# Online Learning: Sufficient Statistics and the Burkholder Method

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

**Alexander Rakhlin** RAKHLIN@MIT.EDU
*Massachusetts Institute of Technology*

**Karthik Sridharan** SRIDHARAN@CS.CORNELL.EDU
*Cornell University*

**Editors:** Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

## Abstract

We uncover a fairly general principle in online learning: If a regret inequality can be (approximately) expressed as a function of certain "sufficient statistics" for the data sequence, then there exists a special *Burkholder function* that 1) can be used algorithmically to achieve the regret bound and 2) only depends on these sufficient statistics, not the entire data sequence, so that the online strategy is only required to keep the sufficient statistics in memory. This characterization is achieved by bringing the full power of the *Burkholder Method*—originally developed for certifying probabilistic martingale inequalities—to bear on the online learning setting.

To demonstrate the scope and effectiveness of the Burkholder method, we develop a novel online strategy for matrix prediction that attains a regret bound corresponding to the variance term in matrix concentration inequalities. We also present a linear-time/space prediction strategy for parameter-free supervised learning with linear classes and general smooth norms.

## 1. Introduction

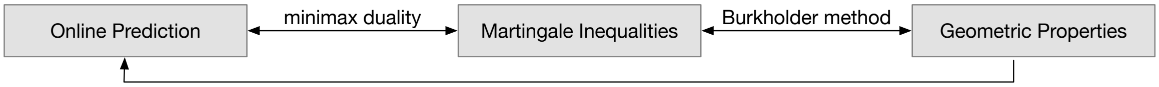
Two of the most appealing features of online learning methods are (a) robustness, due to the absence of assumptions on the data-generating process, and (b) the ability to efficiently incorporate data on the fly. According to this latter desideratum, online methods should not store all the data observed so far in memory, but instead maintain some "compressed" representation, sufficient for making online predictions. The focus of this work is the study of such *sufficient statistics* for online learning, and the design of computationally efficient methods that employ them.

It is natural to turn to Statistics for inspiration: a classical notion of *sufficient statistics* (Fisher, 1922) ensures that a statistician can search for methods that work on "compressed" representations of the data. Sufficient statistics have also been studied in sequential decision theory (Bahadur et al., 1954). However, the very notion of sufficiency is inherently tied to the posited probabilistic model, and the corresponding notion for arbitrary sequences—as postulated by the above desideratum (a)—is all but obvious.

The current theory of online learning offers little guidance as to what summaries of past data should be recorded by an online algorithm. For instance, the Exponential Weights algorithm (Vovk, 1990; Littlestone and Warmuth, 1994) keeps in memory the cumulative losses of the experts, while the general potential-based forecaster (Cesa-Bianchi and Lugosi, 2006) updates the cumulative regret of the algorithm with respect to each expert. The methods from the Follow-the-Regularized-Leader family (also known as Dual Averaging methods) work with the sum of gradients of convex functions, while the Online Newton Step (Hazan et al., 2007) method and the Vovk-Azoury-Warmuth forecaster (Cesa-Bianchi and Lugosi, 2006) also store the "covariance" matrix of outer products. The well-known adaptive gradient descent procedure (e.g. (Rakhlin and Sridharan, 2017)) tunes the step size for online gradient descent according to the cumulative squared norms of gradients, a statistic that appears to be necessary for achieving the adaptive bound, while the ZigZag method of Foster et al. (2017b) keeps track of a sign-transformed sequence of the gradients to achieve the empirical Rademacher complexity as a regret bound.

The question of sufficient statistics for online methods appears to be unexplored and poorly understood, and it will take significant effort to answer it. In this paper we propose an approach that appears to be general yet, inevitably, incomplete. We propose a definition that brings many existing methods under the same umbrella, and allows us to develop new efficient strategies that have otherwise been out of reach. The key workhorse for our development is the Burkholder method, studied in probability theory and harmonic analysis.

Beyond studying a notion of sufficient statistic for online methods, our work can be seen as providing further understanding of emerging connections between online learning, martingale inequalities, and deterministic geometric quantities. At the risk of being imprecise, let us describe the bird's-eye view of our overall approach:



Based on the definition of sufficient statistics for online methods, we first derive corresponding martingale inequalities with the help of the minimax theorem. We then turn to the Burkholder method, and show equivalence of these martingale inequalities for sufficient statistics and existence of a special Burkholder (or Bellman) function, a purely geometric object. We then use this function for the problem of online prediction, thus completing the circle. Crucially, the sufficient statistics we start with are reflected in the Burkholder function, and, hence, the proposed algorithm is only required to update these compressed representations of the data. We exhibit the power of this approach by deriving several new efficient prediction methods.

We remark that (Foster et al., 2017b) studied a particular case of the Burkholder method related to the UMD property for Banach spaces. The present work shows that the approach can be generalized significantly and used to address the question of sufficient statistics. For example, the explicit construction of the UMD-style Burkholder function for certain matrix prediction problems was noted to be challenging in (Foster et al., 2017b) and indeed does not appear to be known in the analysis community (Osękowski, 2017). In spite of this, the approach in the present paper uses different sufficient statistics to attain the same results with an explicit (and efficient) Burkholder function.

## 2. Problem Setup and Sufficient Statistics

Consider the *Online Supervised Learning* setting where, for each round $t = 1, \dots, n$, the forecaster observes side information $x_t \in \mathcal{X}$, makes a prediction $\widehat{y}_t \in \mathcal{Y} \subset \mathbb{R}$, observes an outcome $y_t \in \mathcal{Y}$, and incurs a loss of $\ell(\widehat{y}_t, y_t)$, where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. In a general form, the goal of the forecaster is to ensure that

