}$.

Lemma 7 *For the case of the linear cost function,*

$$\mathcal{A}(\mathcal{F}, B) = \sup_{\mathbf{x}} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \varepsilon_t f(\mathbf{x}_t) - B(f; \mathbf{x}_1, \dots, \mathbf{x}_n) \right]. \tag{28}$$

Therefore, the following are equivalent:

- *For any predictable process* $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \varepsilon_t f(\mathbf{x}_t) - B(f; \mathbf{x}_1, \dots, \mathbf{x}_n) \right] \le 0 \ , \tag{29}$$

3. If the infima are not achieved, a limiting argument can be employed.

- *There exists a strategy* $(\widehat{\boldsymbol{y}}_t)$ *such that the pathwise inequality* (25) *holds.*

Furthermore, the strategy can be assumed to satisfy

$$|\widehat{\boldsymbol{y}}_t| \le \sup_{f\in\mathcal{F}} |f(x_t)|. \tag{30}$$

The in-expectation bound of (29) is a necessary and sufficient condition for the existence of a strategy with the per-sequence bound (25). This latter bound, however, implies a high-probability statement, in the spirit of the other results in the paper. Below, we detail this amplification.

Take any $\mathcal{X}$-valued predictable process $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ with respect to the dyadic filtration. The deterministic inequality (25) applied to $x_t = \mathbf{x}_t(\varepsilon_1, \dots, \varepsilon_{t-1})$ and $y_t = \varepsilon_t$ becomes

$$\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^n \varepsilon_t f(\mathbf{x}_t) - B(f; \mathbf{x}_1, \dots, \mathbf{x}_n) \right\} \le \sum_{t=1}^n \varepsilon_t \widehat{\boldsymbol{y}}_t \tag{31}$$

for any sample path $(\varepsilon_1, \dots, \varepsilon_n)$, and thus we have the comparison of tails

$$P\left(\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^n \varepsilon_t f(\mathbf{x}_t) - B(f; \mathbf{x}_1, \dots, \mathbf{x}_n) \right\} > u \right) \le P\left(\sum_{t=1}^n \varepsilon_t \widehat{\boldsymbol{y}}_t > u \right). \tag{32}$$

Given the boundedness of the increments $\varepsilon_t \widehat{\boldsymbol{y}}_t$, the tail bounds follow immediately from the Azuma-Hoeffding's inequality or from Freedman's inequality (Freedman, 1975). More precisely, we use the fact that the martingale differences are bounded by $|\widehat{\boldsymbol{y}}_t| \le \sup_{f\in\mathcal{F}} |f(\mathbf{x}_t)|$, and conclude:

Lemma 8 *If there exists a prediction strategy* $(\widehat{\boldsymbol{y}}_t)$ *that satisfies* (25) *and* (30)*, then for any predictable process* $\mathbf{x}$*, the Azuma-Hoeffding inequality implies that*

$$P\left(\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^n \varepsilon_t f(\mathbf{x}_t) - B(f; \mathbf{x}_1, \dots, \mathbf{x}_n) \right\} > u \right) \le \exp\left(- \frac{u^2}{4 \max_\varepsilon \sum_{t=1}^n \sup_{f\in\mathcal{F}} f(\mathbf{x}_t(\varepsilon))^2} \right), \tag{33}$$

Freedman's inequality implies

$$P\left(\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^n \varepsilon_t f(\mathbf{x}_t) - B(f; \mathbf{x}_1, \dots, \mathbf{x}_n) \right\} > u, \quad \sum_{t=1}^n \sup_{f\in\mathcal{F}} f(\mathbf{x}_t)^2 \le \sigma^2 \right) \le \exp\left(- \frac{u^2}{2\sigma^2 + 2uM/3} \right), \tag{34}$$

where $M = n \cdot \sup_{f\in\mathcal{F}, \varepsilon\in\{\pm 1\}^n, t\le n} |f(\mathbf{x}_t)|$*, and we also have that for any* $\alpha > 0$*,*

$$P\left(\sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^n \varepsilon_t f(\mathbf{x}_t) - B(f; \mathbf{x}_1, \dots, \mathbf{x}_n) \right\} - \alpha \sum_{t=1}^n \sup_{f\in\mathcal{F}} f(\mathbf{x}_t)^2 > u \right) \le \exp\left(-2\alpha u\right). \tag{35}$$

In view of Lemma 7, a sufficient condition for these inequalities is that (29) *holds for all* $\mathbf{x}$*.*

Let us emphasize the conclusion of the above lemma: *the non-positivity of the expected supremum of a collection of martingales, offset by a function* B*, implies existence of a regret-minimization strategy, which implies a high-probability tail bound.* To close the loop,

we integrate out the tails, obtaining an in-expectation bound of the form (29), but possibly with a somewhat larger B function (this depends on the particular form of B).

