Proceedings of Machine Learning Research vol 65:1–49, 2017

ZigZag: A new approach to adaptive online learning

Dylan J. Foster DJFOSTER@CS.CORNELL.EDU
Cornell University

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

Karthik Sridharan SRIDHARAN@CS.CORNELL.EDU
Cornell University

Abstract

We develop a new family of algorithms for the online learning setting with regret against any data sequence bounded by the *empirical Rademacher complexity* of that sequence. To develop a general theory of when this type of adaptive regret bound is achievable we establish a connection to the theory of *decoupling inequalities* for martingales in Banach spaces. When the hypothesis class is a set of linear functions bounded in some norm, such a regret bound is achievable if and only if the norm satisfies certain decoupling inequalities for martingales. Donald Burkholder's celebrated *geometric characterization* of decoupling inequalities (Burkholder, 1984) states that such an inequality holds if and only if there exists a special function called a *Burkholder function* satisfying certain restricted concavity properties. Our online learning algorithms are efficient in terms of queries to this function.

We realize our general theory by giving new efficient and adaptive algorithms for classes including ℓ_p norms, group norms, and reproducing kernel Hilbert spaces. The empirical Rademacher complexity regret bound implies — when used in the i.i.d. setting — a *data-dependent* complexity bound for excess risk after online-to-batch conversion. To showcase the power of the empirical Rademacher complexity regret bound, we derive improved rates for a supervised learning generalization of the *online learning with low rank experts* task and for the *online matrix prediction* task.

In addition to obtaining tight data-dependent regret bounds, our algorithms enjoy improved efficiency over previous techniques based on Rademacher complexity, automatically work in the infinite horizon setting, and adapt to scale. To obtain such adaptive methods, we introduce novel machinery, and the resulting algorithms are not based on the standard tools of online convex optimization. We conclude with a number of open problems and new directions, both algorithmic and information-theoretic.

1. Introduction

In the *online supervised learning* task, a learner receives data $(x_1, y_1), \ldots, (x_n, y_n)$ in a stream. At time t they receive an instance x_t and must predict y_t given the instance and the previous observations $(x_1, y_1,) \ldots, (x_{t-1}, y_{t-1})$. The learner's prediction, denoted $\hat{y}_t$, is evaluated against y_t according to a loss function $\ell(\hat{y}_t, y_t)$; for classification this is typically a convex surrogate for the zero-one loss $\ell_{01}(\hat{y}, y) = \mathbb{1}\{\hat{y} \neq y\}$ such as the hinge loss $\ell_{\text{hinge}}(\hat{y}, y) = \max\{0, 1 - \hat{y} \cdot y\}$. The learner's overall

performance is measured in terms of their *regret* against a benchmark function class $\mathcal{F}$:

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

In the *statistical setting*, each pair (x_t, y_t) is drawn i.i.d. from some joint distribution $\mathcal{D}$. In this case, a bound on (1) is appealing because it immediately translates to an excess loss bound for the batch statistical learning setting after online-to-batch conversion. At the other extreme is the *fully adversarial* setting, where no generating assumptions on the data are made. We would like to develop methods that enjoy optimal guarantees in both worlds.

Our goal is to come up with prediction strategies that adapt to the "difficulty" of the sequence. In the statistical setting, optimal excess risk behavior has long been understood through empirical process theory and, in particular, Rademacher averages (Bartlett and Mendelson, 2003). Empirical Rademacher averages were shown to be an attractive data-dependent measure of complexity that can be used for model selection and for estimating the excess risk of empirical minimizers. The question considered in this paper is whether there exist prediction strategies such that empirical Rademacher averages control the per-sequence regret (1). As we show below, the empirical Rademacher average is the best sequence-based measure of complexity one can hope for.

Let us formally define the *empirical Rademacher complexity* of the class $\mathcal{F}$:

$$\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}) = \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t), \tag{2}$$

where the Rademacher sequence $\epsilon \in \{\pm 1\}^n$ is drawn uniformly at random and $x_{1:n} = (x_1, \ldots, x_n)$. The questions studied in this paper are:

- **When does there exist a strategy $(\hat{y}_t)$ such that**

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \le \mathbf{D}(\mathcal{F}, n) \cdot \widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}) \tag{3}$$

  for *every sequence* $x_{1:n}, y_{1:n}$?

- **What is the best constant $\mathbf{D}(\mathcal{F}, n)$?**

- **When can the strategy $(\hat{y}_t)$ be efficiently computed?**

We provide a characterization of when the bound (3) is achievable, and, furthermore, develop efficient algorithms based on a new set of techniques. The algorithms are parametrized by a certain special function that has been studied in probability theory and harmonic analysis for the last three decades. Interestingly, the function is neither convex nor concave (see Figure 1), yet it satisfies a property called "zig-zag concavity". The main message of this paper is that this special function can be used for algorithmic purposes and to answer the above questions.

We begin our analysis by showing that $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ is an "optimal" data-dependent regret bound in the following sense:

Lemma 1 (Sequence Optimality) *Let ℓ be the absolute, hinge, or linear loss and let $\mathcal{F}$ be any class of functions with value bounded by 1. Let $\mathcal{B}(x_{1:n})$ be a data-dependent regret bound for which there exists a strategy $(\hat{y}_t)$ guaranteeing*

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \le \mathcal{B}(x_{1:n}). \tag{4}$$

Then

$$\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}) \le \mathcal{B}(x_{1:n}) \quad \forall x_{1:n}.$$

The same result holds for the zero-one loss if we restrict to $\mathcal{F}$ and $(\hat{y}_t)$ with range $\{\pm 1\}$.

Lemma 1 reveals that no data-dependent regret bound can improve upon $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ beyond the factor $\mathbf{D}(\mathcal{F}, n)$. As we will soon show, the question of identifying $\mathbf{D}(\mathcal{F}, n)$ is an extremely rich one. When one restricts to linear function classes, this question is deeply tied to theory of Banach space geometry and, in particular, to martingales in Banach spaces.

In Sections 3-5 we assume that $\mathcal{F}$ is a class of linear functions indexed by a unit ball; Section 6 will concern the general case. For the linear case, we assume that x_t's lie in the unit ball of a separable Banach space $(\mathfrak{B}, \|\cdot\|)$ and

$$\mathcal{F} = \{x \mapsto \langle w, x\rangle \mid w \in \mathfrak{B}^\star, \|w\|_\star \le 1\},$$

with $\|\cdot\|_\star$ being the dual norm and $\mathfrak{B}^\star$ the dual space. We then observe that

$$\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}) = \mathop{\mathbb{E}}_{\epsilon} \sup_{\|w\|_\star \le 1} \sum_{t=1}^{n} \epsilon_t \langle w, x_t\rangle = \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|.$$

Consider the Euclidean setting, where $\mathcal{F}$ is the unit ℓ_2 ball. It is known that gradient descent with an adaptive step size yields a regret bound of order $\sqrt{\sum_{t=1}^n \|x_t\|^2}$ for any sequence. Khintchine's inequality then gives a further upper bound of order $\mathbb{E}_\epsilon \|\sum_{t=1}^n \epsilon_t x_t\|$. Hence, adaptive gradient descent answers the questions posed earlier for the specific case of linear functions indexed by Euclidean ball. This is one of the very few cases known to us where the bound of $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ was previously available.[1]

2. Background

Let $(\mathfrak{B}, \|\cdot\|)$ be a separable Banach space and $(\mathfrak{B}^\star, \|\cdot\|_\star)$ denote its dual. This paper focuses on the problem of online supervised learning described in Protocol 1. Input instances belong to some subset $\mathcal{X} \subseteq \mathfrak{B}$ and predictions $\hat{y}_t$ are real valued. Outcomes y_t's are selected from some abstract label space $\mathcal{Y}$. Throughout this paper we assume that the loss $\ell(\hat{y}, y)$ is convex and 1-Lipschitz in its first argument. We also assume that there exists some bounded domain $[-B, B]$ such that for all $y \in \mathcal{Y}$, $\exists \hat{y} \in [-B, B]$ such that the derivative with respect to the first argument $\ell'(\hat{y}, y) = 0$ (that is, minimum is achievable in the compact set). Call such a loss function *well-behaved*. We remark that this bound B never explicitly appears in our results, and its only purpose is to enable application of the Minimax Theorem, which requires compactness.

Definitions For $p \in (1, \infty)$, let $p' = p/(p-1)$ denote its conjugate, and $p^\star = \max\{p, p'\}$. An $\mathcal{X}$-valued tree $\mathbf{x}$ is a sequence of mappings $(\mathbf{x}_t)_{t=1}^n$ with $\mathbf{x}_t : \{\pm 1\}^{t-1} \to \mathcal{X}$. When $\epsilon_1, \dots, \epsilon_n$ are independent Rademacher random variables, the tree $\mathbf{x}$ is simply a predictable process with respect to the dyadic filtration. Recall that a sequence of random variables $(Z_t)_{t=1}^n$ is a *martingale* if for each t, $\mathbb{E}[Z_t \mid Z_1, \dots, Z_{t-1}] = Z_{t-1}$, and is called a *martingale difference sequence* if $\mathbb{E}[Z_t \mid Z_1, \dots, Z_{t-1}] = 0$. For a given martingale (Z_t), we let (dZ_t) denote its corresponding martingale difference sequence, i.e. $dZ_t = Z_t - Z_{t-1}$. For a matrix $X \in \mathbb{R}^{d \times d}$, let $X_{i,\cdot}$ denote the ith row and $X_{\cdot j}$ denote the jth column. We define its (p, q) group norm as $\|X\|_{p,q} = (\sum_{i \in [d]} \|X_{i,\cdot}\|_q^p)^{1/p} = \|(\|X_{i,\cdot}\|_q)_{i \in [d]}\|_p$. The

1. The other example is $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ for the ℓ_∞ ball, attained by diagonal AdaGrad (Duchi et al., 2011).

Protocol 1 Online Supervised Learning

for $t = 1, \dots, n$ **do**
  Nature provides $x_t \in \mathcal{X}$.
  Learner selects randomized strategy $q_t \in \Delta(\mathbb{R})$.
  Nature provides outcome $y_t \in \mathcal{Y}$.
  Learner draws $\hat{y}_t \sim q_t$ and incurs loss $\ell(\hat{y}_t, y_t)$.
end for

Schatten p-norm is defined as $\|X\|_{S_p} = \mathbf{Tr}((XX^\dagger)^{\frac{p}{2}})^{\frac{1}{p}}$. We let $\|X\|_\sigma$ denote the spectral norm (Schatten S_∞), $\|X\|_\Sigma$ denote the trace norm (Schatten S_1), and $\|X\|_F$ denote the Frobenius norm (Schatten S_2). For a set $\mathcal{A} \subseteq \mathbb{R}^d$, assumed to be symmetric, the atomic norm with respect to $\mathcal{A}$ is given by $\|x\|_{\mathcal{A}} = \min\{\alpha \mid x \in \alpha \cdot \mathrm{conv}(\mathcal{A})\}$.

3. Deriving algorithms: Adaptive relaxations and zig-zag concavity

Let us propose a simple schema for designing algorithms to achieve (3). It will turn out that considering this scheme naturally leads to us to decoupling inequalities for Banach space-valued martingales via a deep result of Burkholder (1984). We begin by observing that by convexity of the loss function,

$$\ell(\hat{y}_t, y_t) - \ell(\langle w, x_t\rangle, y_t) \le \ell'(\hat{y}_t, y_t) \cdot (\hat{y}_t - \langle w, x_t\rangle) \tag{5}$$

and hence, denoting the derivative by $\ell'_t = \ell'(\hat{y}_t, y_t)$,

$$\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{\|w\|_\star \le 1} \sum_{t=1}^n \ell(\langle w, x_t\rangle, y_t) \le \sum_{t=1}^n \hat{y}_t \cdot \ell'_t + \left\| \sum_{t=1}^n \ell'_t x_t \right\|. \tag{6}$$

Rather than aiming for the adaptive bound of empirical Rademacher averages in (3), we shall aim for $\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}, \ell'_{1:n}) = \mathbb{E}_\epsilon \|\sum_{t=1}^n \epsilon_t \ell'_t x_t\|$, a quantity that is always tighter than $\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}) = \mathbb{E}_\epsilon \|\sum_{t=1}^n \epsilon_t x_t\|$ because ℓ is 1-Lipschitz.

Foster et al. (2015) proposed a general framework called *adaptive relaxations* for deriving algorithms to achieve data-dependent regret bounds. Adaptive relaxations are a compact tool for reasoning about minimax strategies on a round-by-round basis.

Definition 2 *An admissible relaxation* $\mathbf{Rel} : \bigcup_{t=0}^n \mathcal{X}^t \times [-1,1]^t \to \mathbb{R}$ *satisfies the initial condition*

$$\mathbf{Rel}(x_{1:n}, \ell'_{1:n}) \ge \left\| \sum_{t=1}^n \ell'_t x_t \right\| - \mathbf{D} \cdot \mathop{\mathbb{E}}_\epsilon \left\| \sum_{t=1}^n \epsilon_t \ell'_t x_t \right\|, \tag{7}$$

and the recursive condition

$$\mathbf{Rel}(x_{1:t-1}, \ell'_{1:t-1}) \ge \sup_{x_t \in \mathcal{X}} \inf_{\hat{y}_t} \sup_{\ell'_t \in [-1,1]} \left[\hat{y}_t \cdot \ell'_t + \mathbf{Rel}(x_{1:t}, \ell'_{1:t}) \right].^2 \tag{8}$$

2. In original game, $\ell'_t = \ell'(\hat{y}_t, y_t)$. We have moved to an upper bound by allowing the adversary to choose ℓ'_t arbitrarily.

Proposition 3 *Suppose* $\mathbf{Rel}$ *is an admissible relaxation. If at each time* t *the learner plays the strategy*

$$\hat{y}_t = \arg\min_{\hat{y}} \sup_{\ell'_t \in [-1,1]} \left[\hat{y} \cdot \ell'_t + \mathbf{Rel}(x_{1:t}, \ell'_{1:t})\right], \tag{9}$$

regret is bounded as

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \le \mathbf{D} \cdot \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \ell'_t x_t \right\| + \mathbf{Rel}(\varnothing).$$

The takeaway from Proposition 3 is that if we can design an adaptive relaxation for which the end value $\mathbf{Rel}(\varnothing)$ is not too large, we will have succeeded in achieving the upper bound of empirical Rademacher complexity[3]. But how should we find such a relaxation? Let us try the simplest possible choice:

$$\mathbf{Rel}(x_{1:t}, \ell'_{1:t}) = \left\| \sum_{s=1}^{t} \ell'_s x_s \right\| - \mathbf{D} \cdot \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{s=1}^{t} \epsilon_s \ell'_s x_s \right\|.$$

This relaxation clearly satisfies the initial condition, but it is not so clear how to demonstrate the recursive condition. The challenge in analyzing this relaxation is that the function $z \mapsto \|A + z\| - \mathbf{D}\|B + \epsilon z\|$ is neither convex nor concave. Virtually all potential functions used in online learning are convex and the absence of such a property makes it difficult to bound the relaxation's growth under possible outcomes for the gradient ℓ'_t. Let us propose a surrogate potential with more tractable analytical properties:

Proposition 4 *Suppose there exists a function* $\mathbf{U} : \mathfrak{B} \times \mathfrak{B} \to \mathbb{R}$ *satisfying*

1. $\mathbf{U}(x, x') \ge \|x\| - \mathbf{D}\|x'\|$.
2. $\mathbf{U}$ *is* **zig-zag concave***:* $z \mapsto \mathbf{U}(x + z, x' + \epsilon z)$ *is concave for all* $x, x' \in \mathfrak{B}$ *and* $\epsilon \in \{\pm 1\}$.
3. $\mathbf{U}(0, 0) \le 0$.

Then the adaptive relaxation

$$\mathbf{Rel}(x_{1:t}, \ell'_{1:t}) = \mathop{\mathbb{E}}_{\epsilon_{1:t}} \mathbf{U}\left(\sum_{s=1}^{t} \ell'_s x_s, \sum_{s=1}^{t} \epsilon_s \ell'_s x_s \right) \tag{10}$$

is admissible.

Property 1 of $\mathbf{U}$ clearly implies that the relaxation satisfies the initial condition, and Property 3 ensures that the end value is at most 0. The zig-zag concavity property (2) is most critical, as it implies that the simple gradient-based strategy

$$\hat{y}_t = -\frac{d}{d\alpha} \mathop{\mathbb{E}}_{\epsilon_{1:t}} \mathbf{U}\left(\sum_{s=1}^{t-1} \ell'_s x_s + \alpha x_t, \sum_{s=1}^{t-1} \epsilon_s \ell'_s x_s + \epsilon_t \alpha x_t \right)\Bigg|_{\alpha=0} \tag{11}$$

achieves admissibility. We remark that this strategy is horizon-independent whenever $\mathbf{U}$ does not depend on n (which we will show is usually the case). Furthermore, one may avoid re-drawing the random signs, and, hence, the computation time is simply the evaluation of the derivative of $\mathbf{U}$.

The full description of the ZigZag algorithm is given in Section 5, but before that let us spend some time deriving such $\mathbf{U}$ functions — called the *Burkholder functions* — and connecting their existence to other properties of the Banach space.

3. We omit proof of Proposition 3 for space, but the proof of Theorem 11, the main algorithm, uses the same technique and is self-contained. See also Foster et al. (2015).

4. Zig-Zag functions, regret, and UMD spaces

What have we gained by reducing our problem to finding a $\mathbf{U}$ function? We will now show that $\mathbf{U}$ exists *if and only if* $(\mathfrak{B}, \|\cdot\|)$ is an *Unconditional Martingale Difference* (UMD) space. Informally, in a UMD space lengths of martingales are comparable to those of random walks with independent increments (see Definition 6). We call $\mathbf{U}$ a *Burkholder function* in reference to Donald Burkholder's central result characterizing UMD spaces in terms of the existence of these functions (Burkholder, 1984).

In Proposition 4 we assumed that the Burkholder function $\mathbf{U}$ satisfies $\mathbf{U}(x, x') \geq \|x\| - \mathbf{D}\|x'\|$. We will soon see that it is often easier to find an efficiently computable zig-zag concave function $\mathbf{U}_p$ that, as before, satisfies $\mathbf{U}_p(0,0) \leq 0$, but the first requirement in Proposition 4 is replaced with

$$\mathbf{U}_p(x, x') \geq \|x\|^p - \mathbf{D}_p^p \|x'\|^p$$

for some $p > 1$ (i.e. $p \neq 1$). However, the simple observation that for any number $a > 0$, $a = \frac{1}{p}\inf_{\eta>0}\{\eta a^p + (p-1)\eta^{-1/(p-1)}\}$ will allow us to algorithmically use a $\mathbf{U}_p$ function for any p to obtain the desired regret bound $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ (this is described in detail in Section 5). This motivates our complete Burkholder function definition:

Definition 5 *A function* $\mathbf{U}_p^{\mathfrak{B}} : \mathfrak{B} \times \mathfrak{B} \to \mathbb{R}$ *is* Burkholder *for* $(\|\cdot\|, p, \mathbf{D}_p)$ *if*

1. $\mathbf{U}_p^{\mathfrak{B}}(x, x') \geq \|x\|^p - \mathbf{D}_p^p \|x'\|^p$.
2. $\mathbf{U}_p^{\mathfrak{B}}$ *is* ***zig-zag concave****:* $z \mapsto \mathbf{U}_p^{\mathfrak{B}}(x + z, x' + \epsilon z)$ *is concave for all* $x, x' \in \mathfrak{B}$ *and* $\epsilon \in \{\pm 1\}$.
3. $\mathbf{U}_p^{\mathfrak{B}}(0,0) \leq 0$.[4]

For concreteness, here is a simple example for the scalar case: The function

$$\mathbf{U}_2^{\mathbb{R}}(x, x') = |x|^2 - |x'|^2$$

is Burkholder for $(|\cdot|, 2, 1)$. The reader can easily verify that this function is zig-zag concave by observing that $\mathbf{U}_2^{\mathbb{R}}(x + z, x' \pm z)$ is in fact linear in z. Perhaps the most famous $\mathbf{U}$ function is Burkholder's construction for general powers in the scalar case: For $p \in (1, \infty)$ the function

$$\mathbf{U}_p^{\mathbb{R}}(x, x') = \alpha_p\big(|x| - \beta_p |x'|\big)\big(|x| + |x'|\big)^{p-1},$$

is a $(|\cdot|, p, \beta_p)$ Burkholder function upper bounding $|x|^p - \beta_p^p |x'|^p$ for appropriate α_p, β_p.