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t)\right] \le \phi(x_1, y_1, \dots, x_n, y_n) \tag{1}$$

for any sequence $(x_1, y_1), \dots, (x_n, y_n)$, where the expectation is with respect to forecaster's randomization. The choice of $\phi$ models the problem at hand, and examples in this paper focus on

$$\phi(x_1, y_1, \dots, x_n, y_n) = \min_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \ell(f(x_t), y_t) + \mathcal{A}(f, x_1, \dots, x_n) \right\}, \tag{2}$$

for some class of functions $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ and an *adaptive bound* $\mathcal{A} : \mathcal{F} \times \mathcal{X}^n \to \mathbb{R}$. In this case, the difference of the cumulative losses of the forecaster and of $f \in \mathcal{F}$ is commonly referred to as *regret*,

$$\mathrm{Reg}_n(f) = \sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \ell(f(x_t), y_t).$$

We assume that $\phi$ is uniformly bounded over $(\mathcal{X} \times \mathcal{Y})^n$. We further assume that $\ell$ is convex and $L$-Lipschitz in the first argument over $\mathcal{Y}$. We denote the derivative (or a subderivative) of $\ell(\cdot, y)$ at $\widehat{y}$ by $\partial \ell(\widehat{y}, y) \in [-L, L]$. We will abbreviate $\delta_t = \partial \ell(\widehat{y}_t, y_t)$ when it is clear from context, but keep in mind that this value depends on the two variables $\widehat{y}_t$ and $y_t$. We assume that for any distribution $p$ on $\mathcal{Y}$, $\arg\min_{\widehat{y} \in \mathbb{R}} \mathbb{E}_{y \sim p}\, \ell(\widehat{y}, y) \in \mathcal{Y}$, and that $\mathcal{Y}$ is compact. We let $\Delta_{\mathcal{Y}}$ denote the space of all Borel probability measures on $\mathcal{Y}$ (more generally, $\Delta_A$ will denote the set of Borel probability measures over some set $A$). Since $\mathcal{Y}$ is compact, Prokhorov's theorem implies that $\Delta_{\mathcal{Y}}$ is compact in the weak topology. This enables application of the minimax theorem as in previous works in this direction (Rakhlin et al., 2010, 2015; Foster et al., 2015).

**Additional notation** Given a function $f : S \to \mathbb{R}$, its Fenchel dual $f^\star$ is defined via $f^\star(w) = \sup_{x \in S}\{\langle w, x \rangle - f(x)\}$. For any norm $\|\cdot\|$, the dual norm will be denoted by $\|\cdot\|_\star$. $\mathsf{B}_p^d$ will denote the $d$-dimension unit $\ell_p$ ball and the shorthand $\Delta_d$ will denote the simplex in $d$ dimensions. For any interval $[a, b]$, we let $\mathrm{proj}_{[a,b]}(x) = \min\{b, \max\{a, x\}\}$.

### 2.1. Sufficient Statistics

Since there is no probabilistic model for data in the online learning setting, the notion of "sufficiency" has to be tied to the particular choice of $\phi$. It is then tempting to define a sufficient statistic as a "compressed" representation which may be used by some strategy to ensure (1). While natural, such a definition does not provide any additional structure to narrow the search for an algorithm.

The definition we propose is as follows:

**Definition 1** *Let $\mathcal{T}$ be some vector space. A function $\mathbf{T} : \mathcal{X} \times \mathcal{Y} \times [-L, L] \to \mathcal{T}$ is an* additive sufficient statistic *for $\phi$ if there exists $V : \mathcal{T} \to \mathbb{R}$ such that*

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \phi(x_1, y_1, \dots, x_n, y_n) \le V\left( \sum_{t=1}^{n} \mathbf{T}(x_t, \widehat{y}_t, \partial \ell(\widehat{y}_t, y_t)) \right) \tag{3}$$

*for any sequence $x_1, \widehat{y}_1, y_1, \dots, x_n, \widehat{y}_n, y_n$. We refer to $(\mathbf{T}, V)$ as a* sufficient statistic pair.

In Appendix A, we consider a more general non-additive definition. All examples in this paper, however, are already covered by Definition 1, and we will drop the word "additive" for now. We will also make the mild assumption that there exists $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathbf{T}(x^0, y^0, 0) = 0 \in \mathcal{T}$.

**Example 1 (Prediction with expert advice)** *Consider $\phi$ as in Eq. (2) with $\mathcal{F}$ as the set of linear functions $f(x) = \langle f, x \rangle$ for $f \in \Delta_d$, with $\mathcal{X} = [-1,1]^d$, and with non-adaptive rate $\mathcal{A} := c\sqrt{n \log d}$. Then the left-hand-side of (3) can be upper bounded via linearization of the convex loss by*

$$\max_{j \in 1,\dots,d} \sum_{t=1}^{n} \partial \ell(\widehat{y}_t, y_t) \cdot (\widehat{y}_t - \langle e_j, x_t \rangle) - c\sqrt{n \log d}.$$

*It follows that $\mathbb{R}^d$-valued map $\mathbf{T}$ defined by $[\mathbf{T}(x_t, \widehat{y}_t, \delta_t)]_j = \delta_t \cdot (\widehat{y}_t - \langle e_j, x_t \rangle)$ is a sufficient statistic.*

**Example 2 (Adaptive Gradient Descent)** *Consider $\phi$ as in Eq. (2) with $\mathcal{F}$ as the set of linear functions $f(x) = \langle f, x \rangle$ for $f \in \mathsf{B}_2^d$, $\mathcal{X} = \mathbb{R}^d$, and adaptive bound $\mathcal{A}(\nabla_1, \dots \nabla_n) := (\sum_{t=1}^n \|\nabla_t\|^2)^{1/2}$, where $\nabla_t := \delta_t x_t$. The left-hand-side of (3) is at most*

$$\max_{f \in \mathsf{B}_2^d} \sum_{t=1}^{n} \delta_t \cdot (\widehat{y}_t - \langle f, x_t \rangle) - (\sum_{t=1}^{n} \|\nabla_t\|^2)^{1/2} = \sum_{t=1}^{n} \delta_t \cdot \widehat{y}_t + \left\| \sum_{t=1}^{n} \nabla_t \right\| - (\sum_{t=1}^{n} \|\nabla_t\|^2)^{1/2}. \tag{4}$$

*This implies that $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = \left(\delta_t \widehat{y}_t, \nabla_t, \|\nabla_t\|^2\right) \in \mathbb{R} \times \mathcal{X} \times \mathbb{R}$ is a sufficient statistic.*

## 3. Martingale Inequalities and the Burkholder Method

The notion of sufficient statistic introduced in the previous section will only be useful if we exhibit a prediction strategy employing this representation. Before doing so, we need to build the two bridges outlined in the diagram on the previous page. These are Lemma 2 and Lemma 3 below.