In addition to describing the equivalence, let us capitalize on it and prove a new tail bound. The most basic B is a constant that depends on the complexity of $\mathcal{F}$, but not on f or the data. Define the worst-case sequential Rademacher averages as

$$\mathcal{R}_n(\mathcal{F}) \triangleq \sup_{\mathbf{x}} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \varepsilon_t f(\mathbf{x}_t). \tag{36}$$

Clearly, $B = \mathcal{R}_n(\mathcal{F})$ satisfies (29) and the following is immediate.

Corollary 9 *For any $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ and an $\mathcal{X}$-valued predictable process $\mathbf{x}$ with respect to the dyadic filtration,*

$$P\left(\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \varepsilon_t f(\mathbf{x}_t) > \mathcal{R}_n(\mathcal{F}) + u \right) \leq \exp\left(- \frac{u^2}{4 \max_{\varepsilon} \sum_{t=1}^{n} \sup_{f \in \mathcal{F}} f(\mathbf{x}_t(\varepsilon))^2} \right). \tag{37}$$

Superficially, (37) looks like a one-sided version of a deviation bound for classical (i.i.d.) Rademacher averages (Boucheron et al., 2013). However, sequential Rademacher averages are not Lipschitz with respect to a flip of a sign, as all of the remaining path may change after a flip. It is unclear to the authors how to prove (37) through other existing methods.

3.3. Square loss

Due to limited space, we will not state the analogue of Lemma 7 and simply outline the implication from existence of regret minimization strategies to high probability tail bounds.

As for the case of the linear loss function, take any $\mathcal{X}$-valued predictable process $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with respect to the dyadic filtration. Fix $\alpha > 0$. The deterministic inequality (26) for $x_t = \mathbf{x}_t(\varepsilon_1, \ldots, \varepsilon_{t-1})$ and $y_t = \frac{1}{\alpha}\varepsilon_t$ becomes

$$\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \left(\frac{2}{\alpha} \varepsilon_t f(\mathbf{x}_t) - f^2(\mathbf{x}_t) \right) - B(f; \mathbf{x}_1, \ldots, \mathbf{x}_n) \right\} \leq \sum_{t=1}^{n} \frac{2}{\alpha} \varepsilon_t \widehat{\boldsymbol{y}}_t - \widehat{\boldsymbol{y}}_t^2 \ . \tag{38}$$

As in the proof of (35), we obtain a tail comparison

$$\begin{aligned} &P\left(\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \left(\frac{2}{\alpha} \varepsilon_t f(\mathbf{x}_t) - f^2(\mathbf{x}_t) \right) - B(f; \mathbf{x}_1, \ldots, \mathbf{x}_n) \right\} > \frac{u}{\alpha} \right) \\ &\leq P\left(\sum_{t=1}^{n} \left(\frac{2}{\alpha} \varepsilon_t \widehat{\boldsymbol{y}}_t - \widehat{\boldsymbol{y}}_t^2 \right) > \frac{u}{\alpha} \right) \leq \exp\left\{ -\frac{\alpha u}{2} \right\} \end{aligned} \tag{39}$$

where the last inequality follows via a standard analysis of the moment generating function.

As an example, consider the Azoury-Vovk-Warmuth forecaster for linear regression (see e.g. (Cesa-Bianchi and Lugosi, 2006, Sec. 11.8)). Take the class $\mathcal{F}$ to be the class of functions $\mathcal{F} = \{x \mapsto \langle f, x \rangle : f \in B_2^d\}$, where B_2^d is the unit Euclidean ball in $\mathbb{R}^d$. Assuming $\mathcal{X} = B_2^d$, the regret bound for the forecaster is known to be

$$B(f; \mathbf{x}_1, \ldots, \mathbf{x}_n) = \|f\|^2 + Y^2 \sum_{t=1}^{n} \mathbf{x}_t^{\mathsf{T}} A_t^{-1} \mathbf{x}_t,$$

where $A_t = I + \sum_{s=1}^{t} \mathbf{x}_t \mathbf{x}_t^\top$ and $Y = \max |y_t|$. However, when $\mathcal{F}$ is indexed by the unit ball, the supremum in (39) has a closed form expression, and the overall probability inequality takes on the form

$$P\left(\left\| \sum_{t=1}^{n} \varepsilon_t \mathbf{x}_t \right\|^2_{A_n^{-1}} \geq \frac{1}{2} \sum_{t=1}^{n} \mathbf{x}_t^\top A_t^{-1} \mathbf{x}_t + u \right) \leq \exp\{-u\} . \tag{40}$$

We point out that, being functions of Rademacher random variables, $\mathbf{x}_t$'s are random variables themselves, and the terms in the above expression are dependent in a non-trivial manner.