4.1. When does a zig-zag concave $\mathbf{U}$ function exist?

It turns out that the most common Banach spaces used in machine learning settings — such as ℓ_p spaces, group norms, Schatten-p classes, and operator norms — all happen to be UMD spaces, and that each UMD space comes with its own $\mathbf{U}$ function. This leaves us with the exciting prospect of using their corresponding $\mathbf{U}$ functions to develop new adaptive online learning algorithms with improved data-dependent regret bounds. Without further ado, let us define a UMD Banach space:

4. This condition is without loss of generality.

Definition 6 *A Banach space $(\mathfrak{B}, \|\cdot\|)$ is called UMD$_p$ for some $1 < p < \infty$, if there is a constant $\mathbf{C}_p$ such that for any finite $\mathfrak{B}$-valued martingale difference sequence $(X_t)_{t=1}^n$ in $L_p(\mathfrak{B})$ and any fixed choice of signs $(\epsilon_t)_{t=1}^n$ (where each $\epsilon_t \in \{\pm 1\}$),*

$$\mathbb{E}\left\|\sum_{t=1}^n \epsilon_t X_t\right\|^p \le \mathbf{C}_p^p \mathbb{E}\left\|\sum_{t=1}^n X_t\right\|^p. \tag{12}$$

The space $(\mathfrak{B}, \|\cdot\|)$ is called UMD$_1$ *if there is a constant $\mathbf{C}_1$ such that*

$$\mathbb{E}\sup_{\tau \le n}\left\|\sum_{t=1}^{\tau} \epsilon_t X_t\right\| \le \mathbf{C}_1 \mathbb{E}\sup_{\tau \le n}\left\|\sum_{t=1}^{\tau} X_t\right\|. \tag{13}$$

Burkholder (1984) proved the following geometric characterization of UMD spaces in terms of existence of appropriate zig-zag concave $\mathbf{U}$ functions.[5]

Theorem 7 (Hytönen et al. (2016), Theorem 4.5.6) *For a Banach space $(\mathfrak{B}, \|\cdot\|)$, the following are equivalent:*

1. *$\mathfrak{B}$ is UMD$_p$ with constant $\mathbf{C}_p$.*
2. *There exists Burkholder function $\mathbf{U}_p^{\mathfrak{B}} : \mathfrak{B} \times \mathfrak{B} \mapsto \mathbb{R}$ for $(\|\cdot\|, p, \mathbf{C}_p)$.*

Theorem 7 is strengthened considerably by the following fact:

Theorem 8 *Let $p \in (1, \infty)$. If* UMD$_p$ *holds with constant $\mathbf{C}_p$, then*

- *For all $q \in (1, \infty)$,* UMD$_q$*, holds with constant $\mathbf{C}_q \le 100\left(\frac{q}{p} + \frac{q'}{p'}\right)\mathbf{C}_p$.*
- UMD$_1$ *holds with $\mathbf{C}_1 = O(\mathbf{C}_p)$.*

Furthermore, if UMD$_1$ *holds with constant $\mathbf{C}_1$, then for all $p \in (1, \infty)$ there is some constant $\mathbf{C}'_p$ for which* UMD$_p$ *holds.*

With these properties of UMD spaces established, we proceed to state our main theorem on achieving the $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ regret bound in these spaces.

Theorem 9 *Let $(\mathfrak{B}, \|\cdot\|)$ satisfy* UMD$_p$ *with constant $\mathbf{C}_p$ for any $p \in [1, \infty)$. Then there exists some randomized strategy achieving the regret bound:*

$$\mathbb{E}\left[\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \ell(f(x_t), y_t)\right] \le O\left(\mathbf{C}_p \mathbb{E}\,\mathbb{E}_{\epsilon} \sup_{\tau \le n}\left\|\sum_{t=1}^{\tau} \epsilon_t \ell'(\hat{y}_t, y_t) x_t\right\|\right) \tag{14}$$

$$\le O\left(\mathbf{C}_p \mathbb{E}\left(\mathbb{E}_{\epsilon}\left\|\sum_{t=1}^n \epsilon_t \ell'(\hat{y}_t, y_t) x_t\right\| + \max_{t \in [n]} \|x_t\| \log(n)\right)\right) \tag{15}$$

$$\le O\left(\mathbf{C}_p \mathbb{E}\left(\mathbb{E}_{\epsilon}\left\|\sum_{t=1}^n \epsilon_t x_t\right\| + \max_{t \in [n]} \|x_t\| \log(n)\right)\right). \tag{16}$$

5. Burkholder (1984) does not work with $\mathbf{U}$ functions directly but rather an equivalent property called ζ-convexity. The $\mathbf{U}$ function presentation first appeared in Burkholder (1986). See Hytönen et al. (2016) or Osekowski (2012) for a modern exposition.

This shows that a bound on $\mathbf{C}_p$ for any p gives $\mathbf{D}(\mathcal{F}, n) \le \mathbf{C}_p$ in (3), up to an extra additive $\log n$ factor[6].

An interesting feature of this theorem is that there are multiple ways through which it can be proven. In the appendix it is proven purely *non-constructively* by plugging the UMD inequality (13) into the minimax analysis framework developed in Foster et al. (2015). In Section 5 it is proven *constructively* by using the existence of the $\mathbf{U}$ function to exhibit a particular strategy for the learner.

Let us remark that the bound in (14) has the desirable property of adapting to scale, in that it does not require an a-priori upper bound on the data norms $\max_{t\in[n]}\|x_t\|$.

With Theorem 9 in mind, we finally state bounds on $\mathbf{C}_p$ for classes of interest.

Theorem 10 *The following* UMD *constants hold:*

- $(\mathbb{R}, |\cdot|)$*:* $\mathbf{C}_p = p^\star - 1 \ \forall p \in (1,\infty)$.
- $(\mathbb{R}^d, \|\cdot\|_p)$, $p \in (1,\infty)$*:* $\mathbf{C}_p = p^\star - 1$.
- $(\mathbb{R}^d, \|\cdot\|_1/\|\cdot\|_\infty)$*:* $\mathbf{C}_2 = O(\log d)$.
- $(\mathbb{R}^d, \|\cdot\|_{\mathcal{A}}/\|\cdot\|_{\mathcal{A}^\star})$*:* $\mathbf{C}_2 = O(\log|\mathcal{A}|)$.
- $(\mathbb{R}^{d\times d}, \|\cdot\|_{S_p})$, $p \in (1,\infty)$*:* $\mathbf{C}_p = O((p^\star)^2)$.
- $(\mathbb{R}^{d\times d}, \|\cdot\|_\sigma/\|\cdot\|_\Sigma)$*:* $\mathbf{C}_2 = O(\log^2 d)$.
- $(\mathbb{R}^{d\times d}, \|\cdot\|_{p,q})$, $p,q \in (1,\infty)$*:* $\mathbf{C}_p = O(p^\star q^\star)$.
- $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ *for Hilbert space* $\mathcal{H}$*:* $\mathbf{C}_2 = 1$.

4.2. Efficient Burkholder functions

Burkholder's geometric characterization, Theorem 7, implies existence of a Burkholder function $\mathbf{U}_p^{\mathfrak{B}}$ whenever a space $(\mathfrak{B}, \|\cdot\|)$ has UMD constant $\mathbf{C}_p$. Unfortunately, the generic $\mathbf{U}$ function construction (see Hytönen et al. (2016), Theorem 4.5.6) is not *efficiently computable*; it is expressed in terms of a supremum over all martingale difference sequences. However, the construction of concrete $\mathbf{U}$ functions has been an active area of research in the three decades since Burkholder's original construction. This is because one can exhibit a $\mathbf{U}$ function to certify that a space is UMD for a specific constant $\mathbf{C}_p$, and discovering *sharp* UMD constants is of general interest to the analysis community (Osekowski, 2012).

Let us begin by stating Burkholder's optimal $\mathbf{U}$ function construction for the scalar setting. This function was originally obtained by solving a particular partial differential equation. This function is graphed in Figure 1.

Example 1 ($|\cdot|^p$, Hytönen et al. (2016), Theorem 4.5.7) *For any $p \in (1,\infty)$, the function*

$$\mathbf{U}_p^{\mathbb{R}}(x,y) \triangleq \alpha_p(|x| - \beta_p|y|)(|x| + |y|)^{p-1} \tag{17}$$

is Burkholder for $(|\cdot|, p, \beta_p)$, where $\alpha_p = p\left(1 - \frac{1}{p^\star}\right)^{p-1}$, $\beta_p = p^\star - 1$. β_p is the sharpest constant possible.

Observe that all of the Burkholder function properties (Definition 5) are preserved under addition. This leads us to a construction for ℓ_p norms in the vector setting, which inherits the optimal constants from Burkholder's scalar construction.

6. All of the $\log n$ factors incurred in this paper arise when passing from bounds of the form $\mathbb{E}\sup_{\tau\le n} F_\tau$ to those of the form $\mathbb{E} F_n$ for some random process (F_t). This is notable technical issue with most martingale inequalities involving the $L_1(\mathfrak{B})$ norm, including for example Doob's well-known maximal inequality.

Example 2 (ℓ_p norm)

$$\mathbf{U}_p^{\ell_p}(x,y) \triangleq \sum_{i\in[d]} \mathbf{U}_p^{\mathbb{R}}(x_i,y_i) \tag{18}$$

is a Burkholder function for $(\|\cdot\|_p^p, p, \beta_p)$, with β_p as in Example 1. $\mathbf{U}_p^{\ell_p}$ can be computed in time $O(d)$.

Example 3 (Weighted ℓ_2 norm) *Let $\|x\|_A = \sqrt{\langle x, Ax\rangle}$ for some PSD matrix A. Then*

$$\mathbf{U}_2^{\ell_{2,A}}(x,y) \triangleq U_2^{\ell_2}(A^{1/2}x, A^{1/2}y)$$

is a Burkholder function for $(\ell_{2,A}, 2, 1)$. $\mathbf{U}_2^{\ell_{2,A}}$ can be computed in time $O(d^2)$.

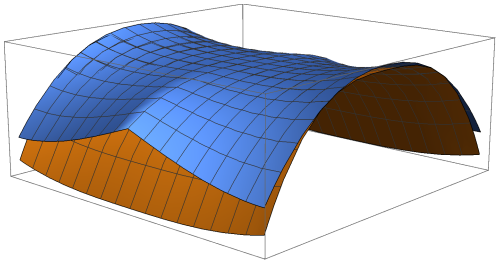

Figure 1: $\mathbf{U}_p^{\mathbb{R}}(x,x')$ (blue) and $|x|^p - \beta_p^p|x'|^p$ (orange) for $p = 3$.

Another useful construction extends Burkholder's scalar function to general Hilbert spaces. This is useful as it applies even to infinite dimensional spaces such as RKHS.

Example 4 (General Hilbert Space, Hytönen et al. (2016), Theorem 4.5.14) *Let $\mathcal{H}$ be some Hilbert space whose norm will be denoted $\|\cdot\|_{\mathcal{H}}$.*

$$\mathbf{U}_p^{\mathcal{H}}(x,y) \triangleq \alpha_p(\|x\|_{\mathcal{H}} - \beta_p\|y\|_{\mathcal{H}})(\|x\|_{\mathcal{H}} + \|y\|_{\mathcal{H}})^{p-1} \tag{19}$$

is a Burkholder function for $(\|\cdot\|_{\mathcal{H}}, p, \beta_p)$ for each $p \in (1,\infty)$, where α_p and β_p, and are as in Example 1. This function works for all Hilbert spaces, even those of infinite dimension. For $p = 2$ this function and its derivatives can be implemented efficiently using the Representer Theorem.

We can lift the former construction to a construction for group norms in the same fashion as in our construction for ℓ_p norms.

Example 5 ($(p,2)$ Group Norm) *In this example we consider group norms over matrices in $\mathbb{R}^{d\times d}$. The function,*

$$\mathbf{U}_p^{(p,2)}(x,y) \triangleq \sum_{i\in[d]} \mathbf{U}_p^{\ell_2}(x,y),$$

where $\mathbf{U}^{\ell_2,p}$ is the general Hilbert space Burkholder function (19), is a Burkholder function for $(\|\cdot\|_{(p,2)}, p, \beta_p)$. $\mathbf{U}_p^{(p,2)}$ can be computed in time $O(d^2)$.

Group norms are used in multi-task learning. Furthermore, Example 5 works not just for $\mathbb{R}^{d\times d}$, but more generally for $\mathbb{R}^d \times \mathcal{H}$ for any Hilbert space $\mathcal{H}$. This makes it well-suited to multiple kernel learning tasks.

As we will show in the sequel, there are a number of algorithmic tricks we can use to achieve $\widehat{\mathbf{Rad}}_{\mathcal{F}}$-type bounds even when we do not exactly have a $\mathbf{U}$ function for a class of interest.

5. Algorithms and applications

Recall that our goal is to design algorithms whose regret is bounded by $\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}, \ell'_{1:n}) = \mathbb{E}_\epsilon \|\sum_{t=1}^n \epsilon_t \ell'_t x_t\|$. We now present an algorithm, ZIGZAG (Algorithm 2), which efficiently achieves a regret bound of this form whenever we have an efficient Burkholder function $\mathbf{U}_p^{\mathfrak{B}}$, even if $p \neq 1$.

Algorithm 2 ZIGZAG

procedure ZIGZAG($\mathbf{U}_p, p, \eta$) ▷ $\mathbf{U}_p$ is Burkholder for $(\|\cdot\|, p, \beta)$. $\eta > 0$ is the learning rate.
  for time $t = 1, \dots, n$ **do**
    Let $G_t(\alpha) = \mathbb{E}_{\sigma_t \in \{\pm 1\}} \frac{\eta}{p} \mathbf{U}_p\big(\sum_{s=1}^{t-1} \ell'_s x_s + \alpha x_t, \sum_{t=1}^{t-1} \epsilon_s \ell'_s x_s + \sigma_t \alpha x_t\big)$.
    Predict $\hat{y}_t = -G'_t(0)$. ▷ More generally, use the supergradient.
    Draw independent Rademacher $\epsilon_t \in \{\pm 1\}$.
  end for
end procedure

Theorem 11 *Denote the prediction of Algorithm 2 as $\hat{y}_t^{\epsilon_{1:t-1}}$ to make the dependence on the sequence $(\epsilon_t)_{t\le n}$ explicit. Algorithm 2 enjoys the regret bound,*

$$\mathbb{E}_\epsilon\left[\sum_{t=1}^n \ell(\hat{y}_t^{\epsilon_{1:t-1}}, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \ell(f(x_t), y_t) - \frac{1}{p}\left(\eta\beta^p \left\|\sum_{t=1}^n \epsilon_t \ell'_t x_t\right\|^p + \frac{1}{p'-1}\eta^{-(p'-1)}\right)\right] \le 0. \tag{20}$$

A few remarks are in order. A naive application of the relaxation technique would yield a bound

$$\mathbb{E}_\epsilon\left[\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \ell(f(x_t), y_t)\right] \le \frac{1}{p}\left(\eta\beta^p \mathbb{E}_\epsilon \left\|\sum_{t=1}^n \epsilon_t \ell'_t x_t\right\|^p + \frac{1}{p'-1}\eta^{-(p'-1)}\right), \tag{21}$$

which falls short of the goal of achieving $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ for the following reason. Observe that for any $p > 1$,

$$x^{1/p} = \frac{1}{p}\inf_{\eta>0}\left(\eta x + \frac{1}{p'-1}\eta^{1-p'}\right) \triangleq \inf_{\eta>0} \Psi_{\eta,p}(x). \tag{22}$$

Recall that $\eta > 0$ is a parameter of Algorithm 2. (22) combined with (21) suggest that if we chose the optimal η in hindsight, the regret of ZIGZAG would be bounded by $\sqrt[p]{\mathbb{E}_\epsilon \|\sum_{t=1}^n \epsilon_t \ell'_t x_t\|^p}$. However, this bound is always worse than $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ via Jensen's inequality, and is indeed sub-optimal for ℓ_p norms. Luckily, (20) reveals that for ZIGZAG, the Rademacher sequence $(\epsilon_t)_{t\le n}$ used by the algorithm and the Rademacher sequence appearing in the regret bound are one and the same, which allows us to adapt η to $\|\sum_{t=1}^n \epsilon_t \ell'_t x_t\|$ for a particular playout of the sequence $(\epsilon_t)_{t\le n}$ to get the desired $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ bound. This tuning of η via doubling is stated in the next result.

Lemma 12 *Define*

$$\Phi(x_{t_1:t_2}, , \ell'_{t_1:t_2}, \epsilon_{t_1:t_2}) = \beta^p \sup_{t_1\le a\le b\le t_2} \left\|\sum_{t=a}^b \epsilon_t \ell'_t x_t\right\|^p.$$

Consider the following strategy:

1. *Choose $\eta_0 = (\beta\cdot p)^{-p}$ for $p \ge 2$ and $\eta_0 = 1$ for $p < 2$. Update with $\eta_i = 2^{-\frac{i}{p'-1}}\eta_0$.*

2. *In phase i, which consists of all $t \in \{s_i, \dots, s_{i+1} - 1\}$, play Algorithm 2,* ZIGZAG, *with learning rate η_i.*

3. *Take $s_1 = 1$, $s_{N+1} = n + 1$, and $s_{i+1} = \inf\{\tau \mid \eta_i \Phi(x_{s_i:\tau-1}, \ell'_{s_i:\tau-1}, \epsilon_{s_i:\tau-1}) > \eta_i^{-(p'-1)}\}$, where N is the index of the last phase (note that whether $t = s_{i+1}$ can be tested using only information available to the learner at time t).*

This strategy achieves

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)\right] \le O\left(\beta^2 \log^2 n \mathop{\mathbb{E}}_{\epsilon,\epsilon'}\left\|\sum_{t=1}^{n} \epsilon'_t \ell'_t x_t\right\| + \min\left\{\log n + (p\cdot\beta)^{\frac{p}{p-1}}, \beta^p \log n\right\}\right). \tag{23}$$

Remark 13 *In the above bound, (x_t) and (ℓ'_t) may adapt to the sequence (ϵ_t) drawn by the algorithm (unless the adversary is oblivious), but may not adapt to (ϵ'_t).*

5.1. ℓ_p norms

We now specialize our generic algorithm to the important special case of ℓ_p norms. We use $\mathbb{E}$ (without subscript) to denote the expectation with respect to the learner's randomization.

Example 6 *Fix $p \in (1, \infty)$. Let $\hat{y}_t$ be the strategy produced by* ZIGZAG *(Algorithm 2) using the Burkholder function $\mathbf{U}_p^{\ell_p}$ from Example 2 with the learning rate tuning strategy from Lemma 12. This strategy achieves*

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)\right] \le O\left(\mathbb{E}\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right\|_p \cdot (p^\star)^2 \log^2 n + (p^\star)^2 \log n\right). \tag{24}$$

This algorithm serves as a generalization of AdaGrad to all powers of p. If we take $p = 2$, the result recovers the regret bound for full matrix AdaGrad (Duchi et al., 2011) up to logarithmic factors:

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)\right] \le \widetilde{O}\left(\mathbb{E}\sqrt{\sum_{t=1}^{n} \|x_t\|_2^2}\right). \tag{25}$$

We can also recover the regret bound for diagonal AdaGrad (Duchi et al., 2011) by taking $p = 1 + 1/\log d$:

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)\right] \le \widetilde{O}\left(\mathbb{E}\sum_{i\in[d]} \|x_{1:n,i}\|_2\right). \tag{26}$$

Here $x_{1:n,i}$ denotes the ith row of the data matrix $(x_1, x_2, \dots, x_n) \in \mathbb{R}^{d\times n}$

There is also a direct construction of a $\mathbf{U}$ function for ℓ_1 due to Osekowski (2016), which is stated in the appendix as Example 9. Using this function we will achieve (26), but without having to use the learning rate tuning strategy, and with only $O(\log d)$ factors in the regret bound instead of $O(\log^2 d)$.

6. Beyond linear function classes: Necessary and sufficient conditions

The aim of our paper is to analyze conditions for the existence of adaptive methods that enjoy per-sequence empirical Rademacher complexity as the regret bound. In this quest, we introduced the UMD property as a necessary condition. In the present section, we consider arbitrary, possibly nonlinear function classes $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ and show that a closely related "probabilistic" UMD property offers both a necessary *and* sufficient condition.

For this section we restrict ourselves to absolute loss $\ell_{\mathrm{abs}}(\hat{y}, y) = |\hat{y} - y|$ and assume that $\mathcal{Y} = [-1,1]$.

Theorem 14 *Let $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$ be any class of predictors. The following statements are equivalent:*

1. *There exists a learning algorithm and constant B such that the following regret bound against any adversary holds:*

$$\sum_{t=1}^{n} \ell_{\mathrm{abs}}(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell_{\mathrm{abs}}(f(x_t), y_t) \le B \mathop{\mathbb{E}}_{\epsilon} \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) + b.$$

2. *For any $\mathcal{X}$ valued tree $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ where each $\mathbf{x}_t : \{\pm 1\}^{t-1} \to \mathcal{X}$, there exists constant C such that*

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1}))\right] \le C \mathop{\mathbb{E}}_{\epsilon,\epsilon'}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1}))\right] + c, \tag{27}$$

where $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ and $\epsilon' = (\epsilon'_1, \dots, \epsilon'_n)$ are independent Rademacher random variables.