First, we show that existence of a prediction strategy that guarantees the regret inequality (1) for all sequences can be ensured by checking a martingale inequality involving only the sufficient statistics. The key tool in proving the lemma is the minimax theorem.

Note that in a slight abuse of notation, we will concatenate the first two arguments of any sufficient statistic $\mathbf{T}$ and write them as $z_t := (x_t, \widehat{y}_t)$ going forward.

**Lemma 2** *Suppose $(\mathbf{T}, V)$ is a sufficient statistic pair for $\phi$. Let $\delta = (\delta_1, \dots, \delta_n)$ be a $[-L, L]$-valued martingale difference sequence (i.e. $\mathbb{E}[\delta_t \mid \mathcal{G}_{t-1}] = 0$, where $\mathcal{G}_{t-1} = \sigma(\delta_1, \dots, \delta_{t-1})$). Let $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)$ be a sequence of functions $\boldsymbol{z}_t : [-L, L]^{t-1} \to \mathcal{X} \times \mathcal{Y}$, each viewed as a predictable process with respect to $\mathcal{G}_{t-1}$. Then a sufficient condition for existence of a prediction strategy such that (1) holds for all sequences $(x_1, y_1), \dots, (x_n, y_n)$ is that*

$$\mathbb{E}\left[V\left(\sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t, \delta_t)\right)\right] \le 0 \tag{5}$$

*holds for any $\boldsymbol{z}$ and any law of $\delta$. Moreover, when $\alpha \mapsto V(\tau + \mathbf{T}(z, \alpha))$ is convex for any $z \in \mathcal{X} \times \mathcal{Y}, \tau \in \mathcal{T}$, it is enough to check (5) for $\delta_t = \epsilon_t \cdot 2L$, $\forall t = 1, \dots, n$, where $\epsilon_t$s are independent Rademacher random variables.*

Lemma 2 is in the spirit of results in (Rakhlin et al., 2010, 2015; Foster et al., 2015) whereby existence of a strategy (or, "learnability") is certified non-constructively by proving a martingale inequality.

The next lemma provides a key insight into existence of certain deterministic functions with "geometric" properties (in particular, *restricted concavity*) and can be seen as a variation on the so-called *Burkholder method* (also sometimes called the *Bellman function method*; see (Osękowski, 2012) for the detailed treatment and examples).

**Lemma 3** *Let $\delta = (\delta_1, \dots, \delta_n)$ be a $[-L, L]$-valued martingale difference sequence with joint law $\boldsymbol{p}$ and let $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)$ be a predictable process ($\boldsymbol{z}_t : [-L, L]^{t-1} \to \mathcal{X} \times \mathcal{Y}$) with respect to $\mathcal{G}_{t-1} = \sigma(\delta_1, \dots, \delta_{t-1})$. The probabilistic inequality*

$$\mathbb{E}\left[V\left(\sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t, \delta_t)\right)\right] \le 0 \tag{6}$$

*holds for any $n \ge 1$, $\boldsymbol{z}$, and $\boldsymbol{p}$* if and only if *one can find a function $\mathbf{U} : \mathcal{T} \to \mathbb{R}$ that satisfies the following three properties:*

$1^o$ $\mathbf{U}(0) \le 0$.

$2^o$ *For any $\tau \in \mathcal{T}$, $\mathbf{U}(\tau) \ge V(\tau)$.*

$3^o$ *For any $\tau \in \mathcal{T}$, $z \in \mathcal{X} \times \mathcal{Y}$, and any mean-zero distribution $p$ on $[-L, L]$,*

$$\mathbb{E}_{\alpha \sim p}\left[\mathbf{U}(\tau + \mathbf{T}(z, \alpha))\right] \le \mathbf{U}(\tau). \tag*{(restricted concavity)}$$

*Furthermore, if for any $\tau \in \mathcal{T}$ and $z \in \mathcal{X} \times \mathcal{Y}$ the mapping $\alpha \mapsto V(\tau + \mathbf{T}(z, \alpha))$ is convex, then the condition (6) is implied by $\mathbb{E}\left[V\left(\sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t, \epsilon_t \cdot 2L)\right)\right] \le 0$, where $(\epsilon_1, \dots, \epsilon_n)$ are Rademacher random variables. For this new condition, property $3^o$ is replaced by*

$3'$ *The mapping $\alpha \mapsto \mathbf{U}(\tau + \mathbf{T}(z, \alpha))$ is convex and:*

$$\forall \tau \in \mathcal{T}, z \in \mathcal{X} \times \mathcal{Y}, \quad \mathbb{E}_\epsilon \mathbf{U}(\tau + \mathbf{T}(z, \epsilon \cdot 2L)) \le \mathbf{U}(\tau),$$

*where $\epsilon$ is a Rademacher random variable.*

**Definition 4** *We call any function $\mathbf{U}$ satsifying the properties $1^o$, $2^o$, and $3^o$/$3'$ a* Burkholder function for $(\mathbf{T}, V)$.

In plain language, the lemma says that one can prove a certain probabilistic inequality if and only if there is a deterministic function with certain properties. The proof of the lemma, in fact, provides a construction for the "optimal" function $\mathbf{U}$, but it is not clear how to directly evaluate the optimal function efficiently (see Appendix A for a discussion of the computational prospects of automating this process).

We remark that the Burkholder functions guaranteed by the lemma are not unique, and some may be easier to find than others. We also note that any Burkholder function $\mathbf{U}$ for $(\mathbf{T}, V)$ yields another sufficient statistic pair $(\mathbf{T}, \mathbf{U})$ guaranteeing the same regret bound. The power of Lemma 3 is to guarantee the existence of a function $\mathbf{U}$ satisfying property $3^o$ when the function $V$ under consideration does not have these properties. This situation, where the choice of $V$ is "obvious" but the discovery of $\mathbf{U}$ requires nontrivial analysis, occurs frequently when one attempts to design adaptive algorithms for a new task.