We would like to refer the reader to the full version of this paper (Rakhlin and Sridharan, 2015) which contains further implications of the equivalence between the existence of deterministic strategies and tail bounds. In particular, the amplification allows us to prove a characterization of a notion of martingale type beyond the linear case.

4. Symmetrization: dyadic filtration is enough

In Section 3, we presented connections between deterministic regret inequalities in the supervised setting and tail bounds for dyadic martingales. One may ask whether these tail bounds can be used for more general martingales indexed by some set. The purpose of this section is to prove that statements for the dyadic filtration can be lifted to general processes via sequential symmetrization. Consider the martingale

$$M_g = \sum_{t=1}^{n} g(Z_t) - \mathbb{E}[g(Z_t)|Z_1, \dots, Z_{t-1}]$$

indexed by $g \in \mathcal{G}$. If (Z_t) is adapted to a dyadic filtration $\mathcal{A}_t = \sigma(\varepsilon_1, \dots, \varepsilon_t)$, each increment $g(Z_t) - \mathbb{E}[g(Z_t)|Z_1, \dots, Z_{t-1}]$ takes on the value

$$f_g(\mathbf{x}_t(\varepsilon_{1:t-1})) \triangleq \left(g(Z_t(\varepsilon_{1:t-1}, +1)) - g(Z_t(\varepsilon_{1:t-1}, -1))\right)/2$$

or its negation, where $\mathbf{x}_t$ is a predictable process with values in $\mathcal{Z} \times \mathcal{Z}$ and $f_g \in \mathcal{F}$ defined by $(z, z') \mapsto g(z) - g(z')$. In Section 3, we worked directly with martingales of the form $M_f = \sum_{t=1}^{n} \varepsilon_t f(\mathbf{x}_t(\varepsilon))$, indexed by an abstract class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ and an abstract $\mathcal{X}$-valued predictable process $\mathbf{x}$.

We extend the symmetrization approach of Panchenko (Panchenko, 2003) to sequential symmetrization for the case of martingales. In contrast to the more frequently-used Giné-Zinn symmetrization proof (via Chebyshev's inequality) (Giné and Zinn, 1984; Van Der Vaart and Wellner, 1996) that allows a direct tail comparison of the symmetrized and the original processes, Panchenko's approach allows for an "indirect" comparison. The following immediate extension of (Panchenko, 2003, Lemma 1) will imply that any $\exp\{-\mu(u)\}$ type tail behavior of the symmetrized process yields the same behavior for the original process.

Lemma 10 *Suppose ξ and ν are random variables and for some $\Gamma \geq 1$ and for all $u \geq 0$*

$$P(\nu \geq u) \leq \Gamma \exp\{-\mu(u)\}.$$

Let $\mu : \mathbb{R}_+ \to \mathbb{R}_+$ be an increasing differentiable function with $\mu(0) = 0$ and $\mu(\infty) = \infty$. Suppose for all $a \in \mathbb{R}$ and $\phi(x) \triangleq \mu([x-a]_+)$ it holds that $\mathbb{E}\phi(\xi) \le \mathbb{E}\phi(\nu)$. Then for any $u \ge 0$,

$$P(\xi \ge u) \le \Gamma \exp\{-\mu(u - \mu^{-1}(1))\}.$$

In particular, if $\mu(b) = cb$, we have $P(\xi \ge u) \le \Gamma \exp\{1 - cu\}$; if $\mu(b) = cb^2$, then $P(\xi \ge u) \le \Gamma \exp\{1 - cu^2/4\}$.

As in (Panchenko, 2003), the lemma will be used with ξ and ν as functions of a single sample and the double sample, respectively. The expression for the double sample will be symmetrized in order to pass to the dyadic filtration. However, unlike (Panchenko, 2003), we are dealing with a dependent sequence $Z_1, \dots, Z_n$, and the meaning ascribed to the "second sample" $Z'_1, \dots, Z'_n$ is that of a conditionally independent *tangent sequence*. That is, Z_t, Z'_t are independent and have the same distribution conditionally on $Z_1, \dots, Z_{t-1}$. Let $\mathbb{E}_{t-1}$ stand for the conditional expectation given $Z_1, \dots, Z_{t-1}$.