Moreover, $B = \Theta(C)$ and $b = \Theta(c)$. More generally 2 implies 1 for any loss ℓ that is 1-Lipschitz and well-behaved as in Section 2, for any choice of $\mathcal{Y}$.

6.1. Function classes with the generalized UMD property

We now give examples of function classes that satisfy the generalized UMD inequality (27).

Example 7 (Kernel Classes) *Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space with kernel K such that $\sup_{x\in\mathcal{X}} \sqrt{K(x,x)} \le B$, and let $\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \le 1\}$. Then there are constants K_1, K_2 such that the generalized UMD property (27) holds with*

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1})) \le K_1 \mathop{\mathbb{E}}_{\epsilon,\epsilon'} \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1})) + K_2 B \log(n).$$

The next example is that of homogenous polynomial classes under an injective tensor norm. The full description of this setting is deferred to Appendix A.

Example 8 (Homogeneous Polynomials) *Consider homogeneous polynomials of degree $2k$, with coefficients under the unit ball of the norm $(\|\cdot\|_{\{1,\dots,k\},\{k+1,\dots,2k\}})_\star$ in $(\mathbb{R}^d)^{\otimes 2k}$. Then there exist constants K_1, K_2 such that the generalized UMD property (27) holds with*

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1})) \le K_1 k^2 \log^2(d) \mathop{\mathbb{E}}_{\epsilon,\epsilon'} \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1})) + K_2 k^2 \log^2(d) \log(n).$$

6.2. Necessary versus sufficient conditions

When we take $\mathcal{F}$ to be the unit ball of the dual norm $\|\cdot\|_\star$ as in previous sections, the inequality in (27) becomes:

$$\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t\mathbf{x}_t(\epsilon_{1:t-1})\right\| \le C\mathop{\mathbb{E}}_{\epsilon,\epsilon'}\left\|\sum_{t=1}^{n}\epsilon'_t\epsilon_t\mathbf{x}_t(\epsilon_{1:t-1})\right\|. \tag{28}$$

This condition is sometimes referred to as a *probabilistic one-sided UMD inequality* for Paley-Walsh martingales (Hytönen et al., 2016). Comparing the condition to the UMD_1 inequality (13) one observes three differences: The Rademacher sequence ϵ' is drawn uniformly rather than being fixed, we only consider Paley-Walsh martingales (trees), and there is no supremum over end times. The supremum in (13) does not present a significant difference, as it can be removed from UMD_1 at a multiplicative cost of $O(\log n)$. The randomization over ϵ' is more interesting. It turns out that if in addition to (28) we require the opposite direction of this inequality to hold, i.e.

$$\mathop{\mathbb{E}}_{\epsilon,\epsilon'}\left\|\sum_{t=1}^{n}\epsilon'_t\epsilon_t\mathbf{x}_t(\epsilon_{1:t-1})\right\| \le C'\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t\mathbf{x}_t(\epsilon_{1:t-1})\right\|,$$

then this is equivalent to the full UMD property (13) up to the presence of the supremum (Hytönen et al., 2016, Theorem 4.2.5). Thus, (28) can be thought of as a *one-sided* version of the UMD inequality.

There are indeed classes for which one-sided UMD inequality holds but the full UMD property does not. A result due to Hitczenko (1994) shows that there is a mild separation between these conditions even in the scalar setting:[7]

Theorem 15 (Hitczenko (1994)) *There exists a constant K independent of p such that for all $p \in [1,\infty)$,*

$$\mathop{\mathbb{E}}_{\epsilon}\left|\sum_{t=1}^{n}\epsilon_t\mathbf{x}_t(\epsilon_{1:t-1})\right|^p \le K^p\mathop{\mathbb{E}}_{\epsilon,\epsilon'}\left|\sum_{t=1}^{n}\epsilon'_t\epsilon_t\mathbf{x}_t(\epsilon_{1:t-1})\right|^p. \tag{29}$$

When $p = 1$ this result is exactly the generalized UMD inequality (27), and for $p > 1$ it gives a one-sided version of the UMD_p condition. This bound is quantitatively stronger than what one would obtain from the UMD_p property, since (Burkholder, 1984) shows that the full two-sided UMD_p condition requires $K \ge p^\star - 1$. In the next section we show that the stronger constants in the one-sided inequality (29) can be used to obtain improved rates for the low-rank experts setting of Hazan et al. (2016) The full UMD_p inequality would not be sufficient for this task due to its larger constant. However, we remark that the gap here is only in logarithmic factors, and that the separation between the one-sided and full UMD properties is very mild for all examples we are aware of.

6.3. Application: Low-rank experts

In this section we consider a supervised learning generalization of the problem of online learning with low-rank experts (Hazan et al., 2016). Within Protocol 1, we take $\mathcal{X} = \left\{x \in \mathbb{R}^d \mid \|x\|_\infty \le 1\right\}$ and take our set of predictors to be the simplex: $\mathcal{F} = \{x \mapsto \langle w, x\rangle \mid w \in \Delta_d\}$. We let $\mathcal{Y} = [-1,+1]$ and take ℓ to be any well-behaved loss.

The challenge stated in (Hazan et al., 2016) is to develop algorithms for this setting whose regret scales not with the dimension d (as in the standard experts bound of $O(\sqrt{n\log d})$), but rather scales

7. See also Hitczenko (1993); Cox and Veraar (2007, 2011).

with the rank of the observed data matrix $X_{1:n} = (x_1 \mid \ldots \mid x_n) \in \mathbb{R}^{d\times n}$. Hazan et al. (2016) gave an algorithm obtaining regret $O(\sqrt{n}\cdot\mathrm{rank}(X_{1:n}))$ and showed a lower bound of $\Omega(\sqrt{n\cdot\mathrm{rank}(X_{1:n})})$. Note that these bounds differ by a factor of $\sqrt{\mathrm{rank}(X_{1:n})}$; improving this gap was stated in (Hazan et al., 2016) as Open Problem (1). Using Hitczenko's decoupling inequality, this gap can be closed for the supervised setting.

Theorem 16 *For the supervised experts setting, there exists a strategy $(\hat{y}_t)$ that attains*

$$\sum_{t=1}^{n}\ell(\hat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t) \le O\Big(\sqrt{n\cdot\mathrm{rank}(X_{1:n})}\Big) + O(\log n\log d). \tag{30}$$

This bound matches the lower bound given in (Hazan et al., 2016) up to a low-order additive $\log d$ term. The result has two main ingredients: First, using Hitczenko's inequality, we show that there exists an algorithm whose regret is bounded by a quantity that closely approximates the empirical Rademacher complexity $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ for the class $\mathcal{F}$. Then, following Hazan et al. (2016), we show that the empirical Rademacher complexity of $\mathcal{F}$ on a sequence $x_{1:n}$ can be bounded as $O(\sqrt{n\cdot\mathrm{rank}(X_{1:n})})$.

Our approach also yields improved rates in terms of *approximate rank* of the matrix $X_{1:n}$, which was stated as Open Problem (3) in (Hazan et al., 2016). Define the γ-approximate rank of X via $\mathrm{rank}_\gamma(X) = \min\{\mathrm{rank}(X') \mid \|X - X'\|_\infty \le \gamma, \|X'\|_\infty \le 1\}$.

Theorem 17 *There exists a strategy $(\hat{y}_t)$ that for all $\gamma > 0$ attains*

$$\sum_{t=1}^{n}\ell(\hat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t) \le O\Big(\sqrt{n\cdot\mathrm{rank}_\gamma(X_{1:n})} + \gamma\sqrt{n\log d}\Big) + O(\log n\log d). \tag{31}$$

Furthermore, the strategy is the same as that of Theorem 16.

A bound matching (31) up to log factors was given in (Hazan et al., 2016), but only for the stochastic setting.

Lastly, we give improved rates for Open Problem (2) of (Hazan et al., 2016), which asks for experts bounds that only depend on the max norm of $X_{1:n}$. Recall that

$$\|X\|_{\max} = \min_{U\in\mathbb{R}^{d\times d}, V\in\mathbb{R}^{n\times d}, X=UV^\dagger} \|U\|_{\infty,2}\|V\|_{\infty,2},$$

where $\|\cdot\|_{\infty,2}$ denotes the group norm.

Theorem 18 *There exists a strategy $(\hat{y}_t)$ that attains*

$$\sum_{t=1}^{n}\ell(\hat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t) \le O\Big(\sqrt{n}\cdot\|X_{1:n}\|_{\max}\Big) + O(\log n\log d). \tag{32}$$

Furthermore, the strategy is the same as that of Theorem 16 and Theorem 17.

For Theorem 16, Theorem 17, and Theorem 18, the key idea is to (almost) achieve the empirical Rademacher complexity in the *online setting*, then apply bounds that had previously been used in the *statistical setting* to get tight data-dependent bounds. Since all of of these theorems are derived as upper bounds on the empirical Rademacher complexity, they are actually achieved simultaneously by

a single algorithm, and this algorithm needs no knowledge of the rank, approximate rank parameter γ, or max norm a-priori.

While our bounds depend on the ambient dimension d, they do so only weakly, through an additive $\log d$ term that does not depend on, for example, $\sqrt{n}$. Therefore, they improve on (Hazan et al., 2016) as long as the dimension d is at most exponential in $\sqrt{n}$.

It is important to note that the new bounds we have stated do not immediately transfer to the online linear optimization setting considered in (Hazan et al., 2016) due to the condition on the loss ℓ. Rather, they act as supervised analogues to the results in that paper. We do not yet have an efficient algorithm that obtains (30) because we do not have an efficient $\mathbf{U}$ function analogue for the one-sided UMD inequality.

6.4. Application: Online matrix prediction

We are not yet aware of a construction for an efficient Burkholder function for matrix classes such as the spectral norm, trace norm, and more generally the Schatten p-norm ball. Nonetheless, the UMD constants for these classes (given by Theorem 10) imply the existence of algorithms with new tradeoffs for online matrix prediction, which we highlight below.

In the online matrix prediction setting (Hazan et al., 2012) one takes $\mathcal{X} = [d] \times [d]$ and the hypothesis class $\mathcal{F}$ to be a set of $d \times d$ matrices. Writing $x_t = (i_t, j_t)$ for the t'th input instance, we let $F(x_t) = F[i_t, j_t]$ denote the (i_t, j_t)'th entry of the matrix. Suppose one wishes to compete with a class $\mathcal{G}$ of low rank — for concreteness, rank r — matrices with entry magnitudes bounded by 1. A standard approach to developing efficient algorithms for this setting is to take $\mathcal{F}$ to be a convex relaxation of $\mathcal{G}$:

$$\mathcal{F} = \left\{ F \in \mathbb{R}^{d \times d} \mid \|F\|_\Sigma \le r\sqrt{d} \right\}.$$

Then $\mathcal{G} \subseteq \mathcal{F}$, but the worst-case sample complexity of $\mathcal{F}$ is larger than that of $\mathcal{G}$. We show the existence of an algorithm for competing with $\mathcal{F}$ whose regret matches that of $\mathcal{G}$ when the data (x_t) is favorable, matches the optimal worst-case behavior of $\mathcal{F}$ for unfavorable data, and more generally interpolates between these regimes.

Let ℓ be any convex 1-Lipschitz loss and $\mathcal{Y} = [-1, +1]$. Let $N_{\text{row}} = \max_i |\{t \mid i_t = i\}|$ and $N_{\text{col}} = \max_j |\{t \mid j_t = j\}|$; these are the maximum number of times an entry appears in a given row or column, respectively.

Theorem 19 *There exists a strategy $(\hat{y}_t)$ that achieves the following regret bound:*

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{F \in \mathcal{F}} \sum_{t=1}^{n} \ell(F(x_t), y_t) \le \widetilde{O}\left(\sqrt{r} \cdot d \cdot \sqrt{\max\{N_{\text{row}}, N_{\text{col}}\}} \right). \tag{33}$$

Remark 20 *Consider the average regret $\mathbf{Reg}_n/n$, which appears as an upper bound on excess risk after online-to-batch conversion.*

- *When entries are drawn from the uniform distribution, $N_{\text{col}}, N_{\text{row}} \approx n/d$, which yields*

$$\frac{\mathbf{Reg}_n}{n} \approx \sqrt{\frac{rd}{n}}.$$

This implies that the algorithm will begin to generalize after seeing a constant number of rows worth of entries, which is the best possible rate in this setting, even if one competes with $\mathcal{G}$ directly.

- *Any entry pattern satisfying $N_{\mathrm{col}}, N_{\mathrm{row}} \approx n/d$, is sufficient to obtain the optimistic $\mathbf{Reg}_{\mathrm{n}}/n \approx \sqrt{rd/n}$ rate. Remarkably, this can happen even when the entries are chosen adaptively, so long as the condition on N_{col} and N_{row} is satisfied once the game ends.*
- *In the worst case $\mathbf{Reg}_{\mathrm{n}}/n \approx \sqrt{r}d/\sqrt{n}$, which is the standard worst-case Rademacher complexity bound for the trace norm, and is obtained when the entry distribution is too "spiky".*

The i.i.d./optimistic bound of $\sqrt{rd/n}$ matches that obtained by (Foygel and Srebro, 2011, Theorem 4) for the statistical learning setting up to logarithmic factors, but the algorithm does not need to know in advance that the entries will be distributed i.i.d.

The worst-case $\sqrt{r}d/\sqrt{n}$ bound is weaker than that of Hazan et al. (2012), which obtains worst-case regret of $\mathbf{Reg}_{\mathrm{n}}/n \approx \sqrt{rd^{3/2}/n}$, because it does not fully exploit that well-behaved losses such as ℓ_{hinge} are effectively bounded (see Shamir and Shalev-Shwartz (2014) for a discussion). One can achieve the best of both worlds by using the standard multiplicative weights strategy to combine the predictions of the two algorithms. One could also combine predictions with the transductive matrix prediction algorithm proposed in Rakhlin et al. (2012), which will obtain a tighter $\sqrt{r}d^{3/2}/n$ rate if there are no repetitions in the observed entries.

6.5. Application: Empirical covering number bounds

Having developed online learning algorithms for which regret is bounded by the empirical Rademacher complexity, we are in the appealing position of being able to apply empirical process tools designed for the *statistical setting* to derive tight regret bounds for the *adversarial setting*. One particularly powerful set of tools is that of covering numbers and, in particular, chaining.

Definition 21 (Empirical Cover) *For a hypothesis class $\mathcal{F} : \mathcal{X} \to \mathbb{R}$, data sequence $x_{1:n}$, and $\alpha > 0$, a set $\mathcal{V} \subseteq \mathbb{R}^n$ is called an empirical covering with respect to ℓ_p, $p \in [1, \infty)$, if*

$$\forall f \in \mathcal{F}\ \exists v \in \mathcal{V}\ \text{ s.t. } \left(\frac{1}{n}\sum_{t=1}^{n}(f(x_t) - v_t)^p\right)^{1/p} \le \alpha. \tag{34}$$

The set $\mathcal{V}$ is a cover with respect to ℓ_∞ if $\forall f \in \mathcal{F}\ \exists v \in \mathcal{V}$ s.t. $|f(x_t) - v_t| \le \alpha\ \forall t \in [n]$.

We let the *empirical covering number* $\mathcal{N}_p(\mathcal{F}, \alpha, x_{1:n})$ denote the size of the smallest α-empirical cover for $\mathcal{F}$ on $x_{1:n}$ with respect to ℓ_p.

Because our task is simply to obtain bounds on the empirical Rademacher complexity on a particular sequence $x_{1:n}$, we can obtain regret bounds that depend on the data-dependent *empirical* covering number defined above, instead of a *worst-case* covering number. Such bounds have proved elusive in the adversarial setting, where most existing results are based on worst-case covering numbers (e.g. Rakhlin et al. (2010)). In particular, we derive two regret bounds based on the classical covering number bound (Pollard, 1990) and Dudley Entropy Integral bound (Dudley, 1967) for Rademacher complexity.

Theorem 22 (Empirical covering bound) *For any class $\mathcal{F} \subseteq [-1, +1]^{\mathcal{X}}$ satisfying the generalized* UMD *inequality (27) with constant C, there exists a strategy $(\hat{y}_t)$ that attains*

$$\sum_{t=1}^{n}\ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t), y_t) \le O\left(C \cdot \inf_{\alpha>0}\left\{\alpha n + \sqrt{\log \mathcal{N}_1(\mathcal{F}, \alpha, x_{1:n})\, n}\right\}\right). \tag{35}$$

Theorem 23 (Empirical Dudley Entropy bound) *For any class $\mathcal{F} \subseteq [-1,+1]^{\mathcal{X}}$ satisfying the generalized* UMD *inequality (27) with constant C, there exists a strategy $(\hat{y}_t)$ that attains*

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \le O\left(C \cdot \inf_{\alpha>0}\left\{\alpha \cdot n + \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \delta, x_{1:n}) n} d\delta\right\}\right). \quad (36)$$

More generally, since our upper bounds depend on the empirical Rademacher complexity conditioned on the data $x_{1:n}$, more powerful techniques — such as Talagrand's generic chaining — may be applied to derive even tighter data-dependent covering bounds than those implied by (36).

Cohen and Mannor (2017) recently obtained bounds in the online learning with expert advice setting that scale with the empirical covering number of the class $\mathcal{F} = \Delta_{\mathbb{N}}$ (the simplex on countably many experts) on the data sequence. They derive regret bounds that scale as

$$\inf_{\alpha>0}\left\{\alpha n + \mathcal{N}_\infty(\Delta_{\mathbb{N}}, \alpha, x_{1:n}) + \sqrt{\mathcal{N}_\infty(\Delta_{\mathbb{N}}, \alpha, x_{1:n}) n}\right\}.$$

This bound falls short of the covering bound (35), which enjoys *logarithmic* scaling in the covering number $\mathcal{N}$. As a corollary of our empirical Rademacher complexity regret bound, we derive a rate with the correct dependence on $\mathcal{N}$ for the supervised learning generalization of the experts setting described in the previous section.

Theorem 24 *For the supervised experts setting, there exists a strategy $(\hat{y}_t)$ that attains*

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\Delta_d} \sum_{t=1}^{n} \ell(f(x_t), y_t) \le O\left(\inf_{\alpha>0}\left\{\alpha n + \sqrt{\log \mathcal{N}_1(\Delta_d, \alpha, x_{1:n}) n}\right\}\right) + O(\log n \log d). \quad (37)$$

This bound does not apply to the countable simplex $\Delta_{\mathbb{N}}$ due to the low-order additive $\log(d)$ term, but offers an improvement on two fronts: First, it has the correct logarithmic dependence on the empirical cover, and second, it scales with the ℓ_1-cover instead of the ℓ_∞-cover. Note that one always has $\mathcal{N}_1 \le \mathcal{N}_\infty$.

The extraneous $\log(d)$ can be replaced by the worst-case data-independent covering number (i.e. $\sup_{x_{1:n}\in\mathcal{X}^n} \log \mathcal{N}_1(\Delta_{\mathbb{N}}, \alpha, x_{1:n})$), and so can apply to the countable simplex $\Delta_{\mathbb{N}}$ if $\mathcal{X}$ possesses additional structure a-priori. We leave replacing $\log(d)$ with an empirical covering number or removing it entirely as an open question.

We conclude this section by noting that one can further derive an improvement on (37) based on the data-dependent Dudley chaining.

Theorem 25 *For the supervised experts setting, there exists a strategy $(\hat{y}_t)$ that attains*

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\Delta_d} \sum_{t=1}^{n} \ell(f(x_t), y_t) \le O\left(\inf_{\alpha>0}\left\{\alpha n + \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_2(\Delta_d, \delta, x_{1:n}) n} d\delta\right\}\right) + O(\log n \log d). \quad (38)$$

7. Discussion and further directions

We considered the task of achieving regret bounded by the empirical Rademacher complexity $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ in the adversarial online learning setting. We showed that $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ satisfies a notion of *sequence*

optimality, and derived necessary and sufficient conditions under which this bound can be achieved based on a connection to decoupling inequalities for martingales, namely the *UMD property*. We leveraged Burkholder's geometric characterization of UMD spaces to derive efficient algorithms based on Burkholder/Bellman functions. Most importantly, we showed that achieving tight data-dependent regret bounds such as $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ reduces to the crisp mathematical task of exhibiting a Burkholder function with the *zig-zag concavity* property. We used this observation to give efficient algorithms for classes based on ℓ_p norms and group norms, and to derive improved rates for settings such as matrix prediction and learning with low-rank experts.

This work leaves open a plethora of new directions centered around applying the Burkholder function method in online learning and optimization.

Related work (Foster et al., 2015) was the first work to explore data-dependent regret bounds via symmetrization techniques, but focused on non-constructive results instead of developing efficient algorithms. The present work extends the algorithmic directions proposed in that paper.