To showcase the power of this lemma, we consider a particular martingale inequality that gives rise to the geometric notions of strong convexity and smoothness. These geometric properties are extensively employed in Online Convex Optimization: to instantiate the Mirror Descent algorithm with a given norm, one needs to exhibit a function that is strongly convex with respect to a given norm of interest. For example, for the $\ell_1$ norm a standard choice is the negative entropy function. The next example shows that for any norm, the optimal strongly convex function is precisely the dual of the special Burkholder function for a particular martingale inequality. This example is the focus of Pisier (1975), yet for us it is one point on the spectrum of sufficient statistics.

**Example 3 (Smoothness and Strong Convexity)** *Assume $L = 1$ for brevity. Suppose $\mathcal{X} = \mathbb{R}^d$ (more generally, we may take $\mathcal{X}$ to be a Banach space), equipped with a norm $\|\cdot\|$. Let $V : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ be defined by $(x, a) \mapsto \|x\|^2 - C \cdot a$ for $C > 0$. Take $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = (\delta_t x_t, \|x_t\|^2)$. Since $\alpha \mapsto V(\tau + \mathbf{T}(x_t, \widehat{y}_t, \alpha))$ is convex, it is enough to consider (6) for independent Rademacher random variables. The martingale inequality (6) then reads*

$$\mathbb{E}\left[\left\|\sum_{t=1}^{n} \epsilon_t \boldsymbol{x}_t\right\|^2 - C\sum_{t=1}^{n}\|\boldsymbol{x}_t\|^2\right] \leq 0 \tag{7}$$

*for any $\mathcal{X}$-valued predictable process $(\boldsymbol{x}_t)$ with respect to the dyadic filtration $\mathcal{F}_{t-1} = \sigma(\epsilon_1, \ldots, \epsilon_{t-1})$. If (7) holds, Lemma 3 guarantees existence of a Burkholder function $\mathbf{U}$, and property $3'$ reads*

$$\mathbb{E}_\epsilon \mathbf{U}(\tau_1 + \epsilon x, \tau_2 + \|x\|^2) \leq \mathbf{U}(\tau_1, \tau_2),$$

*for any $\tau = (\tau_1, \tau_2) \in \mathcal{X} \times \mathbb{R}$ and $x \in \mathcal{X}$. From the construction of $\mathbf{U}$ in the proof of Lemma 3, with our particular choice of $V$, one can deduce that $\mathbf{U}(\tau_1, \tau_2) = \mathbf{U}(\tau_1, 0) + \tau_2$. Hence,*

$$\frac{1}{2}\left(\mathbf{U}(\tau_1 + x, 0) + \mathbf{U}(\tau_1 - x, 0)\right) + C\|x\|^2 = \frac{1}{2}\left(\mathbf{U}(\tau_1 + x, C\|x\|^2) + \mathbf{U}(\tau_1 - x, C\|x\|^2)\right) \leq \mathbf{U}(\tau_1, 0)$$

*and, thus, $x \mapsto \mathbf{U}(x, 0)$ is smooth with respect to the norm and its dual is strongly convex with respect to $\|\cdot\|_\star$. In summary, the Burkholder method captures the geometry necessary for defining Gradient-Descent-style methods, as the dual of $\mathbf{U}(x, 0)$ provides the universal construction for a strongly convex function with respect to a given norm. See Srebro et al. (2011) for an in-depth treatment of Mirror Descent and universal construction of strongly convex regularizers.*

What should an algorithm designer take away from the developments thus far? Let us provide a brief summary. One first starts with a desired regret inequality for the online learning setting, such as (1). The next step is to find an upper bound on the regret inequality that can be expressed in terms of additive sufficient statistics. Lemma 2 and Lemma 3 then guarantee, respectively, that there is a certain martingale inequality that must hold if the upper bound in terms of sufficient statistics is achievable, and that there must exist a Burkholder function with certain geometric properties. In the next section we close the loop by showing that whenever such a Burkholder function can be evaluated efficiently, it yields an efficient algorithm that only keeps the sufficient statistics in memory.

Before proceeding, we briefly remark that if the sufficient statistic expansion $V$ also serves as a lower bound on the regret inequality, then there is a formal sense in which the special Burkholder function exists if and only if there exists a strategy achieving the original regret inequality of interest; this is the focus of Appendix C. In the reverse direction, one may start with a probabilistic inequality and determine the statistics that should be used to define the online prediction goal.[1]

## 4. The Burkholder Algorithm

Example 3 in the previous section already suggests that the Burkholder $\mathbf{U}$ functions capture the "geometry" needed for forming online predictions. Indeed, the method applies to settings in which more complicated sufficient statistics (beyond the norm of the sum and the sum of the squared norms) are necessary. We now define a "universal" algorithm for online prediction based on $\mathbf{U}$.

To define the algorithm, first let $\zeta_{t-1} = \sum_{j=1}^{t-1} \mathbf{T}(x_j, \widehat{y}_j, \delta_j)$ be the cumulative value of the sufficient statistic computed after $t-1$ rounds. Since $\mathcal{T}$ is a vector space, $\zeta_t$s are elements of $\mathcal{T}$, and this is the only information the algorithm stores in memory.