Corollary 11 *Let $\tilde{B} : \mathcal{G} \times \mathcal{Z}^{2n} \to \mathbb{R}$ be a function that is symmetric with respect to the swap of the i-th pair z_i, z'_i, for any $i \in [n]$:*

$$\tilde{B}(g; z_1, z'_1, \dots, z_i, z'_i, \dots, z_n, z'_n) = \tilde{B}(g; z_1, z'_1, \dots, z'_i, z_i, \dots, z_n, z'_n) \tag{41}$$

for all $g \in \mathcal{G}$. Then, under the assumptions of Lemma 10 on μ, a tail behavior

$$\forall (\mathbf{z}, \mathbf{z}'), \quad P\left(\sup_{g \in \mathcal{G}} \sum_{t=1}^{n} \varepsilon_t (g(\mathbf{z}_t) - g(\mathbf{z}'_t)) - \tilde{B}(g; (\mathbf{z}_1, \mathbf{z}'_1), \dots, (\mathbf{z}_n, \mathbf{z}'_n)) > u \right) \le \Gamma \exp\{-\mu(u)\}$$

for all $u > 0$ implies the tail bound

$$P\left(\sup_{g \in \mathcal{G}} \sum_{t=1}^{n} (g(Z_t) - \mathbb{E}_{t-1} g(Z_t)) - \mathbb{E}_{Z'_{1:n}} \tilde{B}(g; Z_1, Z'_1, \dots, Z_n, Z'_n) > u \right) \le \Gamma \exp\{-\mu(u - \mu^{-1}(1))\}$$

for any sequence of random variables $Z_1, \dots, Z_n$ and the corresponding tangent sequence $Z'_1, \dots, Z'_n$. The supremum is taken over a pair of predictable processes $\mathbf{z}, \mathbf{z}'$ with respect to the dyadic filtration. A direct comparison of the expected suprema also holds:

$$\begin{aligned} \mathbb{E} \sup_{g \in \mathcal{G}} \sum_{t=1}^{n} (g(Z_t) - \mathbb{E}_{t-1} g(Z_t)) - \mathbb{E}_{Z'_{1:n}} \tilde{B}(g; Z_1, Z'_1, \dots, Z_n, Z'_n) \\ \le \sup_{\mathbf{z}, \mathbf{z}'} \mathbb{E} \sup_{g \in \mathcal{G}} \sum_{t=1}^{n} \varepsilon_t (g(\mathbf{z}_t) - g(\mathbf{z}'_t)) - \tilde{B}(g; (\mathbf{z}_1, \mathbf{z}'_1), \dots, (\mathbf{z}_n, \mathbf{z}'_n)). \end{aligned} \tag{42}$$

We conclude that it is enough to prove tail bounds for a supremum

$$\sup_{f \in \mathcal{F}} \textstyle\sum_{t=1}^{n} \varepsilon_t f(\mathbf{x}_t) - B(f; \mathbf{x}_1, \dots, \mathbf{x}_n)$$

of a martingale with respect to the dyadic filtration, offset by a function $B(f; \mathbf{x}_1, \dots, \mathbf{x}_n)$, as done in Section 3.

Acknowledgements

Research is supported in part by the NSF under grants no. CDS&E-MSS 1521529 and 1521544.

References

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

B. Acciaio, M. Beiglbck, F. Penkner, W. Schachermayer, and J. Temme. A trajectorial interpretation of Doob's martingale inequalities. *Ann. Appl. Probab.*, 23(4):1494–1505, 08 2013. URL http://dx.doi.org/10.1214/12-AAP878.

M. Beiglböck and M. Nutz. Martingale inequalities and deterministic counterparts. *Electron. J. Probab*, 19(95):1–15, 2014.

M. Beiglböck and P. Siorpaes. Pathwise versions of the burkholder–davis–gundy inequality. *Bernoulli*, 21(1):360–373, 2015.

B. Bercu, B. Delyon, and E. Rio. Concentration inequalities for sums and martingales, 2015.

J. Borwein, A. Guirao, P. Hájek, and J. Vanderwerff. Uniformly convex functions on banach spaces. *Proceedings of the American Mathematical Society*, 137(3):1081–1091, 2009.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

D. Burkholder. The best constant in the davis inequality for the expectation of the martingale square function. *Transactions of the American Mathematical Society*, 354(1):91–105, 2002.

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

T. Cover. Behaviour of sequential predictors of binary sequences. In *Proc. 4th Prague Conf. Inform. Theory, Statistical Decision Functions, Random Processes*, 1965.

V. H de la Peña, T. L. Lai, and Q.-M. Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2008.

D. Foster, A. Rakhlin, and K. Sridharan. Adaptive online learning, 2015. In Submission.

D. A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

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

A. Gushchin. On pathwise counterparts of Doob's maximal inequalities. *Proceedings of the Steklov Institute of Mathematics*, 1(287):118–121, 2014.

D. Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *Annals of Probability*, 31(4):2068–2081, 2003.

I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

A. Rakhlin and K. Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. *arXiv preprint arXiv:1510.03925*, 2015.

A. Rakhlin and K. Sridharan. A tutorial on online supervised learning with applications to node classification in social networks. *CoRR*, abs/1608.09014, 2016. URL http://arxiv.org/abs/1608.09014.