General function classes Much of the existing work on adapting to data in online learning focuses on the experts setting, where of particular interest are small loss or $L^\star$-type bounds. Existing UMD results fall short in this setting because they have only been developed for the symmetric setting of the ℓ_1 ball, a superset of the probability simplex, thus leading to looser bounds. Extending our algorithmic results to non-symmetric sets like the simplex and more generally abstract function classes as in (27) is an interesting direction for future research.

Designing U functions The design of $\mathbf{U}$ functions and related objects called Bellman functions has witnessed significant research activity in areas from harmonic analysis to optimal stopping and stochastic optimal control (Osekowski, 2012; Nazarov and Treil, 1996; Nazarov et al., 2001). The applicability to our setting has been limited so far by a focus on bounds that have sharp constants and are dimension- and horizon-independent. We anticipate that designing new $\mathbf{U}$ functions from a computer science perspective — for example, exploiting that we are tolerant to logarithmic factors in most settings — will allow us to unlock the full power of these techniques for learning applications. One such example — an elementary derivation of a scalar $\mathbf{U}$ function with sub-optimal constants — is given in the appendix as Theorem 41.

Beyond UMD UMD is far from the only martingale inequality that can be certified using Burkholder functions. For example, the textbook (Osekowski, 2012) applies the Burkholder technique to inequalities all across probability, in both discrete and continuous time. We anticipate that this technique will find extensive application in and around online learning for a wide range of settings and performance measures.

Tighter rates for specific losses The $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ bound is not tight for strongly convex losses such as the square loss. *Offset rademacher complexity* techniques have been used to obtain tight worst-case rates in this case (Rakhlin and Sridharan, 2014). Developing UMD-type inequalities for the offset Rademacher complexity and more generally developing martingale inequalities to support other types of loss structure should yield new adaptive algorithms for a number of settings.

Acknowledgements

We thank Elad Hazan and Adam Osękowski for helpful discussions. D.F. is supported in part by the NDSEG fellowship. Research is supported in part by the NSF under grants no. CDS&E-MSS

1521529 and 1521544. Part of this work was performed while D.F. and K.S. were visiting the Simons Institute for the Theory of Computing and A.R. was visiting MIT.

References

Radosław Adamczak and Paweł Wolff. Concentration inequalities for non-lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162(3-4):531–586, 2015.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003. ISSN 1532-4435.

Peter L. Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Donald L. Burkholder. Boundary value problems and sharp inequalities for martingale transforms. *The Annals of Probability*, 12(3):647–702, 1984.

Donald L. Burkholder. Martingales and fourier analysis in banach spaces. In *Probability and analysis*, pages 61–108. Springer, 1986.

Alon Cohen and Shie Mannor. Online learning with many experts. *CoRR*, abs/1702.07870, 2017. URL http://arxiv.org/abs/1702.07870.

Sonja Cox and Mark Veraar. Some remarks on tangent martingale difference sequences in l1-spaces. *Electron. Comm. Probab*, 12(421-433):380, 2007.

Sonja Cox and Mark Veraar. Vector-valued decoupling and the burkholder–davis–gundy inequality. *Illinois Journal of Mathematics*, 55(1):343–375, 2011.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

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

Dylan J. Foster, Alexander Rakhlin, and Karthik Sridharan. Adaptive online learning. In *Advances in Neural Information Processing Systems*, pages 3375–3383, 2015.

Rina Foygel and Nathan Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In *24th Annual Conference on Learning Theory (COLT)*, 2011.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Elad Hazan, Satyen Kale, and Shai Shalev-Shwartz. Near-optimal algorithms for online matrix prediction. *CoRR*, abs/1204.0136, 2012. URL http://arxiv.org/abs/1204.0136.

Elad Hazan, Tomer Koren, Roi Livni, and Yishay Mansour. Online learning with low rank experts. In *29th Annual Conference on Learning Theory*, pages 1096–1114, 2016.

Paweł Hitczenko. Domination inequality for martingale transforms of a rademacher sequence. *Israel Journal of Mathematics*, 84(1-2):161–178, 1993.

Pawel Hitczenko. On a domination of sums of random variables by sums of conditionally independent ones. *The Annals of Probability*, pages 453–468, 1994.

Tuomas Hytönen, Jan van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach Spaces*, volume 1. 2016.

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21*, pages 793–800. MIT Press, 2009.

Fedor Nazarov and Sergei Treil. The hunt for a bellman function: applications to estimates for singular integral operators and to other classical problems of harmonic analysis. 1996.

Fedor Nazarov, Sergei Treil, and Alexander Volberg. Bellman function in stochastic control and harmonic analysis. In *Systems, approximation, singular integral operators, and related topics*, pages 393–423. Springer, 2001.

Adam Osekowski. Sharp martingale and semimartingale inequalities. *Monografie Matematyczne*, 72, 2012.

Adam Osekowski. On the umd constant of the space ℓ_1^N. *Colloquium Mathematicum*, 142:135–147, 2016.

Gilles Pisier. Martingales in banach spaces (in connection with type and cotype). course ihp, feb. 2–8, 2011. 2011.

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

Alexander Rakhlin and Karthik Sridharan. Statistical learning and sequential prediction, 2012. Available at http://stat.wharton.upenn.edu/~rakhlin/book_draft.pdf.

Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression. In *Conference on Learning Theory*, 2014.

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

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

Ohad Shamir and Shai Shalev-Shwartz. Matrix completion with the trace norm: learning, bounding, and transducing. *Journal of Machine Learning Research*, 15(1):3401–3423, 2014.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S Song. Operator norm inequalities between tensor unfoldings on the partition lattice. *arXiv preprint arXiv:1603.05621*, 2016.

Appendix A. Proofs

Proof [Proof of Lemma 1] Recall that $\ell_{\text{hinge}}(\hat{y}, y) = \max\{0, 1 - \hat{y} \cdot y\}$, $\ell_{\text{abs}}(\hat{y}, y) = |\hat{y} - y|$, $\ell_{\text{lin}}(\hat{y}, y) = -\hat{y} \cdot y$. Fix a sequence $x_{1:n}$, and let $y_t = \epsilon_t$ where $\epsilon \in \{\pm 1\}^n$ is a Rademacher sequence. By our hypothesis, we have

$$\mathcal{B}(x_{1:n}) \geq \mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\hat{y}_t, \epsilon_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), \epsilon_t)\right] \geq \mathop{\mathbb{E}}_{\epsilon}\left[-\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), \epsilon_t)\right].$$

For the linear loss, observe that since $\hat{y}_t$ cannot react to ϵ_t, we immediately have

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\hat{y}_t, \epsilon_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), \epsilon_t)\right] = \mathop{\mathbb{E}}_{\epsilon}\left[-\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), \epsilon_t)\right] = \widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}).$$

For the absolute and hinge losses, we will use two facts. First, since $|f(x_t)| \leq 1$, both losses satisfy $\ell(f(x_t), \epsilon_t) = 1 - f(x_t)\epsilon_t$. Second, without any assumption on the range of $\hat{y}_t$, one has $\ell(\hat{y}_t, \epsilon_t) \geq 1 - \hat{y}_t\epsilon_t$. Therefore, whenever ℓ is the absolute or hinge loss, one has

$$\begin{aligned}
\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\hat{y}_t, \epsilon_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), \epsilon_t)\right] &\geq \mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n}(1 - \hat{y}_t\epsilon_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n}(1 - f(x_t)\epsilon_t)\right] \\
&= \mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n} -\hat{y}_t\epsilon_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} -f(x_t)\epsilon_t\right] \\
&= \mathop{\mathbb{E}}_{\epsilon}\left[-\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} -\epsilon_t f(x_t)\right].
\end{aligned}$$

The above is equal to $\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n})$ as in the linear loss case, so we have shown that for each loss our hypothesis implies $\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}) \leq \mathcal{B}(x_{1:n})$. ■

Proof [Proof of Proposition 4] We stress that this proof is meant to serve as a warmup exercise. See the proof of Theorem 11 for the correctness proof for the full ZIGZAG algorithm (Algorithm 2), which is more computationally efficient and attains a stronger performance guarantee.

Recall that the relaxation is given by

$$\mathbf{Rel}(x_{1:t}, \ell'_{1:t}) = \mathop{\mathbb{E}}_{\epsilon_{1:t}} \mathbf{U}\left(\sum_{s=1}^{t} \ell'_s x_s, \sum_{s=1}^{t} \epsilon_s \ell'_s x_s\right).$$

We first show that the initial condition property is satisfied.

Initial Condition

The initial value of the online learning game is:

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - \mathbf{D} \cdot \widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}, \ell'_{1:n}).$$

Linearizing as in (6) and expanding out $\widehat{\mathbf{Rad}}_{\mathcal{F}}$, we have

$$\leq \sum_{t=1}^{n} \hat{y}_t \ell'_t + \left\| \sum_{t=1}^{n} \ell'_t x_t \right\| - \mathbf{D} \cdot \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \ell'_t x_t \right\|$$

Now use property 1 of the function $\mathbf{U}$:

$$\leq \sum_{t=1}^{n} \hat{y}_t \ell'_t + \mathop{\mathbb{E}}_{\epsilon} \mathbf{U}\left(\sum_{t=1}^{n} \ell'_t x_t, \sum_{t=1}^{n} \epsilon_t \ell'_t x_t \right).$$
$$= \sum_{t=1}^{n} \hat{y}_t \ell'_t + \mathbf{Rel}(x_{1:n}, \ell'_{1:n}).$$

This establishes the initial condition.

Admissibility Condition First, observe that we have

$$\sup_{x_t} \inf_{\hat{y}_t} \sup_{\ell'_t} \mathop{\mathbb{E}}_{\epsilon_t} \left[\hat{y}_t \ell'_t + \mathbf{Rel}(x_{1:t}, \ell'_{1:t}, \epsilon_{1:t}) \right]$$
$$= \sup_{x_t} \inf_{\hat{y}_t} \sup_{\ell'_t} \left[\hat{y}_t \ell'_t + \mathop{\mathbb{E}}_{\epsilon_{1:t}} \mathbf{U}\left(\sum_{s=1}^{t} \ell'_s x_s, \sum_{s=1}^{t} \epsilon_t \ell'_s x_s \right) \right].$$

Define a function $G_t : \mathbb{R} \to \mathbb{R}$:

$$G_t(\alpha) = \mathop{\mathbb{E}}_{\epsilon_{1:t}} \mathbf{U}\left(\sum_{s=1}^{t-1} \ell'_s x_s + \alpha x_t, \sum_{s=1}^{t-1} \epsilon_t \ell'_s x_s + \epsilon_t \alpha x_t \right).$$

Zig-zag concavity (property 2 of $\mathbf{U}$) implies that $G_t(\alpha)$ is concave in α. With this definition, the above is equal to

$$= \sup_{x_t} \inf_{\hat{y}_t} \sup_{\ell'_t} \left[\hat{y}_t \ell'_t + G_t(\ell'_t) \right].$$

Observe that the strategy prescribed in (11) is equivalent to $\hat{y}_t = -G'_t(0)$. Moving to an upper bound by replacing the infimum with this choice of $\hat{y}_t$, we have:

$$= \sup_{x_t} \sup_{\ell'_t} \left[-G'_t(0) \cdot \ell'_t + G_t(\ell'_t) \right].$$

By concavity of G_t, this is upper bounded by:

$$\leq \sup_{x_t} G_t(0)$$
$$= \mathbf{Rel}(x_{1:t-1}, \ell'_{1:t-1}, \epsilon_{1:t-1}).$$

Hence, $\mathbf{Rel}$ is an admissible relaxation, and if we play the strategy $\hat{y}_t$ in (11) we will have

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - \mathbf{D} \cdot \widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}, \ell'_{1:n})$$
$$\leq \mathbf{Rel}(x_{1:n}, \ell'_{1:n}) \leq \mathbf{Rel}(x_{1:n-1}, \ell'_{1:n-1}) \leq \ldots \leq \mathbf{Rel}(\varnothing).$$

Finally, by property 3 of $\mathbf{U}$, $\mathbf{Rel}(\varnothing) = \mathbf{U}(0,0) \leq 0$, and so the final value of the game is at most zero. This implies that the regret bound of $\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}, \ell'_{1:n})$ is achieved. ■

A.1. Proofs from Section 4

Proof [Proof of Theorem 8] For the case $p, q \in (1, \infty)$, we appeal to Theorem 34.

Now consider the case $q = 1$, and suppose UMD_p holds for $p \in (1, \infty)$ with $\mathbf{C}_p$. Then by Theorem 34, $\mathbf{C}_2 \leq 200\mathbf{C}_p$. Finally, by Theorem 35, $\mathbf{C}_1 \leq 108\mathbf{C}_2 \leq 108 \cdot 200\mathbf{C}_p$.

For the converse direction, we appeal to Pisier (2011), Remark 8.2.4. ■

Proof [Proof of Theorem 9] Fix some $C > 0$ to be chosen later. Define the minimax value for the a game where the learner's goal is to achieve the $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ regret bound:

$$\mathcal{V} = \left\langle\!\!\left\langle \sup_{x_t} \inf_{q_t \in \Delta([-B,+B])} \sup_{y_t \in [-1,+1]} \mathbb{E}_{\hat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - C \mathbb{E}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\hat{y}_t, y_t) x_t \right\| \right].$$

Here $\langle\!\langle \star \rangle\!\rangle_{t=1}^{n}$ denotes repeated application of the operator $\star$ for $t = 1, \ldots, n$. From this definition, there always exists some randomized strategy making predictions in $[-B, +B]$ whose regret is bounded by

$$C \mathbb{E}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\hat{y}_t, y_t) x_t \right\| + \mathcal{V}.$$

See Foster et al. (2015) for a more detailed discussion of this principle. We will show that for the value of C given in the theorem statement one has $\mathcal{V} \leq 0$. To begin, observe that in view of the linearization inequality (6), the minimax value $\mathcal{V}$ is bounded by

$$\left\langle\!\!\left\langle \sup_{x_t} \inf_{q_t \in \Delta([-B,+B])} \sup_{y_t \in [-1,+1]} \mathbb{E}_{\hat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell'(\hat{y}_t, y_t) \hat{y}_t + \left\| \sum_{t=1}^{n} \ell'(\hat{y}_t, y_t) x_t \right\| - C \mathbb{E}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\hat{y}_t, y_t) x_t \right\| \right].$$

Using the minimax theorem swap technique for regret analysis — see Foster et al. (2015)[8] — the last expression is equal to

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta([-1,+1])} \inf_{\hat{y}_t \in [-B,+B]} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell'(\hat{y}_t, y_t) \hat{y}_t + \left\| \sum_{t=1}^{n} \ell'(\hat{y}_t, y_t) x_t \right\| - C \mathbb{E}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\hat{y}_t, y_t) x_t \right\| \right].$$

Choose $\hat{y}_t^\star = \arg\min_f \mathbb{E}_{y_t \sim p_t}[\ell(f, y_t)]$. By the assumption on the loss, the minimizer is obtained in $[-B, B]$ and so $\mathbb{E}_{y_t \sim p_t}[\ell'(\hat{y}_t^\star, y_t)] = 0$. With this (sub)optimal choice, we obtain an upper bound of

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta([-1,+1])} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell'(\hat{y}_t^\star, y_t) \hat{y}_t^\star + \left\| \sum_{t=1}^{n} \ell'(\hat{y}_t^\star, y_t) x_t \right\| - C \mathbb{E}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\hat{y}_t^\star, y_t) x_t \right\| \right].$$

Since $\hat{y}_t^\star$ is the population minimizer, we have $\mathbb{E}_{y_t \sim p_t}[\ell'(\hat{y}_t^\star, y_t) \hat{y}_t^\star] = \mathbb{E}_{y_t \sim p_t}[\ell'(\hat{y}_t^\star, y_t)] \hat{y}_t^\star = 0$. The proceeding expression is thus equal to

$$\begin{aligned} &\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta([-1,+1])} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\left\| \sum_{t=1}^{n} \ell'(\hat{y}_t^\star, y_t) x_t \right\| - C \mathbb{E}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\hat{y}_t^\star, y_t) x_t \right\| \right] \\ &\leq \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta([-1,+1])} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \ell'(\hat{y}_t^\star, y_t) x_t \right\| - C \mathbb{E}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\hat{y}_t^\star, y_t) x_t \right\| \right]. \end{aligned}$$

8. A word of caution: we use the assumption on the loss that there exists a minimizer for every label within some bounded domain precisely so that we can now use minimax theorem restricting $\hat{y}_t$'s to be in bounded domain.

Observe that we may rewrite the above expression as

$$\sup_{\mathbf{x}} \sup_{P} \mathop{\mathbb{E}}_{y_{1:n}\sim P}\left[\sup_{\tau\le n}\left\|\sum_{t=1}^{\tau}\ell'(\hat{y}_t^\star(p_{1:t}),y_t)\mathbf{x}_t(y_{1:t-1})\right\| - C\mathop{\mathbb{E}}_{\epsilon}\sup_{\tau\le n}\left\|\sum_{t=1}^{\tau}\epsilon_t\ell'(\hat{y}_t^\star(p_{1:t}),y_t)\mathbf{x}_t(y_{1:t-1})\right\|\right],$$

where $P = (p_1,\dots,p_n)$ is a sequence of conditional distributions over $y_{1:n}$, $\mathbf{x}$ is a sequence of mappings $\mathbf{x}_t : \mathcal{Y}^{t-1} \to \mathcal{X}$, and $\hat{y}_t^\star(p_{1:t})$ is the minimizer policy described above. For any fixed choice for P and $\mathbf{x}$, we have that $(\ell'(\hat{y}_t^\star(p_{1:t}),y_t)\mathbf{x}_t(y_{1:t-1}))_{t\le n}$ is a martingale difference sequence, because the choice of $\hat{y}_t^\star$ guarantees $\mathbb{E}[\ell'(\hat{y}_t^\star(p_{1:t}),y_t)\mathbf{x}_t(y_{1:t-1}) \mid y_{1:t-1}] = 0$.

Therefore, if UMD_1 holds with constant $\mathbf{C}_1$, we have (by choosing a uniform random sign sequence in Definition 6) that for any fixed P, $\mathbf{x}$,

$$\mathbb{E}\sup_{\tau\le n}\left\|\sum_{t=1}^{\tau}\ell'(\hat{y}_t^\star(p_{1:t}),y_t)\mathbf{x}_t(y_{1:t-1})\right\| \le \mathbf{C}_1\,\mathbb{E}\mathop{\mathbb{E}}_{\epsilon}\sup_{\tau\le n}\left\|\sum_{t=1}^{\tau}\epsilon_t\ell'(\hat{y}_t^\star(p_{1:t}),y_t)\mathbf{x}_t(y_{1:t-1})\right\|.$$

This implies that the inequality holds for the supremum over P and $\mathbf{x}$, so we have

$$\mathcal{V} \le \left\langle\!\!\left\langle \sup_{x_t}\sup_{p_t\in\Delta([-1,+1])}\mathop{\mathbb{E}}_{y_t\sim p_t}\right\rangle\!\!\right\rangle_{t=1}^{n}\left[\mathbf{C}_1\mathop{\mathbb{E}}_{\epsilon}\sup_{\tau\le n}\left\|\sum_{t=1}^{\tau}\epsilon_t\ell'(\hat{y}_t^\star,y_t)x_t\right\| - C\mathop{\mathbb{E}}_{\epsilon}\sup_{\tau\le n}\left\|\sum_{t=1}^{\tau}\epsilon_t\ell'(\hat{y}_t^\star,y_t)x_t\right\|\right].$$

Thus, if we take $C \ge \mathbf{C}_1$:

$$\le 0.$$

We have established that there exists a strategy $(\hat{y}_t)$ guaranteeing

$$\mathbb{E}\left[\sum_{t=1}^{n}\ell(\hat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right] \le \mathbf{C}_1\,\mathbb{E}\mathop{\mathbb{E}}_{\epsilon}\sup_{\tau\le n}\left\|\sum_{t=1}^{\tau}\epsilon_t\ell'(\hat{y}_t,y_t)x_t\right\|$$

Treating $(\ell'(\hat{y}_t,y_t)x_t)_{t\le n}$ as a fixed sequence, we may now apply Corollary 40 to remove the supremum over end times:

$$\le 4\mathbf{C}_1\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t\ell'(\hat{y}_t,y_t)x_t\right\| + 5\mathbf{C}_1\max_{t\in[n]}\|x_t\|\log(n).$$

By the standard contraction argument for Rademacher complexity, since $|\ell'| \le 1$,

$$\le 4\mathbf{C}_1\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\| + 5\mathbf{C}_1\max_{t\in[n]}\|x_t\|\log(n).$$

Finally, recall that by Theorem 8, $\mathbf{C}_1 \le O(\mathbf{C}_p)$.