The *Burkholder algorithm* is defined by the update:

$$\textbf{Compute } q_t = \underset{q\in\Delta_{\mathcal{Y}}}{\arg\min} \sup_{y\in\mathcal{Y}} \mathbb{E}_{\widehat{y}\sim q} \mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}, \partial\ell(\widehat{y}, y))\Big). \qquad \textbf{Sample } \widehat{y}_t \sim q_t. \tag{8}$$

**Lemma 5** *For a sufficient statistic pair* $(\mathbf{T}, V)$*, if there exists a Burkholder function* $\mathbf{U}$ *satisfying Properties* $1^o$*,* $2^o$*, and* $3^o$ *(or* $3'$*) of Lemma 3, then the Burkholder algorithm (8) obtains the regret bound (1) in expectation for all sequences* $(x_1, y_1), \ldots, (x_n, y_n)$*.*

**Proof** To check that the above strategy works, fix a value $x_t$ and observe that by the minimax theorem,[2]

$$\inf_{q\in\Delta_{\mathcal{Y}}} \sup_{y\in\mathcal{Y}} \mathbb{E}_{\widehat{y}\sim q\in\Delta_{\mathcal{Y}}} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}, \partial\ell(\widehat{y}, y))\right) = \sup_{p\in\Delta_{\mathcal{Y}}} \inf_{\widehat{y}\in\mathcal{Y}} \mathbb{E}_{y\sim p} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}, \partial\ell(\widehat{y}, y))\right)$$

For any fixed $p$, let $\widehat{y}^\star := \arg\min_{\widehat{y}\in\mathcal{Y}} \mathbb{E}_{y\sim p}\, \ell(\widehat{y}, y)$, which implies $\partial\ell(\widehat{y}^\star, y)$ is a mean-zero variable (see the proof of Lemma 2). Taking the worst case value for $p$ and choosing $\widehat{y}^\star$ as the learner's strategy for each $p$ yields an upper bound of $\sup_{p\in\Delta_{\mathcal{Y}}} \mathbb{E}_{y\sim p} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}^\star, \partial\ell(\widehat{y}^\star, y))\right)$, which in turn is upper bounded by

$$\sup_{\widehat{y}^\star\in\mathcal{Y}} \sup_{p\in\Delta_{[-L,L]}\,:\,\mathbb{E}_{\alpha\sim p}[\alpha]=0} \mathbb{E}_{\alpha\sim p} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}^\star, \alpha)\right)$$

by observing that the distribution over $\partial\ell(\widehat{y}^\star, y)$ belongs to the set of all zero-mean distributions supported on $[-L, L]$. The third property of $\mathbf{U}$ now leads to the upper bound,

$$\sup_{\widehat{y}^\star\in\mathcal{Y}} \sup_{p\in\Delta_{[-L,L]}\,:\,\mathbb{E}_{\alpha\sim p}[\alpha]=0} \mathbb{E}_{\alpha\sim p} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}^\star, \alpha)\right) \le \mathbf{U}\left(\zeta_{t-1}\right).$$

Applying this argument from $t = n$ down to $t = 0$ yields the value $\mathbf{U}(0) \le 0$. ■

[1] This was precisely the approach used to develop a matrix prediction method we present in Section 5.

[2] The minimax theorem can be applied because $\Delta_{\mathcal{Y}}$ is compact; see discussion in the proof of Lemma 2.

We remark that the approach presented here extends beyond the relaxation framework of (Rakhlin et al., 2012). In particular, the present approach can handle recursions which cannot be written in the form "$\ell(\widehat{y}_t, y_t) + \mathrm{Rel}(x_{1:t}, y_{1:t})$", e.g. when the potential function depends on past forecasts $(\widehat{y}_t)$.

**Implementation** When $\mathbf{U}$ is convex in $\widehat{y}$ and the set $\mathcal{Y}$ is convex, the minimum over $q$ is achieved at a deterministic strategy, and so the minimization problem simplifies to $\arg\min_{\widehat{y}\in\mathcal{Y}}$. All of the Burkholder functions we explore in this paper enjoy this or similar simplified and efficient representations for the algorithm. These simplifications are detailed in Appendix E. Even without convexity, the general form for the Burkholder algorithm in (8) can be implemented efficiently via convex programming, assuming only Lipschitz continuity of $\mathbf{U}$.

**Proposition 6** *Suppose $\mathbf{U}$ is Lipschitz and bounded and can be evaluated in constant time. Then (8) can be implemented approximately so as to achieve the regret inequality (1) up to additive constants in time poly$(n)$.*

See Proposition 20 in the appendix for a precise version of this statement.

## 5. Example: Matrix Prediction

In this section we focus on linear matrix prediction problems. The side information $x_t$ is now matrix-valued, and we shall denote it by a capital letter $X_t \in \mathbb{R}^{d_1\times d_2}$. Our goal is to achieve a regret inequality as in (2) with a class $\mathcal{F} = \{X \mapsto \langle W, X\rangle \mid W \in \mathcal{W}\}$, where $\mathcal{W} = \left\{W \in \mathbb{R}^{d_1\times d_2} \mid \|W\|_{\Sigma} \le r\right\}$. Here $\langle A, B\rangle = \mathrm{tr}(AB^\top)$ is the standard matrix inner product and $\|\cdot\|_{\Sigma}$ denotes the nuclear norm. We also let $\|\cdot\|_{\sigma}$ denote the spectral norm. As before, the loss $\ell$ is assumed to be $L$-Lipschitz and regret against a matrix $W \in \mathcal{W}$ is given by $\mathrm{Reg}_n(W) \coloneqq \sum_{t=1}^n \ell(\widehat{y}_t, y_t) - \ell(\langle W, X_t\rangle, y_t)$.

In a search for an adaptive bound on regret, we inspect the adaptive bound (4) for the vector case. The direct analogue for matrices would be a bound proportional to $\left(\sum_{t=1}^n \|X_t\|_{\sigma}^2\right)^{1/2}$, and indeed such a bound is possible with Matrix Exponential Weights (Hazan et al., 2012, Theorem 13).[3] However, the matrix version of Khintchine inequality, as well as matrix deviation inequalities, involve—for the case of random centered self-adjoint matrices—the tighter quantity $\left\|\sum_{t=1}^n X_t^2\right\|_{\sigma}^{1/2}$ (see (Tropp, 2012; Mackey et al., 2014)). Given the correspondence between online regret bounds and martingale inequalities, one may wonder if there is an algorithm that achieves this adaptive bound. We shall exhibit such a method using our approach, and the reader might already guess that $\sum_{t=1}^n X_t^2$ should be part of the sufficient statistic for the online algorithm. This is indeed the case, though we present results for general non-square matrices.