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

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

N. Srebro, K. Sridharan, and A. Tewari. On the universality of online mirror descent. In *NIPS*, pages 2645–2653, 2011.

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

Appendix A. Proofs

Lemma 12 *The update in (2) satisfies*

$$\forall z_1, \dots, z_n \in \mathcal{B}, \quad \sum_{t=1}^{n} \langle \widehat{\boldsymbol{y}}_t - f, z_t \rangle \le \sqrt{n}.$$

Proof [**Lemma 12**] The following two-line proof is standard. By the property of a projection,

$$\|\widehat{\boldsymbol{y}}_{t+1} - f\|^2 = \left\|\mathrm{Proj}_{\mathcal{B}}(\widehat{\boldsymbol{y}}_t - n^{-1/2} z_t) - f\right\|^2 \le \left\|(\widehat{\boldsymbol{y}}_t - n^{-1/2} z_t) - f\right\|^2 \tag{43}$$

$$= \|\widehat{\boldsymbol{y}}_t - f\|^2 + \frac{1}{n}\|z_t\|^2 - 2n^{-1/2}\langle \widehat{\boldsymbol{y}}_t - f, z_t\rangle . \tag{44}$$

Rearranging,

$$2n^{-1/2}\langle \widehat{\boldsymbol{y}}_t - f, z_t\rangle \le \|\widehat{\boldsymbol{y}}_t - f\|^2 - \|\widehat{\boldsymbol{y}}_{t+1} - f\|^2 + \frac{1}{n}\|z_t\|^2 .$$

Summing over $t = 1, \dots, n$ yields the desired statement. ■

Proof [**Lemma 6**] Because of the "anytime" property of the regret bound and the strategy definition, we can write (19) as

$$\max_{s=1,\dots,n}\left\{\left\|\sum_{t=1}^{s} Z_t\right\| - \sum_{t=1}^{s}\langle \widehat{\boldsymbol{y}}_t, Z_t\rangle\right\} \le 2R_{\max}\sqrt{V_n} \tag{45}$$

simply because the right-hand side is largest for $s = n$. Sub-additivity of max implies

$$\max_{s=1,\dots,n}\left\|\sum_{t=1}^{s} Z_t\right\| - 2R_{\max}\sqrt{V_n} \le \max_{s=1,\dots,n}\sum_{t=1}^{s}\langle \widehat{\boldsymbol{y}}_t, Z_t\rangle\,. \tag{46}$$

By the Burkholder-Davis-Gundy inequality (with the constant from Burkholder (2002)),

$$\mathbb{E}\max_{s=1,\dots,n}\sum_{t=1}^{s}\langle \widehat{\boldsymbol{y}}_t, Z_t\rangle \le \sqrt{3}\mathbb{E}\left(\sum_{t=1}^{n}\langle \widehat{\boldsymbol{y}}_t, Z_t\rangle^2\right)^{1/2} \le \sqrt{3}\mathbb{E}\sqrt{V_n}\,. \tag{47}$$

■

Proof [**Lemma 2**] Let $\widehat{\boldsymbol{y}}'_{t+1}$ be the unrestricted minimum of (15). Because of the update form,

$$\forall f \in \mathcal{F},\quad \left\langle \widehat{\boldsymbol{y}}'_{t+1} - f, z_t\right\rangle \le \frac{1}{\eta_t}\left(D_{\mathcal{R}}(f,\widehat{\boldsymbol{y}}_t) - D_{\mathcal{R}}(f,\widehat{\boldsymbol{y}}_{t+1}) - D_{\mathcal{R}}(\widehat{\boldsymbol{y}}'_{t+1},\widehat{\boldsymbol{y}}_t)\right).$$

Summing over $t = 1,\dots,n$,

$$\begin{aligned}
\sum_{t=1}^{n}\langle \widehat{\boldsymbol{y}}_{t+1} - f, z_t\rangle &\le \eta_1^{-1}D_{\mathcal{R}}(f,\widehat{\boldsymbol{y}}_1) + \sum_{t=2}^{n}(\eta_t^{-1}-\eta_{t-1}^{-1})D_{\mathcal{R}}(f,\widehat{\boldsymbol{y}}_t) - \sum_{t=1}^{n}\eta_t^{-1}D_{\mathcal{R}}(\widehat{\boldsymbol{y}}'_{t+1},\widehat{\boldsymbol{y}}_t)\\
&\le \eta_1^{-1}R_{\max}^2 + \sum_{t=2}^{n}(\eta_t^{-1}-\eta_{t-1}^{-1})R_{\max}^2 - \sum_{t=1}^{n}\frac{\eta_t^{-1}}{2}\left\|\widehat{\boldsymbol{y}}'_{t+1} - \widehat{\boldsymbol{y}}_t\right\|_*^2\\
&\le R_{\max}^2\eta_n^{-1} - \sum_{t=1}^{n}\frac{\eta_t^{-1}}{2}\left\|\widehat{\boldsymbol{y}}'_{t+1} - \widehat{\boldsymbol{y}}_t\right\|_*^2,
\end{aligned}$$