■

Proof [Proof of Theorem 10] Most of the proofs in this theorem use the following fact: If $(X_t)_{t\le n}$ is a martingale difference sequence, its restriction to a subset of coordinates is also a martingale difference sequence. This allows one to prove the deterministic UMD property (12) for complex spaces by building up from simpler spaces.

- $(\mathbb{R}, |\cdot|)$: Burkholder (1984) shows that for all $p \in (1,\infty)$, $\mathbf{C}_p = p^\star - 1$.
- $(\mathbb{R}^d, \|\cdot\|_p)$, for $p \in (1,\infty)$:

$$\mathop{\mathbb{E}}_X \left\| \sum_{t=1}^n \epsilon_t X_t \right\|_p^p = \sum_{i\in[d]} \mathop{\mathbb{E}}_X \left| \sum_{t=1}^n \epsilon_t X_t[i] \right|^p \le (p^\star - 1) \sum_{i\in[d]} \mathop{\mathbb{E}}_X \left| \sum_{t=1}^n X_t[i] \right|^p = (p^\star - 1) \mathop{\mathbb{E}}_X \left\| \sum_{t=1}^n X_t \right\|_p^p. \quad (39)$$

  The middle inequality here uses the UMD_p constant for the scalar case.

- $(\mathbb{R}^d, \|\cdot\|_p)$, for $p \in \{1,\infty\}$: We will start with ℓ_∞. Set $p = \log d$, and observe that for ℓ_p, by Theorem 34, ℓ_p has $\mathbf{C}_2 = O(\mathbf{C}_p) = O(p^\star)$ (the second bound is from the previous example). Then we have, for any sequence of signs,

$$\begin{aligned}
\mathbb{E} \left\| \sum_{t=1}^n \epsilon_t X_t \right\|_\infty^2 &\le \mathbb{E} \left\| \sum_{t=1}^n \epsilon_t X_t \right\|_p^2 \\
&\le O(p^\star)\, \mathbb{E} \left\| \sum_{t=1}^n X_t \right\|_p^2 \\
&\le O(p^\star)\, \mathbb{E} \left(d^{1/p} \left\| \sum_{t=1}^n X_t \right\|_\infty \right)^2.
\end{aligned}$$

  Since $d^{1/\log d} = O(1)$, the last expression is at most

$$O(p^\star)\, \mathbb{E} \left\| \sum_{t=1}^n X_t \right\|_\infty^2.$$

  Finally, note that $p^\star = O(\log d)$.
  The same argument works for the ℓ_1 norm using $p = 1 + 1/\log d$. Alternatively, the constant can be deduced from duality using Theorem 37. That these constants are optimal follows from Hytönen et al. (2016), Proposition 4.2.19.

- $(\mathbb{R}^d, \|\cdot\|_{\mathcal{A}}/\|\cdot\|_{\mathcal{A}^\star})$. Let us focus on $\|\cdot\|_{\mathcal{A}^\star}$. Assume $\mathcal{A} = \{a_1, \dots, a_N\}$. Observe that

$$\begin{aligned}
\|x\|_{\mathcal{A}^\star} &= \max\{\langle y, x\rangle \mid y \in \mathrm{conv}(\mathcal{A})\} \\
&= \max\left\{ \sum_{i\in[N]} \theta_i \langle a_i, x_i\rangle \mid \theta \in \Delta(N) \right\}
\end{aligned}$$

  Since we assumed $\mathcal{A}$ is symmetric:

$$\begin{aligned}
&= \left\| (\langle a_i, x_i\rangle)_{i\in[N]} \right\|_\infty \\
&= \|Ax\|_\infty, \text{ where } A \in \mathbb{R}^{N\times d} \text{ is the matrix of elements of } \mathcal{A} \text{ stacked as rows.}
\end{aligned}$$

  For any martingale difference sequence $(X_t)_{t\le n}$, $(AX_t)_{t\le n}$ is also a martingale difference. Therefore, we can deduce the UMD_2 property for $\|\cdot\|_{\mathcal{A}^\star}$ from our result for $\|\cdot\|_\infty$. The UMD_2 property for $\|\cdot\|_{\mathcal{A}}$ follows from Theorem 37.

- $(\mathbb{R}^{d\times d}, \|\cdot\|_{S_p})$, for $p \in (1,\infty)$: Hytönen et al. (2016) Theorem 5.2.10 and Proposition 5.5.5.

- $(\mathbb{R}^{d\times d}, \|\cdot\|_\sigma)$: $\mathbf{C}_2 = O(\log^2 d)$. We will build up from the Schatten p-norms in the same fashion as for the ℓ_p spaces. Let $p = \log d$. For any sequence of signs,

$$\mathbb{E}\left\|\sum_{t=1}^n \epsilon_t X_t\right\|_\sigma^2 \le \mathbb{E}\left\|\sum_{t=1}^n \epsilon_t X_t\right\|_{S_p}^2.$$

Using Theorem 34 to get $C_2 \le O((p^\star)^2)$ for S_p:

$$\le O((p^\star)^2)\,\mathbb{E}\left\|\sum_{t=1}^n X_t\right\|_{S_p}^2$$
$$\le O((p^\star)^2)\,\mathbb{E}\left(d^{1/p}\left\|\sum_{t=1}^n X_t\right\|_\sigma\right)^2.$$

Since $d^{1/\log d} = O(1)$, the preceding expression is at most

$$O((p^\star)^2)\,\mathbb{E}\left\|\sum_{t=1}^n X_t\right\|_\sigma^2.$$

Once again, $p^\star \le \log d$. The constant for $\|\cdot\|_\Sigma$ follows from Theorem 37, since the trace norm is dual to the spectral norm.

- $(\mathbb{R}^{d\times d}, \|\cdot\|_{p,q})$, for $p, q \in (1,\infty)$: For any sequence of signs, we apply the UMD property for ℓ_p row-wise:

$$\mathbb{E}\left\|\sum_{t=1}^n \epsilon_t X_t\right\|_{p,q}^p = \sum_{i\in[d]} \mathbb{E}\left\|\sum_{t=1}^n \epsilon_t (X_t)_{i\cdot}\right\|_q^p.$$

We know ℓ_q has $\mathbf{C}_q \le O(q^\star)$. By Theorem 34, this implies that $\mathbf{C}_p$ for ℓ_q has $\mathbf{C}_p \le O(p^\star \cdot q^\star)$.

$$\le O(p^\star \cdot q^\star) \sum_{i\in[d]} \mathbb{E}\left\|\sum_{t=1}^n (X_t)_{i\cdot}\right\|_q^p$$
$$= O(p^\star \cdot q^\star)\,\mathbb{E}\left\|\sum_{t=1}^n X_t\right\|_{p,q}^p.$$

- $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ for any Hilbert space $\mathcal{H}$: See Example 4.

■

A.2. Proofs from Section 5

A.2.1. PROOFS FOR ALGORITHM 2

Proof [Proof of Theorem 11] We will show that the strategy achieves the regret bound

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^n \ell(\hat{y}_t^{\epsilon_{1:t-1}}, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \ell(f(x_t), y_t) - \Psi_{\eta,p}\left(\beta^p \left\|\sum_{t=1}^n \epsilon_t \ell'(\hat{y}_t^{\epsilon_{1:t-1}}, y_t) x_t\right\|^p\right)\right] \le 0. \tag{40}$$

Our proof technique is to define a relaxation

$$\mathbf{Rel}(x_{1:t}, \ell'_{1:t}, \epsilon_{1:t}) = \frac{\eta}{p}\mathbf{U}_p\left(\sum_{s=1}^{t} \ell'_s x_s, \sum_{t=1}^{t} \epsilon_s \ell'_s x_s\right).$$

and show that the relaxation is admissible for the following game:

$$\left\langle\!\!\left\langle \sup_{x_t} \inf_{\hat{y}_t} \sup_{\ell'_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \hat{y}_t \ell'_t - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} f(x_t)\ell'_t - \Psi_{\eta,p}\left(\beta^p \left\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right\|^p\right)\right]. \tag{41}$$

This relaxation is slightly generalized compared to Definition 2 in that Rademacher sequence $(\epsilon_t)_{t\le n}$ also appears as an argument. This is essential to accomplish the coupling of the algorithm's randomness and the regret functional $\widehat{\mathbf{Rad}}_{\mathcal{F}}$.

With the game defined we can proceed to showing that the relaxation satisfies the admissibility and initial conditions, with one extra step of linearization in the initial condition.

Initial Condition In view of (6),

$$\begin{aligned}
&\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - \Psi_{\eta,p}\left(\beta^p \left\|\sum_{t=1}^{n} \epsilon_t \ell'(\hat{y}_t, y_t) x_t\right\|^p\right) \\
&\le \sum_{t=1}^{n} \hat{y}_t \ell'_t + \left\|\sum_{t=1}^{n} \ell'_t x_t\right\| - \Psi_{\eta,p}\left(\beta^p \left\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right\|^p\right) \\
&\le \sum_{t=1}^{n} \hat{y}_t \ell'_t + \Psi_{\eta,p}\left(\left\|\sum_{t=1}^{n} \ell'_t x_t\right\|^p\right) - \Psi_{\eta,p}\left(\beta^p \left\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right\|^p\right) \\
&= \sum_{t=1}^{n} \hat{y}_t \ell'_t + \frac{\eta}{p}\left(\left\|\sum_{t=1}^{n} \ell'_t x_t\right\|^p - \beta^p \left\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right\|^p\right) \\
&\le \sum_{t=1}^{n} \hat{y}_t \ell'_t + \frac{\eta}{p}\mathbf{U}_p\left(\sum_{t=1}^{n} \ell'_t x_t, \sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right) \\
&= \sum_{t=1}^{n} \hat{y}_t \ell'_t + \mathbf{Rel}(x_{1:n}, \ell'_{1:n}, \epsilon_{1:n}).
\end{aligned}$$

Admissibility Condition

$$\begin{aligned}
&\sup_{x_t} \inf_{\hat{y}_t} \sup_{\ell'_t} \mathbb{E}_{\epsilon_t}\left[\hat{y}_t \ell'_t + \mathbf{Rel}(x_{1:t}, \ell'_{1:t}, \epsilon_{1:t})\right] \\
&= \sup_{x_t} \inf_{\hat{y}_t} \sup_{\ell'_t} \mathbb{E}_{\epsilon_t}\left[\hat{y}_t \ell'_t + \frac{\eta}{p}\mathbf{U}_p\left(\sum_{s=1}^{t} \ell'_s x_s, \sum_{t=1}^{t} \epsilon_s \ell'_s x_s\right)\right] \\
&= \sup_{x_t} \inf_{\hat{y}_t} \sup_{\ell'_t} \left[\hat{y}_t \ell'_t + \mathbb{E}_{\epsilon_t} \frac{\eta}{p}\mathbf{U}_p\left(\sum_{s=1}^{t} \ell'_s x_s, \sum_{t=1}^{t} \epsilon_s \ell'_s x_s\right)\right] \\
&= \sup_{x_t} \inf_{\hat{y}_t} \sup_{\ell'_t} \left[\hat{y}_t \ell'_t + G_t(\ell'_t)\right]
\end{aligned}$$

Plugging in the strategy specified by Algorithm 2, the last expression is at most

$$\begin{aligned}
&\sup_{x_t}\sup_{\ell'_t}\left[-G'_t(0)\cdot\ell'_t + G_t(\ell'_t)\right]\\
&\le \sup_{x_t} G_t(0)\\
&= \mathbf{Rel}(x_{1:t-1}, \ell'_{1:t-1}, \epsilon_{1:t-1}).
\end{aligned}$$

Finally, since $\mathbf{U}_p$ is Burkholder we have $\mathbf{Rel}(\varnothing) \propto \mathbf{U}_p(0,0) \le 0$, and so the final value of the game is at most zero. This implies that (40) is achieved. ■

Proof [Proof of Lemma 12] In what follows we will leave the dependence of $\hat{y}_t, x_t, \ell'_t$ on $\epsilon_{1:t-1}$ implicit for notational convenience. We will handle this dependence at the end of the proof.

Assume $N > 1$. Otherwise, the algorithm's regret is bounded as $2\eta_1^{-(p'-1)} = 4\eta_0^{-(p'-1)}$.

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n}\ell(\hat{y}_t^{\epsilon_{1:t-1}}, y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t), y_t)\right] \le \mathop{\mathbb{E}}_{\epsilon}\left[\sum_{i=1}^{N}\left[\sum_{t=s_i}^{s_{i+1}-1}\ell(\hat{y}_t^{\epsilon_{1:t-1}}, y_t) - \inf_{f\in\mathcal{F}}\sum_{t=s_i}^{s_{i+1}-1}\ell(f(x_t), y_t)\right]\right]$$

Using the regret bound for Algorithm 2 (note that that algorithm has an anytime regret guarantee) given by Theorem 11:

$$\le \mathop{\mathbb{E}}_{\epsilon}\left[\frac{1}{p}\sum_{i=1}^{N}\left[\eta_i\beta_p^p\left\|\sum_{t=s_i}^{s_{i+1}-1}\epsilon_t\ell'_t x_t\right\|^p + \frac{1}{p'-1}\eta_i^{-(p'-1)}\right]\right].$$

Introducing a new supremum:

$$\le \mathop{\mathbb{E}}_{\epsilon}\left[\frac{1}{p}\sum_{i=1}^{N}\left[\eta_i\Phi(x_{s_i:s_{i+1}-1}, \ell'_{s_i:s_{i+1}-1}, \epsilon_{s_i:s_{i+1}-1}) + \frac{1}{p'-1}\eta_i^{-(p'-1)}\right]\right].$$

The doubling condition implies that $\eta_i\Phi(x_{s_i:s_{i+1}-2}, \ell'_{s_i:s_{i+1}-2}, \epsilon_{s_i:s_{i+1}-2}) \le \eta_i^{-(p'-1)}$. To use this fact, observe that since $\|x_t\| \le 1$, we have that for any $C > 0$,

$$\begin{aligned}
&\eta_i\Phi(x_{s_i:s_{i+1}-1}, \ell'_{s_i:s_{i+1}-1}, \epsilon_{s_i:s_{i+1}-1})\\
&= \eta_i\beta_p^p \sup_{s_i\le a\le b\le s_{i+1}-1}\left\|\sum_{t=a}^{b}\epsilon_t\ell'_t x_t\right\|^p\\
&\le \eta_i(1+1/C)^p\beta_p^p \sup_{s_i\le a\le b\le s_{i+1}-2}\left\|\sum_{t=a}^{b}\epsilon_t\ell'_t x_t\right\|^p + \eta_i C^p\beta_p^p.
\end{aligned}$$

For $C = p$:

$$\begin{aligned}
&\le \eta_i e\Phi(x_{s_i:s_{i+1}-2}, \epsilon_{s_i:s_{i+1}-2}) + \eta_i p^p\beta_p^p.\\
&= e\eta_i^{-(p'-1)} + \eta_i p^p\beta_p^p.
\end{aligned}$$

Returning to the regret bound, we have

$$\begin{aligned}
&\leq \mathop{\mathbb{E}}_{\epsilon}\left[\frac{1}{p}\sum_{i=1}^{N}\left[e\eta_i^{-(p'-1)} + \eta_i p^p \beta_p^p + \frac{1}{p'-1}\eta_i^{-(p'-1)}\right]\right] \\
&\leq \mathop{\mathbb{E}}_{\epsilon}\left[e\sum_{i=1}^{N}\eta_i^{-(p'-1)} + p^p\beta_p^p\eta_i\right]
\end{aligned}$$

We will handle with the left-hand term first. Now observe that

$$\eta_{N-1}\Phi(x_{s_{N-1}:s_N}, \ell'_{s_{N-1}:s_N}, \epsilon_{s_{N-1}:s_N}) > \eta_{N-1}^{-(p'-1)}.$$

Rearranging further implies

$$\eta_{N-1}^{-(p'-1)} \leq \Phi(x_{s_{N-1}:s_N}, \ell'_{s_{N-1}:s_N}, \epsilon_{s_{N-1}:s_N})^{1/p} \leq \Phi(x_{1:n}, \ell'_{1:n}, \epsilon_{1:n})^{1/p}.$$

Finally, since $\eta_i^{-(p'-1)} = 2\eta_{i-1}^{-(p'-1)}$,

$$\sum_{i=1}^{N}\eta_i^{-(p'-1)} = \eta_0^{-(p'-1)}\sum_{i=1}^{N}2^i \leq 2\cdot 2^N\eta_0^{-(p'-1)} \leq 4\Phi(x_{1:n}, \ell'_{1:n}, \epsilon_{1:n})^{1/p} = 4\beta_p \sup_{1\leq a\leq b\leq n}\left\|\sum_{t=a}^{b}\epsilon_t\ell'_t x_t\right\|.$$

For the second term, observe that $\eta_i \leq \eta_0$ for all i, so

$$\sum_{i=1}^{N}p^p\beta_p^p\eta_i \leq p^p\beta_p^p\eta_0\cdot N.$$

Finally, by the invariant $2^{N-1}\eta_0^{-(p'-1)} \leq \Phi(x_{1:n}, \epsilon_{1:n})^{1/p}$ we established earlier,

$$N \leq \log\left(\Phi(x_{1:n}, \ell'_{1:n}, \epsilon_{1:n})^{1/p}\eta_0^{(p'-1)}\right) + 1$$

Putting everything together, the regret is bounded as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\epsilon}\max\left\{2e\beta_p \sup_{1\leq a\leq b\leq n}\left\|\sum_{t=a}^{b}\epsilon_t\ell'_t x_t\right\| + p^p\beta_p^p\eta_0\left(\log\left(\sup_{1\leq a\leq b\leq n}\left\|\sum_{t=a}^{b}\epsilon_t\ell'_t x_t\right\|\eta_0^{(p'-1)}\right) + 1\right), 4\eta_0^{-(p'-1)}\right\} \\
&\leq \mathop{\mathbb{E}}_{\epsilon}\left[2e\beta_p \sup_{1\leq a\leq b\leq n}\left\|\sum_{t=a}^{b}\epsilon_t\ell'_t x_t\right\| + p^p\beta_p^p\eta_0\left(\log\left(\sup_{1\leq a\leq b\leq n}\left\|\sum_{t=a}^{b}\epsilon_t\ell'_t x_t\right\|\eta_0^{(p'-1)}\right) + 1\right) + 4\eta_0^{-(p'-1)}\right]
\end{aligned}$$

Using that $\|x_t\| \leq 1$:

$$\leq 2e\beta_p \mathop{\mathbb{E}}_{\epsilon}\sup_{1\leq a\leq b\leq n}\left\|\sum_{t=a}^{b}\epsilon_t\ell'_t x_t\right\| + p^p\beta_p^p\eta_0\log\left(n\cdot\eta_0^{(p'-1)}\right) + 4\eta_0^{-(p'-1)}.$$

For the choice $\eta_0 = (\beta_p\cdot p)^{-p}$:

$$\leq 2e\beta_p \mathop{\mathbb{E}}_{\epsilon}\sup_{1\leq a\leq b\leq n}\left\|\sum_{t=a}^{b}\epsilon_t\ell'_t x_t\right\| + \log(n) + (p\cdot\beta_p)^{\frac{p}{p-1}}.$$

For the choice $\eta_0 = 1$:

$$\leq 2e\beta_p \mathop{\mathbb{E}}_{\epsilon} \sup_{1\leq a\leq b\leq n} \left\| \sum_{t=a}^{b} \epsilon_t \ell_t' x_t \right\| + p^p \beta_p^p \log(n) + 4.$$

Writing $x_t(\epsilon_{1:t-1})$ and $\ell_t'(\epsilon_{1:t-1})$ to make the adversary's dependence on the sequence ϵ explicit, the main term of interest in the above quantity is

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{1\leq a\leq b\leq n} \left\| \sum_{t=a}^{b} \epsilon_t \ell_t'(\epsilon_{1:t-1}) x_t(\epsilon_{1:t-1}) \right\|.$$

It remains to remove the supremum and decouple the data sequences (x_t) and (ℓ_t') from the Rademacher sequence ϵ. Since $\ell_t' x_t$ can only react to $\epsilon_{1:t-1}$, the sequence $(\epsilon_t \ell_t' x_t)_{t\leq n}$ is a martingale difference sequence. Since $\left\|\sum_{t=a}^{b} \epsilon_t \ell_t' x_t\right\| \leq n$, we may apply Corollary 33 to arrive at an upper bound of

$$\leq O\left(\log(n) \mathop{\mathbb{E}}_{\epsilon} \sup_{1\leq b\leq n} \left\| \sum_{t=1}^{b} \epsilon_t \ell_t'(\epsilon_{1:t-1}) x_t(\epsilon_{1:t-1}) \right\| \right).$$

Now observe that since Algorithm 2 uses a Burkholder function $\mathbf{U}_p$ for $(\|\cdot\|, p, \beta_p)$, Theorem 7 and Theorem 8 together imply that the UMD_1 inequality (13) holds with constant $O(\beta_p)$, therefore, the above is bounded as

$$\leq O\left(\beta_p \log(n) \mathop{\mathbb{E}}_{\epsilon} \mathop{\mathbb{E}}_{\epsilon'} \sup_{1\leq b\leq n} \left\| \sum_{t=1}^{b} \epsilon_t' \ell_t'(\epsilon_{1:t-1}) x_t(\epsilon_{1:t-1}) \right\| \right).$$