Let $\mathbb{S}^d$ denote the set of symmetric matrices in $\mathbb{R}^{d\times d}$, $\mathbb{S}_+^d$ denote the set of positive-semidefinite matrices, and $\mathbb{S}_{++}^d$ denote the set of positive-definite matrices. For $X \in \mathbb{S}^d$ we let $\lambda(X) \in \mathbb{R}^d$ denote its eigenvalues arranged in decreasing order, so that $\lambda_1(X)$ is the largest eigenvalue. For any matrix $X \in \mathbb{R}^{d_1\times d_2}$ we define its Hermitian dilation $\mathcal{H}(X) \in \mathbb{S}^{d_1+d_2}$ and $\mathcal{M}(X) \in \mathbb{S}^{d_1+d_2}$ as:

$$\mathcal{H}(X) = \begin{pmatrix} 0 & X \\ X^\top & 0 \end{pmatrix} \qquad \mathcal{M}(X) = \mathcal{H}(X)^2 = \begin{pmatrix} XX^\top & 0 \\ 0 & X^\top X \end{pmatrix}. \tag{9}$$

It is well-known that for any matrix $X$, $\lambda_1(\mathcal{H}(X)) = \|X\|_{\sigma}$.

[3] With more work it is possible to obtain a bound of $\left(\max_t \|X_t\|_{\sigma} \cdot \|\sum_{t=1}^n X_t\|_{\sigma}\right)^{1/2}$; this is still weaker than our result, and seems to only be possible when the constraint set and $X_t$s are restricted to be positive-semidefinite.

With these definitions in place, the desired adaptive regret bound takes the form

$$\mathcal{A}_\eta(X_1,\ldots,X_n) = \frac{\eta r L^2}{2}\left\|\sum_{t=1}^n \mathcal{M}(X_t)\right\|_\sigma + \frac{c}{\eta} \tag{10}$$

for some fixed $\eta > 0$ and constant $c > 0$. The sufficient statistic takes values in $\mathcal{T} = \mathbb{R}\times\mathbb{S}^{d_1+d_2}\times\mathbb{S}_+^{d_1+d_2}$ and incorporates the matrix variance terms $\mathcal{M}(X_t)$.

**Proposition 7** *The pair* $(\mathbf{T}, V)$ *defined via* $\mathbf{T}(X_t,\widehat{y}_t,\delta_t) = (\delta_t\cdot\widehat{y}_t, \delta_t\cdot\mathcal{H}(X_t), \mathcal{M}(X_t)) \in \mathbb{R}\times\mathbb{S}^{d_1+d_2}\times\mathbb{S}_+^{d_1+d_2}$ *and*

$$V(a,H,M) = a + r\lambda_1\big(H - \tfrac{1}{2}\eta L^2 M\big) - \frac{c}{\eta}, \tag{11}$$

*form a sufficient statistic pair for the adaptive regret bound* $\mathcal{A}_\eta$.

Now that we proposed a sufficient statistic, Lemma 2 and Lemma 3 give a specific form for a martingale inequality and a construction for the special function (if the martingale inequality holds). Since the function constructed in the proof of Lemma 3 may not be efficiently computable, we embark on a search for a function that *can* be evaluated efficiently. The next theorem presents such a Burkholder function. The proof rests on Lieb's Concavity Theorem (Lieb, 1973), which states that for any fixed $A \in \mathbb{S}^d$, the function $X \mapsto \operatorname{tr}\exp(A + \log X)$ is concave over $\mathbb{S}^d_{++}$.

**Theorem 8** *Define* $\mathbf{U} : \mathbb{R}\times\mathbb{S}^{d_1+d_2}\times\mathbb{S}_+^{d_1+d_2} \to \mathbb{R}$ *via*

$$\mathbf{U}(a,H,M) = a + \frac{r}{\eta}\log\operatorname{tr}\exp\big(\eta H - \tfrac{1}{2}\eta^2 L^2 M\big) - \frac{c}{\eta}.$$

*Then* $\mathbf{U}$ *is a Burkholder function, for the pair* $(\mathbf{T}, V)$ *in* (11) *when* $c \geq r\log(d_1+d_2)$.

This Burkholder function construction immediately implies both existence of a prediction strategy (via Lemma 5) and that a probabilistic inequality for matrix-values martingales holds. We will present both in detail. The matrix prediction strategy granted by the Burkholder algorithm is particularly simple due to extra convexity properties of $\mathbf{U}$; see Appendix E.

**Corollary 9 (Matrix prediction algorithm)** *Suppose that* $\mathcal{Y} = [-B, B]$ *for some* $B > 0$. *Then the deterministic strategy*

$$\widehat{y}_t = \operatorname{proj}_{[-B,B]}\left(-\frac{r}{L\eta}\mathbb{E}_{\sigma\in\{\pm1\}}\left[\sigma\log\operatorname{tr}\exp\left(\eta\sigma L\mathcal{H}(X_t) + \eta\sum_{s=1}^{t-1}\delta_s\mathcal{H}(X_s) - \tfrac{1}{2}\eta^2L^2\sum_{s=1}^{t}\mathcal{M}(X_s)\right)\right]\right) \tag{12}$$

*leads to a regret bound of*

$$\sum_{t=1}^n \ell(\widehat{y}_t, y_t) - \inf_{W\in\mathcal{W}}\sum_{t=1}^n \ell(\langle W, X_t\rangle, y_t) \leq \tfrac{1}{2}\eta L^2 r\left\|\sum_{t=1}^n\mathcal{M}(X_t)\right\|_\sigma + \frac{r\log(d_1+d_2)}{\eta}.$$

Since this regret bound is monotonically increasing with time, it is easy to tune $\eta$ to obtain a fully adaptive strategy.