where we used strong convexity of $\mathcal{R}$ and the fact that η_t is nonincreasing. Next, we write

$$\sum_{t=1}^{n}\langle \widehat{\boldsymbol{y}}_t - f, z_t\rangle = \sum_{t=1}^{n}\left\langle \widehat{\boldsymbol{y}}'_{t+1} - f, z_t\right\rangle + \sum_{t=1}^{n}\left\langle \widehat{\boldsymbol{y}}_t - \widehat{\boldsymbol{y}}'_{t+1}, z_t\right\rangle$$

and upper bound the second term by noting that

$$\left\langle \widehat{\boldsymbol{y}}_t - \widehat{\boldsymbol{y}}'_{t+1}, z_t\right\rangle \le \left\|\widehat{\boldsymbol{y}}_t - \widehat{\boldsymbol{y}}'_{t+1}\right\|_* \cdot \|z_t\| \le \frac{\eta_t^{-1}}{2}\left\|\widehat{\boldsymbol{y}}_t - \widehat{\boldsymbol{y}}'_{t+1}\right\|_*^2 + \frac{\eta_t}{2}\|z_t\|^2\,.$$

Combining the bounds,

$$\sum_{t=1}^{n}\langle \widehat{\boldsymbol{y}}_t - f, z_t\rangle \le R_{\max}^2\eta_n^{-1} + \sum_{t=1}^{n}\frac{\eta_t}{2}\|z_t\|^2\,. \tag{48}$$

Using the fact (Cesa-Bianchi and Lugosi, 2006, Lemma 11.8) that $\sum_{t=1}^n \frac{\alpha_t}{\sqrt{\sum_{s=1}^t \alpha_s}} \le 2\sqrt{\sum_{t=1}^n \alpha_t}$ for nonnegative (α_t) and the definition of η_t,

$$\sum_{t=1}^n \langle \widehat{\boldsymbol{y}}_t - f, z_t \rangle \le 2R_{\max} \sqrt{\sum_{t=1}^n \|z_t\|^2}. \tag{49}$$

■

Proof [**Lemma 5**] Let $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}_{t-1}[\cdot | Z_{1:t-1}]$ denote conditional expectation. We have

$$\begin{aligned}
&\mathbb{E}_{t-1} \exp\left\{\lambda \left\langle \widehat{\boldsymbol{y}}_t, Z_t - \mathbb{E}_{t-1} Z_t' \right\rangle - \lambda^2 (\|Z_t\|^2 + \mathbb{E}_{t-1} \left\|Z_t'\right\|^2)\right\} \\
&\le \mathbb{E}_{t-1} \exp\left\{\lambda \left\langle \widehat{\boldsymbol{y}}_t, Z_t - Z_t' \right\rangle - \lambda^2 (\|Z_t\|^2 + \left\|Z_t'\right\|^2)\right\} \\
&\le \mathbb{E}_{t-1} \mathbb{E}_\epsilon \exp\left\{\lambda\epsilon \left\langle \widehat{\boldsymbol{y}}_t, Z_t - Z_t' \right\rangle - \lambda^2 (\|Z_t\|^2 + \left\|Z_t'\right\|^2)\right\}.
\end{aligned}$$

Since exp is a convex function, the expression is

$$\begin{aligned}
&\mathbb{E}_{t-1} \mathbb{E}_\epsilon \exp\left\{\frac{1}{2}\left(2\lambda\epsilon \langle \widehat{\boldsymbol{y}}_t, Z_t \rangle - 2\lambda^2 \|Z_t\|^2\right) + \frac{1}{2}\left(2\lambda\epsilon \left\langle \widehat{\boldsymbol{y}}_t, -Z_t' \right\rangle - 2\lambda^2 \left\|Z_t'\right\|^2\right)\right\} \\
&\le \frac{1}{2}\mathbb{E}_{t-1} \mathbb{E}_\epsilon \exp\left\{2\lambda\epsilon \langle \widehat{\boldsymbol{y}}_t, Z_t \rangle - 2\lambda^2 \|Z_t\|^2\right\} + \frac{1}{2}\mathbb{E}_{t-1} \mathbb{E}_\epsilon \exp\left\{2\lambda\epsilon \left\langle \widehat{\boldsymbol{y}}_t, -Z_t' \right\rangle - 2\lambda^2 \left\|Z_t'\right\|^2\right\} \\
&= \mathbb{E}_{t-1} \mathbb{E}_\epsilon \exp\left\{2\lambda\epsilon \langle \widehat{\boldsymbol{y}}_t, Z_t \rangle - 2\lambda^2 \|Z_t\|^2\right\} \\
&\le \mathbb{E}_{t-1} \exp\left\{2\lambda^2 |\langle \widehat{\boldsymbol{y}}_t, Z_t \rangle|^2 - 2\lambda^2 \|Z_t\|^2\right\} \le 1
\end{aligned}$$

since $\|\widehat{\boldsymbol{y}}_t\|_* \le 1$. Repeating this argument for $t = n$ to $t = 1$ yields the statement.