Note that the variables (x_t) and (ℓ_t') no longer depend on the Rademacher sequence appearing in the sum. Lastly, we apply Corollary 33 once more to remove the remaining supremum and arrive at the bound,

$$\leq O\left(\beta_p \log^2(n) \mathop{\mathbb{E}}_{\epsilon} \mathop{\mathbb{E}}_{\epsilon'} \left\| \sum_{t=1}^{b} \epsilon_t' \ell_t'(\epsilon_{1:t-1}) x_t(\epsilon_{1:t-1}) \right\| \right).$$

■

Proof [Proof of Example 6] (24) is obtained by plugging the optimal UMD constant $p^\star - 1$ into the bound for Lemma 12. For (25), observe that for any sequence z_t we have $\mathbb{E}_\epsilon \|\sum_{t=1}^{n} \epsilon_t z_t\|_2 \leq \sqrt{\mathbb{E}_\epsilon \|\sum_{t=1}^{n} \epsilon_t z_t\|_2^2} = \sqrt{\mathbb{E}_\epsilon \sum_{t=1}^{n} \|z_t\|_2^2}$. Applying this fact with the algorithm's bound for $p = 2$ gives the regret bound

$$O\left(\sqrt{\sum_{t=1}^{n} \|\ell_t' x_t\|_2^2} \cdot \log^2 n + \log n \right).$$

For (26), observe that with $p = 1/\log d$ we have the regret bound

$$O\left(\mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \ell_t' x_t \right\|_p \cdot \log d \log^2 n + \log^2 d \log n \right).$$

However for any X, $\|X\|_p \le d^{1-1/p}\|X\|_1$. For our choice of $p = 1 + 1/\log d$ we have $d^{1-1/p} = O(1)$.

$$\begin{aligned}
&\le O\left(\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t \ell_t' x_t\right\|_1 \cdot \log d \log^2 n + \log^2 d \log n\right)\\
&\le O\left(\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_1 \cdot \log d \log^2 n + \log^2 d \log n\right)\\
&= O\left(\sum_{i\in[d]}\mathop{\mathbb{E}}_{\epsilon}\left|\sum_{t=1}^{n}\epsilon_t x_t[i]\right| \cdot \log d \log^2 n + \log^2 d \log n\right)\\
&\le O\left(\sum_{i\in[d]}\sqrt{\sum_{t=1}^{n}(x_t[i])^2} \cdot \log d \log^2 n + \log^2 d \log n\right)\\
&= O\left(\sum_{i\in[d]}\|x_{1:n,i}\|_2 \cdot \log d \log^2 n + \log^2 d \log n\right).
\end{aligned}$$

■

A.2.2. SIMPLIFIED DOUBLING TRICK

In this section we derive a variant of the doubling trick given in Lemma 12 which achieves an upper bound on $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ rather than $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ itself, but does so with improved dependence on constants and low-order terms. This strategy will be used as a subroutine in subsequent algorithms.

Lemma 26 *Suppose we have an anytime regret minimization algorithm $(\hat{y}_t)$ that guarantees a regret bound of the form*

$$\sum_{t=1}^{n}\ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t), y_t) \le \frac{1}{p}\left[\eta K^p \mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|^p + \frac{1}{p'-1}\eta^{-(p'-1)}\right],$$

where $p > 1$ is fixed and η is a parameter of the algorithm. Define

$$\Phi(x_{t_1:t_2}) = K^p \mathop{\mathbb{E}}_{\epsilon}\sup_{t_1\le a\le b\le t_2}\left\|\sum_{t=a}^{b}\epsilon_t x_t\right\|^p.$$

Consider the following strategy

1. *Choose $\eta_0 < 1$ arbitrary. Update with $\eta_i = 2^{-\frac{1}{p'-1}}\eta_{i-1}$.*
2. *In phase i, which consists of all $t \in \{s_i, \dots, s_{i+1} - 1\}$, play strategy $(\hat{y}_t)$ with learning rate η_i.*
3. *Take $s_1 = 1$, $s_{N+1} = n + 1$, and $s_{i+1} = \inf\left\{\tau \mid \eta_i\Phi(x_{s_i:\tau}) > \eta_i^{-(p'-1)}\right\}$, where N is the index of the last phase.*

This strategy achieves

$$\begin{aligned}
\sum_{t=1}^{n}\ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t), y_t) &\le K\left(\mathop{\mathbb{E}}_{\epsilon}\sup_{1\le a\le b\le n}\left\|\sum_{t=a}^{b}\epsilon_t x_t\right\|^p\right)^{1/p} + \eta_0^{-(p'-1)}\\
&\le C\cdot(p')^2\cdot K\left(\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|^p\right)^{1/p} + \eta_0^{-(p'-1)}.
\end{aligned}$$

Proof [Proof of Lemma 26] We assume $N > 1$. Otherwise, the algorithm's regret is bounded as $2\eta_1^{-(p'-1)} = 4\eta_0^{-(p'-1)}$.

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \le \sum_{i=1}^{N}\left[\sum_{t=s_i}^{s_{i+1}-1} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=s_i}^{s_{i+1}-1} \ell(f(x_t), y_t)\right]$$

Using the assumed regret bound (note that that algorithm has an anytime regret guarantee):

$$\le \frac{1}{p}\sum_{i=1}^{N}\left[\eta_i K^p \mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=s_i}^{s_{i+1}-1} \epsilon_t x_t\right\|^p + \frac{1}{p'-1}\eta_i^{-(p'-1)}\right]$$

Introducing a new supremum:

$$\le \frac{1}{p}\sum_{i=1}^{N}\left[\eta_i \Phi(x_{s_i:s_{i+1}-1}) + \frac{1}{p'-1}\eta_i^{-(p'-1)}\right]$$

Using the invariant $\eta_i\Phi(x_{s_i:s_{i+1}-1}) \le \eta_i^{-(p'-1)}$:

$$\le \frac{1}{p}\left(1 + \frac{1}{p'-1}\right)\sum_{i=1}^{N}\eta_i^{-(p'-1)}$$
$$= \sum_{i=1}^{N}\eta_i^{-(p'-1)}$$

Observe that $\eta_{N-1}\Phi(x_{s_{N-1}:s_N}) > \eta_{N-1}^{-(p'-1)}$, and so rearrangingimplies

$$\eta_{N-1}^{-(p'-1)} \le \Phi(x_{s_{N-1}:s_N})^{1/p} \le \Phi(x_{1:n})^{1/p}.$$

Finally, we can check that $\eta_i^{-(p'-1)} = 2\eta_{i-1}^{-(p'-1)}$, so $2^N\eta_0^{-(p'-1)} \le \Phi(x_{1:n})^{1/p}$. Now,

$$\sum_{i=1}^{N}\eta_i^{-(p'-1)} = \eta_0^{-(p'-1)}\sum_{i=1}^{N}2^i \le 2\cdot 2^N\eta_0^{-(p'-1)} \le \Phi(x_{1:n})^{1/p} = K\left(\mathop{\mathbb{E}}_{\epsilon}\sup_{1\le a\le b\le n}\left\|\sum_{t=a}^{b}\epsilon_t x_t\right\|^p\right)^{1/p}.$$

This gives the first inequality. For the second, apply Doob's maximal inequality (Theorem 32). In particular, let $Z_b = \sup_{1\le a\le b}\left\|\sum_{t=a}^{b}\epsilon_t x_t\right\|$. Then Z_b is a sub-martingale, so Doob's maximal inequality implies $\mathbb{E}_\epsilon \sup_{b\le n} Z_b^p \le (p')^p\,\mathbb{E}_\epsilon Z_n^p$. Applying Doob's inequality once more shows that $\mathbb{E}_\epsilon Z_n^p \le (p')^p\,\mathbb{E}_\epsilon\|\sum_{t=1}^{n}\epsilon_t x_t\|$, which gives the result. ■

A.3. Proofs from Section 6

Since we restrict to the absolute loss in this section and restrict to $y_t \in [-1, +1]$, we can also restrict to $\hat{y}_t \in [-1, +1]$ without loss of generality, since for any value of y_t the loss may always be decreased by clipping $\hat{y}_t$ into this range. In the proof below, any infimum over $\hat{y}_t$ is understood to be over this range.

Proof [Proof of Theorem 14] We shall first show that 2 implies 1, specifically for constant $B = 2C$. We can write down the minimax value for the proposed regret bound and check if it indeed is achievable. To this end, note that

$$\begin{aligned}
\mathcal{V} &= \left\langle\!\!\left\langle \sup_{x_t} \inf_{\hat{y}_t} \sup_{y_t \in [-1,+1]} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - 2C \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \inf_{\hat{y}_t} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} (\ell(\hat{y}_t, y_t) - \ell(f(x_t), y_t)) - 2C \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&\leq \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \inf_{\hat{y}_t} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} \ell'(\hat{y}_t, y_t)(\hat{y}_t - f(x_t)) - 2C \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right]
\end{aligned}$$

setting $\hat{y}_t^*$ to be minimizer of $\mathbb{E}\, \ell(\hat{y}_t, y_t)$, we have

$$\begin{aligned}
&\leq \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \inf_{\hat{y}_t} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} \ell'(\hat{y}_t^*, y_t)(\hat{y}_t^* - f(x_t)) - 2C \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \inf_{\hat{y}_t} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} -\ell'(\hat{y}_t^*, y_t) f(x_t) - 2C \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} (\mathop{\mathbb{E}}_{y_t' \sim p_t} \ell'(\hat{y}_t^*, y_t') - \ell'(\hat{y}_t^*, y_t)) f(x_t) - 2C \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&\leq \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \mathop{\mathbb{E}}_{y_t, y_t' \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} (\ell'(\hat{y}_t^*, y_t') - \ell'(\hat{y}_t^*, y_t)) f(x_t) - 2C \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \mathop{\mathbb{E}}_{y_t, y_t' \sim p_t} \mathop{\mathbb{E}}_{\epsilon_t'} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} \epsilon_t' (\ell'(\hat{y}_t^*, y_t') - \ell'(\hat{y}_t^*, y_t)) f(x_t) - 2C \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&\leq \left\langle\!\!\left\langle \sup_{x_t} \mathop{\mathbb{E}}_{\epsilon_t'} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} 2\epsilon_t' f(x_t) - 2C \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&= \sup_{\mathbf{x}} \mathop{\mathbb{E}}_{\epsilon'} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} 2\epsilon_t' f(\mathbf{x}_t(\epsilon_{1:t-1}')) - 2C \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t(\epsilon_{1:t-1}') \right].
\end{aligned}$$

However by 2, we have that the above is bounded by 0 and so we can conclude that the minimax strategy does attain the regret bound proposed in 1.

Now to prove that 1 implies 2 (with constant B), notice that we have an algorithm that guarantees regret bound:

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq B \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t)$$

Assume now that the adversary at time first provides input instance $\mathbf{x}_t(\epsilon_{1:t-1})$ where $\mathbf{x}$ is any arbitrary $\mathcal{X}$ valued binary tree. Also assume that y_t is picked to be ϵ_t a draw of a coin flip. In this case, we have from the regret bound that

$$\sum_{t=1}^{n} \ell(\hat{y}_t, \epsilon_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(\mathbf{x}_t(\epsilon_{1:t-1})), \epsilon_t) \leq B \mathop{\mathbb{E}}_{\epsilon'} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t' f(\mathbf{x}_t(\epsilon_{1:t-1}))$$

Taking expectation we find that,

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n}\ell(\hat{y}_t,\epsilon_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(\mathbf{x}_t(\epsilon_{1:t-1})),\epsilon_t)\right]\le B\mathop{\mathbb{E}}_{\epsilon,\epsilon'}\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1}))$$

Now notice that irrespective of what $\hat{y}_t$ the algorithm picks, $\mathbb{E}_{\epsilon_t}\ell(\hat{y}_t,\epsilon_t)=1$. Hence,

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}(1-\ell(f(\mathbf{x}_t(\epsilon_{1:t-1})),\epsilon_t))\right]\le B\mathop{\mathbb{E}}_{\epsilon,\epsilon'}\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1}))$$

However note that when $y\in\{\pm1\}$ and $a\in[-1,1]$, we have that $\ell(a,y)=|a-y|=1-ay$. Hence from above we conclude that,

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1}))\right]\le B\mathop{\mathbb{E}}_{\epsilon,\epsilon'}\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1}))$$

Since the above is true for any choice of $\mathbf{x}$, we have shown that 1 implies 2 with constant B. ■

Proof [Proof of Example 7] Let $\mathbf{x}$ be some $\mathcal{X}$-valued tree. Observe that by the reproducing property,

$$\mathop{\mathbb{E}}_{\sigma}\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\sigma_t f(\mathbf{x}_t(\sigma))=\mathop{\mathbb{E}}_{\sigma}\left\|\sum_{t=1}^{n}\sigma_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}},$$

and likewise $\mathbb{E}_{\sigma,\epsilon}\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(\mathbf{x}_t(\sigma))=\mathbb{E}_{\sigma,\epsilon}\|\sum_{t=1}^{n}\epsilon_t K(\cdot,\mathbf{x}_t(\sigma))\|_{\mathcal{H}}$.

Since $\mathcal{H}$ is a Hilbert space the deterministic UMD property for power 2 is trivial. For any fixed sequence $\epsilon\in\{\pm1\}^n$,

$$\mathop{\mathbb{E}}_{\sigma}\left\|\sum_{t=1}^{n}\sigma_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}}^2=\mathop{\mathbb{E}}_{\sigma}\left\|\sum_{t=1}^{n}\epsilon_t\sigma_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}}^2.$$

By Corollary 36, this implies there is some C such that

$$\mathop{\mathbb{E}}_{\sigma}\sup_{\tau\le n}\left\|\sum_{t=1}^{\tau}\sigma_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}}=C\mathop{\mathbb{E}}_{\sigma}\sup_{\tau\le n}\left\|\sum_{t=1}^{\tau}\epsilon_t\sigma_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}}.$$

Now suppose ϵ is drawn uniformly at random. For a fixed draw of σ, Corollary 40 implies that the RHS enjoys the bound

$$\begin{aligned}\mathop{\mathbb{E}}_{\epsilon}\sup_{\tau\le n}\left\|\sum_{t=1}^{\tau}\epsilon_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}}&\le 2\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}}+5\max_{t\in[n]}\|K(\cdot,\mathbf{x}_t(\sigma))\|_{\mathcal{H}}\log(n)\\&\le 2\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}}+5B\log(n).\end{aligned}$$

■

A.3.1. POLYNOMIALS

Suppose we receive data $x_1, \dots, x_n \in \mathbb{R}^d$ and want to compete with a class $\mathcal{F}$ of homogeneous polynomials of degree k. Any homogeneous degree k polynomial f may be represented via a coefficient tensor M in $(\mathbb{R}^d)^{\otimes k}$ via

$$f(x) = \left\langle M, x^{\otimes k} \right\rangle.$$

We may take M to be symmetric, so that $M_{1,\dots,k} = M_{\pi(1),\dots,\pi(k)}$ for any permutation. We may thus work with a class $\mathcal{M} \subseteq (\mathbb{R}^d)^{\otimes k}$ of symmetric tensors, then take $\mathcal{F} = \left\{ x \mapsto \left\langle M, x^{\otimes k} \right\rangle \mid M \in \mathcal{M} \right\}$. Our task is then to decide which norm to place on $\mathcal{M}$. Following, e.g., Adamczak and Wolff (2015); Wang et al. (2016), we define a class of general tensor norms. Let $\mathcal{J} = \{J_1, \dots, J_N\}$ be a partition of $[k]$. For some $\alpha \in [d]^k$ and $J \subseteq [k]$, let $\alpha_J = (\alpha_i)_{i \in J}$. We then define

$$\|M\|_{\mathcal{J}} = \sup \left\{ \sum_{\alpha \in [d]^k} M_\alpha \prod_{l=1}^{N} x^l_{\alpha_{J_l}} \mid \left\|x^l\right\|_2 \le 1 \ \forall l \in [N] \right\}, \tag{42}$$

where $x^l \in (\mathbb{R}^d)^{\otimes |J_l|}$. Under this notation we have $\|M\|_{\{1\},\{2\}}$ as the spectral norm and $\|M\|_{\{1,2\}}$ as the Frobenius norm when $k = 2$ and M is a matrix. In general, $\|M\|_{\{1\},\{2\},\dots,\{k\}}$ is called the *injective tensor norm*.

Proof [Proof of Example 8] Fix an $\mathcal{X}$-valued tree $\mathbf{x}$. Then we have

$$\mathop{\mathbb{E}}_{\sigma} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \sigma_t f(\mathbf{x}_t(\sigma)) = \mathop{\mathbb{E}}_{\sigma} \sup_{M \in \mathcal{M}} \sum_{t=1}^{n} \sigma_t \left\langle M, \mathbf{x}_t(\sigma)^{\otimes 2k} \right\rangle = \mathop{\mathbb{E}}_{\sigma} \left\| \sum_{t=1}^{n} \sigma_t \mathbf{x}_t(\sigma)^{\otimes 2k} \right\|_{\{1,\dots,k\},\{k+1,\dots,2k\}}$$

For some tensor $T \in (\mathbb{R}^d)^{\otimes 2k}$, we can define its flattening $\overline{T}$ into a $\mathbb{R}^{d^k \times d^k}$ matrix and verify that in fact

$$\|T\|_{\{1,\dots,k\},\{k+1,\dots,2k\}} = \max_{u,v \in \mathbb{R}^{d^k} \mid \|u\|_2, \|v\|_2 \le 1} \sum_{\alpha \in [d]^k, \beta \in [d]^k} T_{\alpha,\beta} u_\alpha v_\beta = \left\langle u, \overline{T} v \right\rangle = \|\overline{T}\|_\sigma,$$

so in fact this is the spectral norm of the flattened matrix. Let $\mathbf{X}_t \in \mathbb{R}^{d^k \times d^k}$ be the flattening of $(\mathbf{x}_t)^{\otimes 2k}$. Then

$$\mathop{\mathbb{E}}_{\sigma} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \sigma_t f(\mathbf{x}_t(\sigma)) = \mathop{\mathbb{E}}_{\sigma} \left\| \sum_{t=1}^{n} \sigma_t \mathbf{X}_t(\sigma) \right\|_\sigma,$$

so we can prove the desired inequality by applying the UMD inequality for the spectral norm. Recall from Theorem 10 that the UMD inequality for the spectral norm has a constant of order $\log^2(\dim)$, which for this application translates into a constant of order $O(k^2 \log^2(d))$. We finally apply Corollary 40 as in Example 7 to get the result. ■

A.3.2. LOW-RANK EXPERTS

In this section we prove Theorem 16. The proof relies on the following key lemma, which is proven using the one-sided UMD property for scalars.

Lemma 27 *There exists a strategy $(\hat{y}_t)$ for the experts setting that guarantees*

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \le O\left(\mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\infty^{\log d} \right)^{1/\log d}, \tag{43}$$

where ℓ is any well-behaved 1-Lipschitz loss.

With this lemma, we need one more fact to prove Theorem 16, which is a corollary of John's theorem about the volume of a minimum-volume enclosing ellipsoid.

Lemma 28 (Hazan et al. (2016), Lemma 12) *Let K be a symmetric convex set in $\mathbb{R}^d$. There exists a positive semidefinite matrix Ξ such that for all $x \in K$,*

$$\langle x, \Xi x\rangle \le \sup_{f\in K^\star} |\langle f, x\rangle|^2 \le d \cdot \langle x, \Xi x\rangle. \tag{44}$$

Applying Lemma 28 to the intersection of the ℓ_∞ ball and $\mathrm{span}(x_{1:n})$ gives a Euclidean approximation to the ℓ_∞ norm in terms of the rank of $X_{1:n}$.

Corollary 29
There exists some positive semidefinite $\Xi \in \mathbb{R}^{d\times d}$ such that for all $S \in \mathrm{span}(x_{1:n})$,

$$\langle S, \Xi S\rangle \le \|S\|_\infty^2 \le \mathrm{rank}(X_{1:n}) \cdot \langle S, \Xi S\rangle. \tag{45}$$

We can now proceed to the proof of the main theorem.
Proof [Proof of Theorem 16] By Lemma 27, there exists a strategy whose regret is bounded by

$$O\left(\mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\infty^{\log d} \right)^{1/\log d}.$$

We now complete the upper bound using concentration. Let $Z = \|\sum_{t=1}^n \epsilon_t x_t\|_\infty$. Then we can write $\left(\mathbb{E}_\epsilon \|\sum_{t=1}^n \epsilon_t x_t\|_\infty^{\log d}\right)^{1/\log d}$ as $\left(\mathbb{E}\, Z^{\log d}\right)^{1/\log d}$, where the expectation is over the sequence ϵ. We will upper bound this quantity in terms of the rank. First observe that by Corollary 29, there exists a PSD matrix Ξ such that

$$\mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\infty \le \sqrt{\mathrm{rank}(X_{1:n})}\, \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\Xi,$$

where $\|x\|_\Xi = \langle x, \Xi x\rangle$.