**Proposition 10** *Let $R = \max_t \|X_t\|_\sigma$ be known. By tuning $\eta$ through the standard doubling trick, we arrive at a regret bound of*

$$\sum_{t=1}^n \ell(\widehat{y}_t, y_t) - \inf_{W\in\mathcal{W}} \sum_{t=1}^n \ell(\langle W, X_t\rangle, y_t)$$
$$\leq O\left(r\sqrt{\max\left\{\left\|\sum_{t=1}^n X_t X_t^\top\right\|_\sigma, \left\|\sum_{t=1}^n X_t^\top X_t\right\|_\sigma\right\}\log(d_1+d_2)} + Rr\log(n)\right).$$

Let us briefly discuss the result. First, the computation in (12) involves an SVD, and does not scale with $t$ since the method only keeps in memory the cumulative statistics. The regret bound gives a *sequence-optimal* rate for the problem of *Online Matrix Completion*, where each $X_t$ is an indicator $e_{i_t}e_{j_t}^\top$ corresponding to—for example—a user-movie pair for which the learner must predict a score. Here the regret bound obtained by (12) interpolates between the worst-case configuration of the entries $(i_t, j_t)$ and "spread-out" (e.g. uniform) sampling of the entries. The result improves on (Foster et al., 2017b), which showed that this type of bound is possible by invoking the UMD inequality for Schatten norms but did not provide an efficient algorithm. See that paper for further discussion of the setting and problem.

We now deliver on the second promise, namely a probabilistic martingale inequality. This inequality is stated for $\mathbb{R}^{d_1+d_2}$-valued Paley-Walsh martingale difference sequences $(\epsilon_t \boldsymbol{X}_t(\epsilon))_{t\leq n}$, where each $\boldsymbol{X}_t(\epsilon) = \boldsymbol{X}_t(\epsilon_1, \ldots, \epsilon_{t-1})$ is predictable with respect to $\mathcal{F}_{t-1} = \sigma(\epsilon_1, \ldots, \epsilon_{t-1})$ for Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$.

**Corollary 11 (Martingale Matrix Square Function Inequality)** *For all Paley-Walsh martingale difference sequences $(\epsilon_t \boldsymbol{X}_t(\epsilon))_{t\leq n}$ it holds that*

$$\mathbb{E}_\epsilon\left\|\sum_{t=1}^n \epsilon_t \boldsymbol{X}_t(\epsilon)\right\|_\sigma \leq \sqrt{2\,\mathbb{E}_\epsilon \max\left\{\left\|\sum_{t=1}^n \boldsymbol{X}_t(\epsilon)\boldsymbol{X}_t(\epsilon)^\top\right\|_\sigma, \left\|\sum_{t=1}^n \boldsymbol{X}_t(\epsilon)^\top \boldsymbol{X}_t(\epsilon)\right\|_\sigma\right\}\log(d_1+d_2)}. \tag{13}$$

In the special case where $\boldsymbol{X}_t(\epsilon) = X_t$ is a fixed sequence, this square function inequality (13) recovers the Matrix Khintchine inequality (Mackey et al., 2014), including constants. A similar martingale inequality can be obtained from the Matrix Freedman/Bennett inequalities of Tropp (2011), but this will depend on almost sure bounds on spectral norms of $(\boldsymbol{X}_t(\epsilon))_{t\leq n}$.

## 6. Further Examples

### 6.1. ZigZag Algorithm and the UMD Property

Pisier (1975) used martingale techniques to provide a characterization of super-reflexive Banach spaces as those admitting an equivalent uniformly convex norm. As already described in Example 3, the essential ingredient of this analysis is a construction of a function $\mathbf{U}$ with the desired restricted concavity property (which turns out to be equivalent to uniform smoothness) for the martingale inequality (7). The corresponding notion in the world of online learning is that of an adaptive gradient (or mirror) descent.

Burkholder (1981) provided a geometrical characterization of UMD spaces, and a key ingredient of the approach was to establish existence of (and sometimes to compute in closed form) the function

$\mathbf{U}$ with corresponding geometric properties ($\zeta$-convexity, which is equivalent to "zigzag concavity" (Osękowski, 2012)). As shown in (Foster et al., 2017b), in the online learning world the corresponding adaptive regret bound is that of empirical Rademacher averages:

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \min_{\|w\| \le 1} \sum_{t=1}^{n} \ell(\langle w, x_t \rangle, y_t) - C\,\mathbb{E}\left\| \sum_{t=1}^{n} \epsilon_t \delta_t x_t \right\|.$$

By linearizing the loss, it suffices to use the sufficient statistic $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = (\delta_t \widehat{y}_t, \delta_t x_t, \epsilon_t x_t)$ where $(\epsilon_t)$ is taken to be a sequence drawn by the algorithm. The corresponding martingale inequality is

$$\mathbb{E}\left[ \left\| \sum_{t=1}^{n} \epsilon_t \boldsymbol{x}_t(\epsilon) \right\|^p - C \left\| \sum_{t=1}^{n} \epsilon'_t \boldsymbol{x}_t(\epsilon) \right\|^p \right] \le 0, \tag{14}$$

where the process in the subtracted term is decoupled and $p > 1$ is arbitrary. We refer the reader to (Foster et al., 2017b) for more details.

We would like to emphasize that both smoothness/strong convexity (as in Pisier's work) and the UMD property (as in Burkholder's work) are two distinct notions with distinct sets of sufficient statistics. Since the fundamental works of Pisier and Burkholder, the so-called "Burkholder method" has been employed to prove a wide range of martingale inequalities and discover the corresponding geometric properties of the special function (Osękowski, 2012; Hytönen et al., 2016). The goal of this paper is to present a unifying approach for working with arbitrary sufficient statistics in online learning, and to show that the Burkholder approach is in fact *algorithmic*.

### 6.2. AdaGrad and Square Function Inequalities

The Burkholder method can be used to recover efficient algorithms that obtain regret bounds in the vein of diagonal AdaGrad and full-matrix AdaGrad (Duchi et al., 2011), with optimal constants. We thank Adam Osękowski for suggesting this example to us (Osękowski, 2017).