If Z_t are conditionally symmetric, then $\langle \widehat{\boldsymbol{y}}_t, Z_t \rangle$ are also conditionally symmetric. Hence,

$$\begin{aligned}
\mathbb{E}_{t-1} \exp\left\{\lambda \langle \widehat{\boldsymbol{y}}_t, Z_t \rangle - \frac{\lambda^2}{2}\|Z_t\|^2\right\} &= \mathbb{E}_{t-1} \mathbb{E}_\epsilon \exp\left\{\lambda\epsilon \langle \widehat{\boldsymbol{y}}_t, Z_t \rangle - \frac{\lambda^2}{2}\|Z_t\|^2\right\} \\
&\le \mathbb{E}_{t-1} \exp\left\{\frac{\lambda^2}{2}|\langle \widehat{\boldsymbol{y}}_t, Z_t \rangle|^2 - \frac{\lambda^2}{2}\|Z_t\|^2\right\} \le 1.
\end{aligned}$$

■

Proof [**Lemma 7**] For binary outcomes $y \in \{\pm 1\}$ and either absolute loss or linear loss,

$$\mathcal{A}(\mathcal{F}, B) = \left\langle\!\!\left\langle \sup_{x_t} \inf_{\hat{y}_t} \max_{y_t} \right\rangle\!\!\right\rangle_{t=1}^n \left\{\sum_{t=1}^n -y_t \hat{y}_t + \sup_{f \in \mathcal{F}} \left\{\sum_{t=1}^n y_t f(x_t) - B(f; x_1, \dots, x_n)\right\}\right\},$$

where we shall restrict $\hat{y}_t$ to range over the interval $|\hat{y}_t| \le \sup_{f\in\mathcal{F}} |f(x_t)|$ and y_t in $\{\pm 1\}$. Consider the last step $t = n$. Given $x_{1:n}$, $\hat{y}_{1:n-1}$, and $y_{1:n-1}$, we solve

$$\inf_{\hat{y}_n} \max_{y_n} \{-\hat{y}_n y_n + \phi_n(x_{1:n}, y_{1:n})\} \tag{50}$$

where

$$\phi_n(x_{1:n}, y_{1:n}) \triangleq \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{n} y_t f(x_t) - B(f; x_1, \ldots, x_n) \right\}. \tag{51}$$

Since there are two possibilities for y_n, the closed form solution for $\hat{y}_n$ is given by

$$\hat{y}_n = \frac{1}{2}\left(\phi_n(x_{1:n}, y_{1:n-1}, 1) - \phi_n(x_{1:n}, y_{1:n-1}, -1)\right). \tag{52}$$

Importantly, this value satisfies $|\hat{y}_n| \le \sup_{f\in\mathcal{F}} |f(x_n)|$. With this optimal choice, (50) is equal to $\mathbb{E}_{\varepsilon_n}\phi_n(x_{1:n}, y_{1:n-1}, \varepsilon_n)$. We now include the supremum over x_n in the definition of ϕ_{n-1}

$$\phi_{n-1}(x_{1:n-1}, y_{1:n-1}) \triangleq \sup_{x_n} \mathbb{E}_{\varepsilon_n}\phi_n(x_{1:n}, y_{1:n-1}, \varepsilon_n)$$

and repeat the argument for $t = n-1$. Since all the steps are equalities,

$$\mathcal{A}(\mathcal{F}, B) = \phi_0(\varnothing) = \sup_{x_1}\mathbb{E}_{\varepsilon_1} \ldots \sup_{x_n}\mathbb{E}_{\varepsilon_n} \sup_{f\in\mathcal{F}} \left\{ \sum_{t=1}^{n} \varepsilon_t f(x_t) - B(f; x_1, \ldots, x_n) \right\},$$

which can be written as (28). ■

Proof [**Lemma 10**] We have

$$P(\xi \ge u) \le \frac{\mathbb{E}\phi(\xi)}{\phi(u)} \le \frac{\mathbb{E}\phi(\nu)}{\phi(u)} \le \frac{1}{\phi(u)}\left(\phi(0) + \int_0^\infty \phi'(x) P(\nu \ge x) dx\right).$$