Observe that since $\|\cdot\|_\Xi$ is Euclidean,

$$\mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\Xi \sqrt{\mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\Xi^2} = \sqrt{\sum_{t=1}^{n} \|x_t\|_\Xi^2} \le \sqrt{\sum_{t=1}^{n} \|x_t\|_\infty^2} \le \sqrt{n},$$

where the second-to-last inequality uses Corollary 29. This establishes that

$$\mathbb{E}\, Z \le \sqrt{\mathrm{rank}(X_{1:n})n}.$$

Now, since $\|x_t\|_\infty \le 1$, Lemma 38 implies that with probability at least $1-\delta$ over the draw of ϵ,

$$Z \le O(\mathbb{E}\, Z + \log(1/\delta)).$$

By the law of total expectation, this establishes that for all $\delta > 0$,

$$\left(\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_\infty^{\log d}\right)^{1/\log d} \le O\left(\left((\sqrt{\mathrm{rank}(X_{1:n})n} + \log(1/\delta))^{\log d} + n^{\log d}\delta\right)^{1/\log d}\right).$$

Taking $\delta = n^{-\log d}$, the above quantity is bounded by

$$O\left(\left((\sqrt{\mathrm{rank}(X_{1:n})n} + \log(n)\log(d))^{\log d}\right)^{1/\log d}\right),$$

which is further bounded as

$$O\left(\sqrt{\mathrm{rank}(X_{1:n})n} + \log(n)\log(d)\right).$$

■

Proof [Proof of Theorem 17] This result is proven from the same starting point as in Theorem 16. Recall from Lemma 27 that there is a strategy whose regret is bounded by

$$O\left(\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_\infty^{\log d}\right)^{1/\log d}.$$

Suppose $\mathrm{rank}_\gamma(X_{1:n}) = r$. Then there exist matrices $X'_{1:n} \in \mathbb{R}^{d\times n}$ and $Z_{1:n} \in \mathbb{R}^{d\times n}$ such that

$$X_{1:n} = X'_{1:n} + Z_{1:n},$$

with $\mathrm{rank}(X'_{1:n}) = r$ and $\|Z\|_\infty \le \gamma$. Using x'_t to denote the tth column of $X'_{1:n}$ and z_t to denote the tth column of $Z_{1:n}$, triangle inequality implies

$$\begin{aligned}
\left(\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_\infty^{\log d}\right)^{1/\log d} &= \left(\mathop{\mathbb{E}}_{\epsilon}\|X_{1:n}\epsilon\|_\infty^{\log d}\right)^{1/\log d} \\
&\le O\left(\left(\mathop{\mathbb{E}}_{\epsilon}\|X'_{1:n}\epsilon\|_\infty^{\log d}\right)^{1/\log d} + \left(\mathop{\mathbb{E}}_{\epsilon}\|Z_{1:n}\epsilon\|_\infty^{\log d}\right)^{1/\log d}\right) \\
&= O\left(\left(\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x'_t\right\|_\infty^{\log d}\right)^{1/\log d}\right) + O\left(\left(\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t z_t\right\|_\infty^{\log d}\right)^{1/\log d}\right).
\end{aligned}$$

Since the loss matrix in the first term has rank r, this term can be bounded exactly as in Theorem 16. We now show how to bound the second term. First, observe that since $\|Z_{1:n}\|_\infty \le \gamma$, the standard estimate on the maximum of d subgaussian random variables (e.g. Kakade et al. (2009)) gives

$$\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t z_t\right\|_\infty \le O(\gamma\sqrt{n\log d}).$$

Lemma 38 implies that with probability at least $1-\delta$ over the draw of ϵ

$$\left\|\sum_{t=1}^{n}\epsilon_t z_t\right\|_\infty \le O(\gamma\sqrt{n\log d}+\gamma\log(1/\delta)).$$

Applying the law of total expectation (and recalling that $\gamma \le 1$), this implies that for all $\delta > 0$

$$\left(\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t z_t\right\|_\infty^{\log d}\right)^{1/\log d} \le O\left(\left((\gamma\sqrt{n\log d}+\gamma\log(1/\delta))^{\log d}+n^{\log d}\delta\right)^{1/\log d}\right)$$

Taking $\delta = n^{-\log d}$, the above is finally bounded as

$$O(\gamma\sqrt{n\log d}+\gamma\log n\log d).$$

■

Proof [Proof of Theorem 18] This proof follows the same structure as Theorem 16 and Theorem 17. Starting from Lemma 27, we have that there is a strategy whose regret is bounded by

$$O\left(\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_\infty^{\log d}\right)^{1/\log d}.$$

Observe that $\mathbb{E}_\epsilon\|\sum_{t=1}^n \epsilon_t x_t\|_\infty = \mathbb{E}_\epsilon\|X_{1:n}\epsilon\|_\infty$. From the definition of the max norm, there exist $U \in \mathbb{R}^{d\times d}$, $V \in \mathbb{R}^{n\times d}$ such that $X_{1:n} = UV^\dagger$ and $\|U\|_{\infty,2}\|V\|_{\infty,2} = \|X_{1:n}\|_{\max}$. With this observation, we have

$$\mathop{\mathbb{E}}_{\epsilon}\|X_{1:n}\epsilon\|_\infty = \mathop{\mathbb{E}}_{\epsilon}\|UV^\dagger\epsilon\|_\infty = \mathop{\mathbb{E}}_{\epsilon}\left\|U\sum_{t=1}^{n}v_t\epsilon_t\right\|_\infty,$$

where v_t denotes the tth row of V. Now, observe that

$$\|U\|_{\infty,2} = \max_{i\in[d]}\|u_i\|_2 = \max_{i\in[d]}\max_{x:\|x\|_2\le 1}\langle u_i, x\rangle = \max_{x:\|x\|_2\le 1}\|Ux\|_\infty = \|U\|_{2\to\infty},$$

so $\|\cdot\|_{\infty,2}$ is actually the $2\to\infty$ operator norm. This implies that

$$\mathop{\mathbb{E}}_{\epsilon}\left\|U\sum_{t=1}^{n}v_t\epsilon_t\right\|_\infty \le \|U\|_{\infty,2}\cdot\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}v_t\epsilon_t\right\|_2.$$

Proceeding with the standard Euclidean calculation for Rademacher complexity (e.g. Kakade et al. (2009)), and using that $\|v_t\|_2 \le \|V\|_{\infty,2}\ \forall t$, the above implies that

$$\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_\infty \le \|U\|_{\infty,2}\|V\|_{\infty,2}\sqrt{n} = \|X_{1:n}\|_{\max}\sqrt{n}.$$

Once again, we appeal to Lemma 38, which implies that with probability at least $1-\delta$ over the draw of ϵ,

$$\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_\infty \le O(\|X\|_{\max}\cdot\sqrt{n}+\log(1/\delta)).$$

Again using the law of total expectation, this implies that for all $\delta > 0$

$$\left(\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_{\infty}^{\log d}\right)^{1/\log d} \le O\left(\left((\|X\|_{\max}\cdot\sqrt{n}+\log(1/\delta))^{\log d}+n^{\log d}\delta.\right)^{1/\log d}\right)$$

Taking $\delta = n^{-\log d}$, we have

$$O(\|X\|_{\max}\cdot\sqrt{n}+\log n\log d).$$

■

We now focus on proving Lemma 27. The structure of this proof will follow that of Theorem 14, which gives an upper bound on regret in terms of $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ whenever the one-sided UMD inequality holds. To achieve the desired bound in this framework, we will need the following corollary of Hitczenko's decoupling inequality Theorem 15.

Corollary 30 (One-sided UMD inequality for ℓ_p norms) *There exists some constant K such that for all $p \ge 1$,*

$$\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t\mathbf{x}_t(\epsilon)\right\|_p^p \le K^p \mathop{\mathbb{E}}_{\epsilon,\epsilon'}\left\|\sum_{t=1}^{n}\epsilon'_t\epsilon_t\mathbf{x}_t(\epsilon)\right\|_p^p, \tag{46}$$

where $\mathbf{x}$ is any $\mathcal{X}$-valued tree.

Proof [Proof of Corollary 30] Simply apply Theorem 15 coordinate-wise. ■

With this inequality, we proceed to prove Lemma 27.

Proof [Proof of Lemma 27] Let $p = \log d$. Recall that we have defined

$$\Psi_{\eta,p}(x) = \frac{1}{p}\left(\eta x + \frac{1}{p'-1}\eta^{1-p'}\right).$$

We first will prove that there is a strategy $(\hat{y}_t)$ that achieves

$$\sum_{t=1}^{n}\ell(\hat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t) \le \Psi_{\eta,p}\left(C\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_{\infty}^{p}\right)$$

for some $C > 0$. This portion of the proof will closely follow Theorem 14. Fix C to be decided later and define

$$\mathcal{V} = \left\langle\!\!\left\langle \sup_{x_t}\inf_{\hat{y}_t}\sup_{y_t\in[-1,+1]}\right\rangle\!\!\right\rangle_{t=1}^{n}\left[\sum_{t=1}^{n}\ell(\hat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t) - \Psi_{\eta,p}\left(C\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_{\infty}^{p}\right)\right].$$

The infimum over $\hat{y}_t$ is understood to range over $[-B,+B]$, as guaranteed by the assumption on the loss in Section 2. Observe that the regret bound we desired is achievable if there is a value for C such that $\mathcal{V} \le 0$. Using the minimax theorem, the value is equal to

$$\mathcal{V} = \left\langle\!\!\left\langle \sup_{x_t}\sup_{p_t\in\Delta[-1,+1]}\inf_{\hat{y}_t}\mathop{\mathbb{E}}_{y_t\sim p_t}\right\rangle\!\!\right\rangle_{t=1}^{n}\sup_{f\in\mathcal{F}}\left[\sum_{t=1}^{n}(\ell(\hat{y}_t,y_t)-\ell(f(x_t),y_t)) - \Psi_{\eta,p}\left(C\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_{\infty}^{p}\right)\right].$$

Linearizing the loss, the above is bounded by

$$\leq \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \inf_{\hat{y}_t} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} \ell'(\hat{y}_t, y_t)(\hat{y}_t - f(x_t)) - \Psi_{\eta,p}\left(C \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\infty^p \right) \right].$$

Setting $\hat{y}_t^*$ to be minimizer of $\mathbb{E}\,\ell(\hat{y}_t, y_t)$, we have

$$\begin{aligned}
&\leq \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \inf_{\hat{y}_t} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} \ell'(\hat{y}_t^*, y_t)(\hat{y}_t^* - f(x_t)) - \Psi_{\eta,p}\left(C \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\infty^p \right) \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \inf_{\hat{y}_t} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} -\ell'(\hat{y}_t^*, y_t) f(x_t) - \Psi_{\eta,p}\left(C \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\infty^p \right) \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} (\mathop{\mathbb{E}}_{y_t' \sim p_t} \ell'(\hat{y}_t^*, y_t') - \ell'(\hat{y}_t^*, y_t)) f(x_t) - \Psi_{\eta,p}\left(C \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\infty^p \right) \right] \\
&\leq \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \mathop{\mathbb{E}}_{y_t, y_t' \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} (\ell'(\hat{y}_t^*, y_t') - \ell'(\hat{y}_t^*, y_t)) f(x_t) - \Psi_{\eta,p}\left(C \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\infty^p \right) \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \mathop{\mathbb{E}}_{y_t, y_t' \sim p_t} \mathop{\mathbb{E}}_{\epsilon_t'} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} \epsilon_t' (\ell'(\hat{y}_t^*, y_t') - \ell'(\hat{y}_t^*, y_t)) f(x_t) - \Psi_{\eta,p}\left(C \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\infty^p \right) \right] \\
&\leq \left\langle\!\!\left\langle \sup_{x_t} \mathop{\mathbb{E}}_{\epsilon_t'} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} 2\epsilon_t' f(x_t) - \Psi_{\eta,p}\left(C \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\infty^p \right) \right] \\
&= \sup_{\mathbf{x}} \mathop{\mathbb{E}}_{\epsilon'} \sup_{f \in \mathcal{F}} \left[\sum_{t=1} 2\epsilon_t' f(\mathbf{x}_t(\epsilon')) - \Psi_{\eta,p}\left(C \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t(\epsilon') \right\|_\infty^p \right) \right].
\end{aligned}$$

Using that the simplex Δ_d is a subset of the ℓ_1 ball:

$$\leq \sup_{\mathbf{x}} \mathop{\mathbb{E}}_{\epsilon'} \left[2 \left\| \sum_{t=1} \epsilon_t' \mathbf{x}_t(\epsilon') \right\|_\infty - \Psi_{\eta,p}\left(C \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t(\epsilon') \right\|_\infty^p \right) \right].$$

Using (22), this is upper bounded by

$$= \sup_{\mathbf{x}} \mathop{\mathbb{E}}_{\epsilon'} \frac{\eta}{p} \left[2 \left\| \sum_{t=1} \epsilon_t' \mathbf{x}_t(\epsilon') \right\|_\infty^p - C \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t(\epsilon') \right\|_\infty^p \right].$$

We can replace the left ℓ_∞ norm with the ℓ_p norm as an upper bound:

$$\leq \sup_{\mathbf{x}} \mathop{\mathbb{E}}_{\epsilon'} \frac{\eta}{p} \left[2 \left\| \sum_{t=1} \epsilon_t' \mathbf{x}_t(\epsilon') \right\|_p^p - C \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t(\epsilon') \right\|_\infty^p \right].$$

We now apply the one-sided UMD property for the ℓ_p norm Corollary 30:

$$\leq \sup_{\mathbf{x}} \mathop{\mathbb{E}}_{\epsilon, \epsilon'} \frac{\eta}{p} \left[2K^p \left\| \sum_{t=1} \epsilon_t \mathbf{x}_t(\epsilon') \right\|_p^p - C \left\| \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t(\epsilon') \right\|_\infty^p \right].$$

Finally, since $p = \log d$, there is some constant A such that $\|x\|_p \leq A\|x\|_\infty$ pointwise. Therefore, if we take $C = O(K)^p$, the expression is bounded by zero.

Now, to achieve the bound stated in the theorem, simply using the doubling trick given in Lemma 26 on top of the strategy described above. Since $p' = O(1)$, the doubling strategy will guarantee a regret bound of

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \le O\left(K \left(\mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_\infty^p \right)^{1/p} \right).$$

■

Proof [Proof of Theorem 15] This theorem is an immediate corollary of (Hitczenko, 1994), Theorem 1.1. We will spend a moment to explain this in detail, as that theorem is stated in terms of *tangent sequences*, which are a concept that otherwise does not appear in the present paper.

Given an adapted sequence $(Z_t)_{t \le n}$, we define its *decoupled tangent sequence* $(Z'_t)_{t \le n}$ as follows: At time t, conditioned on $Z_{1:t-1}$, sample Z'_t as an i.i.d. copy of Z_t under the conditional distribution $\Pr(Z_t \mid Z_1, \dots, Z_{t-1})$. Then $(Z'_t)_{t \le n}$ satisfies

1. Identical conditional distribution: $\Pr(Z'_t \mid Z_1, \dots, Z_{t-1}) = \Pr(Z_t \mid Z_1, \dots, Z_{t-1})$
2. Conditional independence: $\Pr(Z'_1, \dots, Z'_n \mid Z_1, \dots, Z_n) = \prod_{t=1}^{n} \Pr(Z'_t \mid Z_1, \dots, Z_n)$

With this definition, (Hitczenko, 1994), Theorem 1.1 is stated as follows:
There is some universal constant K such that for any adapted sequence (Z_t) and its decoupled tangent sequence (Z'_t), for any $1 \le p < \infty$,

$$\mathbb{E} \left| \sum_{t=1}^{n} Z_t \right|^p \le K^p \left| \sum_{t=1}^{n} Z'_t \right|^p. \tag{47}$$

We now show how to conclude Theorem 15 from this result. Observe that for a Paley-Walsh martingale $(\epsilon_t \mathbf{x}_t(\epsilon_{t:t-1}))_{t=1}^n$, its decoupled tangent sequence is given by $(\epsilon'_t \mathbf{x}_t(\epsilon_{t:t-1}))_{t=1}^n$, where ϵ' is an independent sequence of Rademacher random variables. Furthermore, this sequence is distributed identically to $(\epsilon'_t \epsilon_t \mathbf{x}_t(\epsilon_{t:t-1}))_{t=1}^n$. Therefore Theorem 15 follows from specializing (47) to Paley-Walsh martingales. ■

A.3.3. ONLINE MATRIX PREDICTION

Proof [Proof of Theorem 19] Recall that $\mathcal{F} = \left\{ F \in \mathbb{R}^{d \times d} \mid \|F\|_\Sigma \le \tau \right\}$. We will take $\tau = r\sqrt{d}$, which implies that $\mathcal{F}$ contains all matrices F with $\operatorname{rank}(F) \le r$ and $\|F\|_\infty \le 1$.

Let $X_t = e_{i_t} \otimes e_{j_t}$ be the incidence matrix for the entry (i_t, j_t). Then we may write $F(x_t) = \langle F, X_t \rangle$. With this notation, we have

$$\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}) = \tau \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t X_t \right\|_\sigma.$$

From Theorem 10, the class $\mathcal{F}$ has UMD constant $\mathbf{C}_2 \le \log^2(d)$. Therefore, by Theorem 9, there exists a randomized strategy that guarantees[9]

$$\mathbb{E}\left[\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{F\in\mathcal{F}} \sum_{t=1}^n \ell(F(x_t), y_t)\right] \le O\left(\tau \mathbf{C}_p \mathbb{E}\left(\mathbb{E}_\epsilon \left\|\sum_{t=1}^n \epsilon_t X_t\right\|_\sigma + \max_{t\in[n]} \|X_t\|_\sigma \log(n)\right)\right)$$
$$= \widetilde{O}\left(r\sqrt{d}\,\mathbb{E}\,\mathbb{E}_\epsilon \left\|\sum_{t=1}^n \epsilon_t X_t\right\|_\sigma\right).$$

We now apply concentration to remove the expectation over ϵ. Observe that the spectral norm of each X_t is bounded by 1 (since each X_t is an indicator matrix). Hence by Theorem 6.1 of Tropp (2012) we have, letting $\sigma^2 = \max\left\{\left\|\sum_t X_t X_t^\dagger\right\|_\sigma, \left\|\sum_t X_t^\dagger X_t\right\|_\sigma\right\}$, we have that with probability at least $1-\delta$ over the draw of ϵ,

$$\left\|\sum_{t=1}^n \epsilon_t X_t\right\|_\sigma \le O(\sigma \log(d/\delta)).$$

Since each X_t has $\|X_t\|_\sigma \le 1$, the law of total expectation then implies that

$$\mathbb{E}_\epsilon \left\|\sum_{t=1}^n \epsilon_t X_t\right\|_\sigma \le O(\sigma \log(nd)).$$

(X_t) are incidence matrices, and so $\sum_t X_t X_t^\dagger$ and $\sum_t X_t^\dagger X_t$. A straightforward calculation reveals:

$$\sigma = \sqrt{\max\left\{\max_i |\{t \mid i_t = i\}|, \max_j |\{t \mid j_t = j\}|\right\}} = \sqrt{\max\{N_{\text{row}}, N_{\text{col}}\}}.$$

■

A.3.4. EMPIRICAL COVERING NUMBER BOUNDS

Proof [Proof of Theorem 22 and Theorem 23] Theorem 14 proves that when the one-sided UMD-property (27) holds, there exists a strategy whose regret is bounded as

$$C\,\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t).$$

Since this quantity is the statistical Rademacher complexity, we may apply the classical covering number bound (Rakhlin and Sridharan, 2012, Proposition 12.3):

$$\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t) \le O\left(\inf_{\alpha>0}\left\{\alpha n + \sqrt{\log \mathcal{N}_1(\Delta_d, \alpha, x_{1:n})n}\right\}\right).$$

Likewise, the classical Dudley entropy integral bound (Rakhlin and Sridharan, 2012, Theorem 12.4) yields:

$$\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t) \le O\left(\inf_{\alpha>0}\left\{\alpha \cdot n + \int_\alpha^1 \sqrt{\log \mathcal{N}_2(\mathcal{F}, \delta, x_{1:n})n}d\delta\right\}\right).$$

9. One can deduce the existence of a deterministic strategy achieving this bound from Theorem 14.

■

Proof [Proof of Theorem 24 and Theorem 25] By Lemma 27, there exists a strategy whose regret is bounded by

$$O\left(\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_\infty^{\log d}\right)^{1/\log d}.$$

Observe that

$$\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_\infty = \mathop{\mathbb{E}}_{\epsilon}\sup_{f\in\Delta_d}\sum_{t=1}^{n}\epsilon_t f(x_t).$$

We prove the theorem by appealing to the following classical empirical process bounds (Rakhlin and Sridharan, 2012, Proposition 12.3, Theorem 12.4). For Theorem 24:

$$\mathop{\mathbb{E}}_{\epsilon}\sup_{f\in\Delta_d}\sum_{t=1}^{n}\epsilon_t f(x_t) \le O\left(\inf_{\alpha>0}\left\{\alpha n + \sqrt{\log \mathcal{N}_1(\Delta_d,\alpha,x_{1:n})n}\right\}\right).$$

For Theorem 25:

$$\mathop{\mathbb{E}}_{\epsilon}\sup_{f\in\Delta_d}\sum_{t=1}^{n}\epsilon_t f(x_t) \le O\left(\inf_{\alpha>0}\left\{\alpha n + \int_\alpha^1 \sqrt{\log \mathcal{N}_2(\Delta_d,\delta,x_{1:n})n}d\delta\right\}\right).$$

To show the final bound, proceed with the concentration argument used in the proof of Theorem 16.

■

Appendix B. UMD spaces and martingale inequalities

B.1. Stopping inequalities

Let (Z_t) be a martingale. For two stopping times τ_1, τ_2, we define its stopped version as $Z_t^{\tau_1:\tau_2}$ via

$$dZ_t^{\tau_1:\tau_2} = dZ_t\mathbb{1}\{t > \tau_1\}\mathbb{1}\{t \le \tau_2\}.$$

Proposition 31 (Hytönen et al. (2016), Proposition 3.1.14) *For any $p \in [1,\infty)$,*

$$\mathbb{E}\|Z_n^{\tau_1:\tau_2}\|^p \le 2^p\,\mathbb{E}\|Z_n\|^p.$$

Theorem 32 (Doob's Maximal Inequality) *For any martingale $(Z_t)_{t\ge1}$ taking values in $(\mathfrak{B},\|\cdot\|)$ and any $p\in(1,\infty]$,*

$$\mathbb{E}\sup_{\tau\le n}\left\|\sum_{t=1}^{\tau}dZ_t\right\|^p \le (p')^p\,\mathbb{E}\left\|\sum_{t=1}^{n}dZ_t\right\|^p. \tag{48}$$

Furthermore

$$\Pr\left(\sup_{\tau\le n}\left\|\sum_{t=1}^{\tau}dZ_t\right\| > \lambda\right) \le \frac{1}{\lambda}\,\mathbb{E}\left\|\sum_{t=1}^{n}dZ_t\right\| \quad \forall\lambda>0. \tag{49}$$

More generally, (48) and (49) hold when the sequence $(\|\sum_{t=1}^{\tau}Z_t\|)_{\tau\ge1}$ is replaced by any non-negative submartingale $(F_\tau)_{\tau\ge1}$.

Corollary 33 *If (F_n) is a non-negative submartingale and $F_n \le A$ almost surely then for all $\eta > 0$,*

$$\mathbb{E}\left[\max_{\tau\le n} F_\tau\right] \le (\log A + \log \eta)\cdot \mathbb{E}[F_n] + \frac{1}{\eta}.$$

Proof [Proof of Corollary 33]

$$\begin{aligned}
\mathbb{E}\left[\max_{\tau\le n} F_\tau\right] &= \int_0^\infty \Pr\left(\max_{\tau\le n} F_\tau > \lambda\right) d\lambda \\
&= \int_0^A \Pr\left(\max_{\tau\le n} F_\tau > \lambda\right) d\lambda \\
&\le \int_{1/\eta}^A \Pr\left(\max_{\tau\le n} F_\tau > \lambda\right) d\lambda + \frac{1}{\eta} \\
&\le \mathbb{E}[F_n] \int_{\frac{1}{\eta}}^A \frac{1}{\lambda} d\lambda + \frac{1}{\eta} \\
&= (\log A + \log \eta)\cdot \mathbb{E}[F_n] + \frac{1}{\eta}.
\end{aligned}$$

■

B.2. UMD inequalities

Theorem 34 (Hytönen et al. (2016), Theorem 4.2.7) *Suppose $(\mathfrak{B}, \|\cdot\|)$ is such that the determinstic UMD inequality*

$$\mathbb{E}\left\|\sum_{t=1}^n \epsilon_t dZ_t\right\|^p \le \mathbf{C}_p^p\, \mathbb{E}\left\|\sum_{t=1}^n dZ_t\right\|^p$$

holds for $p \in (1,\infty)$. Then the determinstic UMD inequality

$$\mathbb{E}\left\|\sum_{t=1}^n \epsilon_t dZ_t\right\|^q \le \mathbf{C}_q^q\, \mathbb{E}\left\|\sum_{t=1}^n dZ_t\right\|^q$$

holds for any $q \in (1,\infty)$, with

$$\mathbf{C}_q \le 100\left(\frac{q}{p} + \frac{q'}{p'}\right)\mathbf{C}_p.$$

Theorem 35 (Pisier (2011), Theorem 8.23) *Suppose that the deterministic UMD inequality*

$$\sup_n \mathbb{E}\left\|\sum_{t=1}^n \epsilon_t dZ_t\right\|^2 \le \mathbf{C}_2^2 \sup_n \mathbb{E}\left\|\sum_{t=1}^n dZ_t\right\|^2$$

holds for any sign sequence. Then the L_1 UMD inequality

$$\mathbb{E}\sup_n \left\|\sum_{t=1}^n \epsilon_t dZ_t\right\| \le 54\mathbf{C}_2\, \mathbb{E}\sup_n \left\|\sum_{t=1}^n dZ_t\right\|$$

holds as well.

Corollary 36 *If deterministic UMD inequality*

$$\mathbb{E}\left\|\sum_{t=1}^{n}\epsilon_t dZ_t\right\|^2 \le \mathbf{C}_2^2\,\mathbb{E}\left\|\sum_{t=1}^{n} dZ_t\right\|^2$$

holds for any sign sequence, then the L_1 UMD inequality

$$\mathbb{E}\sup_n\left\|\sum_{t=1}^{n}\epsilon_t dZ_t\right\| \le 108\mathbf{C}_2\,\mathbb{E}\sup_n\left\|\sum_{t=1}^{n} dZ_t\right\|$$

holds as well.

Theorem 37 (Hytönen et al. (2016), Proposition 4.2.17) *If $(\mathfrak{B}, \|\cdot\|)$ is UMD_p with constant $\mathbf{C}_p$, then $(\mathfrak{B}^\star, \|\cdot\|_\star)$ is $\mathrm{UMD}_{p'}$ with constant $\mathbf{C}_{p'} = \mathbf{C}_p$.*

B.3. Concentration for Rademacher complexity

Lemma 38 (Bartlett et al. (2005), Theorem A.2) *With probability at least $1-\delta$ over the draw of ϵ,*

$$\begin{aligned}\left\|\sum_{t=a}^{b}\epsilon_t y_t\right\| &\le \mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=a}^{b}\epsilon_t y_t\right\| + \sqrt{\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=a}^{b}\epsilon_t y_t\right\| \cdot 2\max_{t\in[n]}\|y_t\|\log(1/\delta)} + \frac{\max_{t\in[n]}\|y_t\|\log(1/\delta)}{3}\\ &\le 2\,\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=a}^{b}\epsilon_t y_t\right\| + \max_{t\in[n]}\|y_t\|\log(1/\delta).\end{aligned}$$

Lemma 39 *For any fixed sequence $y_1, \ldots, y_n$, with probability at least $1-\delta$ over the draw of ϵ,*

$$\sup_{1\le a\le b\le n}\left\|\sum_{t=a}^{b}\epsilon_t y_t\right\| \le 4\,\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t y_t\right\| + 2\max_{t\in[n]}\|y_t\|\log(n/\delta).$$

Corollary 40

$$\mathop{\mathbb{E}}_{\epsilon}\sup_{1\le a\le b\le n}\left\|\sum_{t=a}^{b}\epsilon_t y_t\right\| \le 4\,\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t y_t\right\| + 5\max_{t\in[n]}\|y_t\|\log(n).$$

Proof [Proof of Lemma 39] Consider $Z = \left\|\sum_{t=a}^{b}\epsilon_t y_t\right\|$ for fixed a, b and a fixed sequence $y_1, \ldots, y_n$. Applying Lemma 38 and taking a union bound over all possible pairs (a, b), of which there are strictly less than n^2, we have that with probability at least $1-\delta$,

$$\sup_{1\le a\le b\le n}\left\|\sum_{t=a}^{b}\epsilon_t y_t\right\| \le 2\sup_{1\le a\le b\le n}\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=a}^{b}\epsilon_t y_t\right\| + 2\max_{t\in[n]}\|y_t\|\log(n/\delta).$$

By Proposition 31:

$$\le 4\,\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t y_t\right\| + 2\max_{t\in[n]}\|y_t\|\log(n/\delta).$$

■

Appendix C. Burkholder/Bellman functions

C.1. Elementary design of U functions

The following construction for the scalar case does not obtain optimal constants, but should give the reader a taste of how one can construct a $\mathbf{U}$ function from first principles.

Theorem 41 (Elementary Scalar U Function) *Let $k \geq 4$ be an even integer. Then the function*

$$\mathbf{U}(x,y) = \frac{k}{2}\left(x^k - 2\binom{k}{2}x^{k-2}y^k - \frac{1}{k-2}\binom{k}{2}^{-1}\left(4\binom{k}{2}\binom{k-2}{2}\right)^{k-2}y^k\right).$$

is Burkholder for $|\cdot|^k$, with UMD constant

$$\mathbf{C}_k \leq \alpha k^4$$

for some constant α.

Proof Let $\widetilde{\mathbf{U}}(x,y) = x^k - Cx^{k-2}y^2 - By^k$. We will show that $\widetilde{\mathbf{U}}$ is Burkholder for an appropriate choice of constants B and C.

Fix $h \in \mathbb{R}$ and let $G(t) = \widetilde{\mathbf{U}}(x + ht, y + \epsilon ht)$ for $\epsilon \in \{\pm 1\}$. By direct calculation we have

$$G''(0) = 2h^2\left[\binom{k}{2}x^{k-2} - C\left(\binom{k-2}{2}x^{k-4}y^2 + 2\binom{k-2}{2}\epsilon x^{k-3}y + x^{k-2}\right) - B\binom{k}{2}y^{k-2}\right]$$

Since k is even, $x^{k-4}y^2$ is a square; we will simply drop this term.

$$\leq 2h^2\left[\binom{k}{2}x^{k-2} - C\left(2\binom{k-2}{2}\epsilon x^{k-3}y + x^{k-2}\right) - B\binom{k}{2}y^{k-2}\right]$$
$$\leq 2h^2\left[\binom{k}{2}x^{k-2} + 2C\binom{k-2}{2}|x|^{k-3}|y| - Cx^{k-2} - B\binom{k}{2}y^{k-2}\right]$$

By Young's inequality, we have

$$2C\binom{k-2}{2}|x|^{k-3}|y| = \underbrace{\left(2C\binom{k-2}{2}|y|\right)}_{a} \cdot \underbrace{|x|^{k-3}}_{b} \leq \frac{1}{k-2}\left((2C\binom{k-2}{2})^{k-2}y^{k-2} + (k-3)x^{k-2}\right),$$

where we have applied $a \cdot b \leq \frac{1}{k-2}a^{k-2} + \frac{k-3}{k-2}b^{\frac{k-2}{k-3}}$.

Returning to $G''(0)$, we now have

$$G''(0) \leq 2h^2\left[\left(\binom{k}{2} + \frac{k-3}{k-2} - C\right)x^{k-2} + \left(\frac{1}{k-2}\left(2C\binom{k-2}{2}\right)^{k-2} - B\binom{k}{2}\right)y^{k-2}\right].$$

In particular, we can take $C \geq 2\binom{k}{2}$ and $B \geq \frac{1}{k-2}\left(2C\binom{k-2}{2}\right)^{k-2}\binom{k}{2}^{-1}$.

$$\leq 0.$$

This certifies that G is zig-zag concave. To see the upper bound property, observe by that Young's inequality,

$$x^k - Cx^{k-2}y^2 - By^k \geq \frac{2}{k}x^k - \left(\frac{2}{k}C^{\frac{k}{2}} + B\right)y^k.$$

Hence, if we take $\mathbf{U}(x,y) = \frac{k}{2}\widetilde{\mathbf{U}}(x,y)$, we have

$$\mathbf{U}(x,y) \geq x^k - \left(C^{\frac{k}{2}} + \frac{k}{2}B\right)y^k.$$

■

C.2. U functions for $p = 1$

Definition 42 ($(1,1)$ **Weak Type Burkholder Function**) *A function* $\mathbf{U} : \mathfrak{B} \times \mathfrak{B} \to \mathbb{R}$ *is* $(\|\cdot\|, \beta)$ *Burkholder for weak type if*

1. $\mathbf{U}(x, x') \geq \mathbb{1}\{\|x\| \geq 1\} - \beta\|x'\|$.
2. $\mathbf{U}$ *is* zig-zag concave*:* $z \mapsto \mathbf{U}(x + \epsilon z, x' + z)$ *is concave for all* $x, x' \in \mathcal{X}$ *and* $\epsilon \in \{\pm 1\}$.
3. $\mathbf{U}(0,0) \leq 0$.

Lemma 43 *Suppose we are given a weak type Burkholder function* $\mathbf{U}_{\|\cdot\|,\text{weak}}$ *for* $(\|\cdot\|, \beta)$*. Then for all arguments* x, y *with* $\|x\|, \|y\| \leq B$*, the following function is Burkholder for* $(\|\cdot\|, 1, C\beta\log(B/\epsilon))$ *up to additive slack* ϵ*:*

$$\mathbf{U}_{\|\cdot\|,1}(x,y) \triangleq \epsilon \sum_{k=1}^{N} \mathbf{U}_{\|\cdot\|,\text{weak}}(x/\lambda_k, y/\lambda_k), \tag{50}$$

where $N = \lceil B/\epsilon \rceil$ *and* $\lambda_k = k\epsilon$*.*

Proof [Proof of Lemma 43] Let $V(x,y) = \|x\| - C'\beta\log(B/\epsilon)\|y\| - \epsilon$. We will show that $\mathbf{U}(x,y) \geq V(x,y)$ when $\|x\|, \|y\| \leq B$.

$$\begin{aligned}
V(x,y) &= \|x\| - C'\beta\log(B/\epsilon)\|y\| - \epsilon \\
&\leq \epsilon + \epsilon\sum_{k=1}^{N} \mathbb{1}\{\|x\| \geq \lambda_k\} - C'\beta\log(B/\epsilon)\|y\| - \epsilon \\
&\leq \epsilon\sum_{k=1}^{N}\left[\mathbf{U}_{\|\cdot\|,\text{weak}}(x/\lambda_k, y/\lambda_k) + \frac{\beta}{\lambda_k}\|y\|\right] - C'\beta\log(B/\epsilon)\|y\| \\
&= \mathbf{U}_{\|\cdot\|,1}(x,y) + \epsilon\sum_{k=1}^{N}\frac{\beta}{\lambda_k}\|y\| - C'\beta\log(B/\epsilon)\|y\| \\
&= \mathbf{U}_{\|\cdot\|,1}(x,y) + \beta\|y\|\sum_{k=1}^{N}\frac{1}{k} - C'\beta\log(B/\epsilon)\|y\| \\
&\leq \mathbf{U}_{\|\cdot\|,1}(x,y) + C\beta\|y\|\log(N) - C'\beta\log(B/\epsilon)\|y\|
\end{aligned}$$

For sufficiently large C':

$$\leq \mathbf{U}_{\|\cdot\|,1}(x,y).$$

It can be seen immediately that $\mathbf{U}_{\|\cdot\|,1}(x,y)$ is zig-zag concave and has $\mathbf{U}_{\|\cdot\|,1}(0,0) \leq 0$. ■

C.2.1. ζ-CONVEXITY

Definition 44 *Say $(\mathfrak{B}, \|\cdot\|)$ is ζ-convex if there exists $\zeta : \mathfrak{B} \times \mathfrak{B} \to \mathbb{R}$ such that*

1. *ζ is biconvex.*
2. *$\zeta(x, y) \le \|x + y\|$ if $\|x\| = \|y\| = 1$,*

Given a such a function ζ, we can construct a "canonical" function u which satisfies some additional properties

Definition 45

$$u(x, y) \triangleq \begin{cases} \max\{\zeta(x, y), \|x + y\|\}, & \max\{\|x\|, \|y\|\} < 1 \\ \|x + y\|, & \max\{\|x\|, \|y\|\} \ge 1. \end{cases}.$$

Then u is biconvex, has $\zeta(0, 0) \le u(0, 0)$, and satisfies

$$u(x, y) \le \|x + y\| \quad \text{if } \max\{\|x\|, \|y\|\} \ge 1.$$

Also, $u(x, y) = u(-x, -y)$.

Assumption 1 $u(x, -x) \le 0$.

The ζ function given in Example 9 satisfies this condition. More generally, most ζ functions can be made to satisfy this property with a slight blowup in the UMD constant they imply (c.f. (Burkholder, 1986, Lemma 8.5)).

By (Burkholder, 1986, 8.6) Assumption 1 implies $u(x, y) \le u(0, 0) + \|x + y\|$. The following argument due to (Burkholder, 1986) shows how to create a $\mathbf{U}$ function from the function u.

Theorem 46 *Suppose $\|\cdot\|$ is ζ-convex and u satisfies Assumption 1. Then this space is UMD with weak type estimate*

$$\Pr\left(\left\|\sum_{t=1}^{n} dZ_t\right\| \ge 1\right) \le \frac{2}{u(0, 0)} \mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t dZ_t\right\|$$

for any martingale difference sequence (dZ_t). Furthermore, the function

$$\mathbf{U}(x, y) = 1 - \frac{u(x + y, y - x)}{u(0, 0)}$$

is weak-type Burkholder for $(\|\cdot\|, \frac{2}{\zeta(0,0)})$, in the sense of Definition 42.

Proof [Proof of Theorem 46] For the weak type estimate, we will start with the base function

$$V(x, y) = \mathbb{1}\{\|x\| \ge 1\} - \frac{2}{u(0, 0)}\|y\|.$$

We will now show that $V(x, y) \le \mathbf{U}(x, y)$. First, observe that

$$\mathbb{1}\{\|x\| \ge 1\} = \mathbb{1}\{\|(x + y) + (x - y)\| \ge 2\} \le \mathbb{1}\{\max\{\|x + y\|, \|y - x\|\} \ge 1\} \le \mathbb{1}\{2\|y\| \ge u(x + y, y - x)\},$$

where the last inequality follows from the additional property of u from Definition 45. We have now established

$$\begin{aligned} V(x,y) &\le \mathbb{1}\{2\|y\| \ge u(x+y,y-x)\} - \frac{2}{u(0,0)}\|y\| \\ &= \mathbb{1}\{2\|y\| - u(x+y,y-x) + u(0,0) \ge u(0,0)\} - \frac{2}{u(0,0)}\|y\| \end{aligned}$$

By the second additional property of u from Definition 45, $2\|y\| - u(x+y,y-x) + u(0,0) \ge 0$, and so we may apply Markov's inequality

$$\begin{aligned} &\le \frac{2\|y\| - u(x+y,y-x) + u(0,0)}{u(0,0)} - \frac{2}{u(0,0)}\|y\| \\ &= \mathbf{U}(x,y). \end{aligned}$$

Observe that $\mathbf{U}(0,0) = 0$ and, since u is biconvex, $-u(x+y,y-x)$ is zig-zag concave, and so $\mathbf{U}$ is itself zig-zag concave. We can now prove that the UMD property holds with constant $\frac{2}{u(0,0)} \le \frac{2}{\zeta(0,0)}$ using the standard step-by-step peeling argument with $\mathbf{U}$ described in Hytönen et al. (2016), Theorem 4.5.6. ■

Example 9 (ℓ_1^d **Osekowski (2016)**) *Define*

$$z(x,y) = \begin{cases} \frac{a\langle x,y\rangle}{2} - \frac{1}{2a}, & \|x+y\| + \|x-y\| \le 2/a \\ \frac{\|x+y\|}{2}\log\Big(\frac{a}{2}\big(\|x+y\| + \|x-y\|\big)\Big) - \frac{\|x-y\|}{2}, & \|x+y\| + \|x-y\| > 2/a \end{cases}.$$

Then define

$$\zeta(x,y) = \frac{2}{\log(3a)}\left(1 + \sum_{i=1}^{d} z(x_i,y_i)\right).$$

For $a \ge d\log d$ the ζ-convexity properties are satisfied and the bound

$$\zeta(0,0) \le \frac{2}{\log d + \log(2\log d)}\left(1 - \frac{1}{2\log d}\right)$$

is achieved.