Choose $a = u - \mu^{-1}(1)$, where μ^{-1} is the inverse function. If $a < 0$, the conclusion of the lemma is true since $\Gamma \ge 1$. In the case of $a \ge 0$, we have $\phi(0) = 0$. The above upper bound becomes

$$\begin{aligned} P(\xi \ge u) &\le \frac{\Gamma}{\phi(u)} \int_0^\infty \phi'(x)\exp\{-\mu(x)\}dx = \frac{\Gamma}{\phi(u)} \int_a^\infty \mu'(x)\exp\{-\mu(x)\}dx \\ &= \frac{\Gamma}{\mu(u-a)}\left[-\exp\{-\mu(x)\}\right]_a^\infty = \Gamma\exp\{-\mu(a)\} = \Gamma\exp\{-\mu(u-\mu^{-1}(1))\}. \end{aligned}$$

If $\mu(b) = cb$, we have

$$P(\xi \ge u) \le \Gamma\exp\{-c(u - 1/c)\} = \Gamma\exp\{1 - cu\}.$$

If $\mu(b) = cb^2$, we have

$$P(\xi \ge u) \le \Gamma\exp\{-c(u - 1/\sqrt{c})^2\} \le \Gamma\exp\{-cu^2/4\}$$

whenever $u \ge 2/\sqrt{c}$. If $u \le 2/\sqrt{c}$, the conclusion is valid since $\Gamma \ge 1$. ■

Proof [**Corollary 11**] Let

$$\xi(Z_1, \ldots, Z_n, Z_1', \ldots, Z_n') = \sup_g \sum_{t=1}^{n} (g(Z_t) - g(Z_t')) - \tilde{B}(g; Z_1, Z_1', \ldots, Z_n, Z_n')$$

and

$$\nu(Z_1,\dots,Z_n) = \sup_g \sum_{t=1}^n (g(Z_t) - \mathbb{E}_{t-1} g(Z_t')) - \mathbb{E}_{Z'_{1:n}} \tilde{B}(g; Z_1, Z_1', \dots, Z_n, Z_n').$$

Then for any convex $\phi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}\phi(\nu) \le \mathbb{E}\phi(\xi)$$

using convexity of the supremum. The problem is now reduced to obtaining tail bounds for

$$P\left(\sup_f \sum_{t=1}^n (g(Z_t) - g(Z_t')) - \tilde{B}(g; Z_1, Z_1', \dots, Z_n, Z_n') > u\right).$$

Write the probability as

$$\mathbb{E}\mathbf{1}\left\{\xi(Z_1,\dots,Z_n,Z_1',\dots,Z_n') > u\right\}.$$

We now proceed to replace the random variables from n backwards with a dyadic filtration. Let us start with the last index. Renaming Z_n and Z_n' we see that

$$\begin{aligned}
&\mathbb{E}\mathbf{1}\left\{\sup_g \sum_{t=1}^n (g(Z_t) - g(Z_t')) - \tilde{B}(g; Z_1, Z_1', \dots, Z_n, Z_n') > u\right\} \\
&= \mathbb{E}\mathbf{1}\left\{\sup_g \sum_{t=1}^{n-1} (g(Z_t) - g(Z_t')) + (g(Z_n') - g(Z_n)) - \tilde{B}(g; Z_1, Z_1', \dots, Z_n, Z_n') > u\right\} \\
&= \mathbb{E}\mathbb{E}_{\epsilon_n}\mathbf{1}\left\{\sup_g \sum_{t=1}^{n-1} (g(Z_t) - g(Z_t')) + \epsilon_n(g(Z_n) - g(Z_n')) - \tilde{B}(g; Z_1, Z_1', \dots, Z_n, Z_n') > u\right\} \\
&\le \mathbb{E} \sup_{z_n, z_n'} \mathbb{E}_{\epsilon_n}\mathbf{1}\left\{\sup_g \sum_{t=1}^{n-1} (g(Z_t) - g(Z_t')) + \epsilon_n(g(z_n) - g(z_n')) - \tilde{B}(g; Z_1, Z_1', \dots, Z_{n-1}, Z_{n-1}', z_n, z_n') > u\right\}.
\end{aligned}$$

Proceeding in this manner for step $n-1$ and back to $t=1$, we obtain an upper bound of

$$\begin{aligned}
&\sup_{z_1, z_1'} \mathbb{E}_{\epsilon_1} \dots \sup_{z_n, z_n'} \mathbb{E}_{\epsilon_n}\mathbf{1}\left\{\sup_g \sum_{t=1}^n \epsilon_t(g(z_t) - g(z_t')) - \tilde{B}(g; z_1, z_1', \dots, z_n, z_n') > u\right\} \\
&= \sup_{\mathbf{x}} \mathbb{E}\mathbf{1}\left\{\sup_g \sum_{t=1}^n \epsilon_t f_g(\mathbf{x}_t) - B(g; \mathbf{x}_1, \dots, \mathbf{x}_n) > u\right\}.
\end{aligned}$$

■