Define a function $\mathbf{U}_{\text{square}}(x, y) : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}$ (Osękowski, 2005, 2012) via

$$\mathbf{U}_{\text{square}}(x, y) = \begin{cases} -\sqrt{2y^2 - \|x\|_2^2}, & y \ge \|x\|_2. \\ \|x\|_2 - 2y, & y < \|x\|_2. \end{cases}$$

$\mathbf{U}_{\text{square}}$ satisfies three properties as in Lemma 3: **1.** $\mathbf{U}_{\text{square}}(x, y) \ge \|x\|_2 - 2y$, **2.** $\mathbf{U}_{\text{square}}(x, \|x\|_2) \le 0$, and **3.** $\mathbf{U}_{\text{square}}(x + d, \sqrt{y^2 + \|d\|_2^2}) \le \mathbf{U}_{\text{square}}(x, y) + \langle \partial_x \mathbf{U}_{\text{square}}(x, y), d \rangle$. This function consequently leads to two algorithms in the style of AdaGrad (Duchi et al., 2011) but with optimal constants, and which we now sketch.

The first regret bound is for $\ell_2$ classes, as in full-matrix AdaGrad, and has the form

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \min_{\|w\|_2 \le 1} \sum_{t=1}^{n} \ell(\langle w, x_t \rangle, y_t) - 2L \sqrt{\sum_{t=1}^{n} \|x_t\|_2^2} \le 0.$$

The associated martingale inequality is $\mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t \boldsymbol{x}_t(\epsilon)\right\|_2 \le 2\,\mathbb{E}\sqrt{\sum_{t=1}^{n} \|\boldsymbol{x}_t(\epsilon)\|_2^2}$, which was shown to be optimal in Osękowski (2005).[4] The second regret bound is for $\ell_\infty$ classes, as in diagonal

[4] Note that the expectation is outside the square root, so this is stronger than the ubiquitous inequality $\mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t \boldsymbol{x}_t(\epsilon)\right\|_2 \le \sqrt{\mathbb{E}\sum_{t=1}^{n} \|\boldsymbol{x}_t(\epsilon)\|_2^2}$.

AdaGrad, and has the form

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \min_{\|w\|_\infty \le 1} \sum_{t=1}^{n} \ell(\langle w, x_t \rangle, y_t) - 2L \left\| \left( \sum_{t=1}^{n} x_t^2 \right)^{1/2} \right\|_1 \le 0,$$

where $x_t^2$ denotes the element-wise square. This is obtained by applying the scalar version of $\mathbf{U}_{\text{square}}$ coordinate-wise. The associated martingale inequality is $\mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t \boldsymbol{x}_t(\epsilon)\right\|_1 \le 2\,\mathbb{E}\left\|\left(\sum_{t=1}^{n} \boldsymbol{x}_t(\epsilon)^2\right)^{1/2}\right\|_1$. Both regret bounds require no prior knowledge of the range of $(x_t)_{t\le n}$.

### 6.3. Strongly Convex Losses

In this section we take $\mathcal{F} = \left\{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d\right\}$ and equip this space with a regularizer $\Phi(w) = \frac{1}{2}\|w\|_2^2$. We assume that the loss $\ell(\widehat{y}, y)$ is $\rho$-strongly convex and $L$-Lipschitz. We adopt the shorthand $z_t = (x_t, -\widehat{y}_t)$, and our goal is to obtain a data- and comparator- dependent regret bound of the form

$$\mathcal{A}_\lambda(w; z_1, \ldots, z_n) = \Phi((w,1)) + c \log \det\left(\rho \sum_{t=1}^{n} z_t z_t^\top + \lambda I\right) - c \log \det(\lambda I).$$

for some $c > 0$. Here we recover the classical Vovk-Azoury-Warmuth-type bound for strongly convex losses (Vovk, 1998; Azoury and Warmuth, 2001). This example is important because it shows that the Burkholder method in full generality can both obtain fast rates for curved losses and obtain bounds that jointly depend on the comparator and data; the UMD-type Burkholder functions used in Foster et al. (2017b) do not obtain such results. The right sufficient statistic for this problem should be familiar: In addition to storing a sum of gradients, we also store the empirical covariance $\sum_{t=1}^{n} z_t z_t^\top$. We introduce one last piece of notation: For $A \ge 0$, $\Psi_A(w) = \frac{1}{2}\langle w, Aw\rangle$.

**Proposition 12** *The sufficient statistic* $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = (\delta_t z_t, z_t z_t^\top) \in \mathbb{R}^{d+1} \times \mathbb{S}_+^{d+1}$ *and*

$$V(x, A) = \Psi^\star_{\rho A + \lambda I}(x) - c \log(\det(\rho A + \lambda I)/\det(\lambda I)) \tag{15}$$

*forms a sufficient statistic pair for the adaptive regret bound* $\mathcal{A}_\lambda$.

**Theorem 13** *For the sufficient statistic pair* $(\mathbf{T}, V)$ *in Proposition 12,* $\mathbf{U} = V$ *is a Burkholder function whenever* $c \ge L^2/\rho$.

Note that for this setting the natural choice for $V$ turned out to be a Burkholder function itself.

## Discussion

Due to space constraints the following additional results have been deferred to the appendix: Discussion of further directions (Appendix A), algorithms for parameter-free online learning in Banach spaces (Appendix B), and necessary conditions for existence of Burkholder functions (Appendix C).

## Acknowledgements

We thank Adam Osękowski for helpful discussions and for suggesting the example in Section 6.2. DF is supported in part by the NDSEG fellowship. Research was supported in part by the NSF under grants no. CDS&E-MSS 1521529 and 1521544. KS additionally acknowledges support from NSF CAREER Award 1750575 and a Sloan Research Fellowship.

## References

Katy S. Azoury and Manfred K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, June 2001.

Raghu Raj Bahadur et al. Sufficiency and statistical decision functions. *Ann. Math. Statist*, 25(3): 423–462, 1954.

Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *arXiv preprint arXiv:1404.5236*, 2014.

Ahron Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001.

Donald L. Burkholder. A geometrical characterization of banach spaces in which martingale difference sequences are unconditional. *The Annals of Probability*, pages 997–1011, 1981.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Ashok Cutkosky and Kwabena A Boahen. Online convex optimization with unconstrained domains and losses. In *Advances in Neural Information Processing Systems 29*, pages 748–756. 2016.

Ashok Cutkosky and Kwabena A. Boahen. Online learning without prior information. *The 30th Annual Conference on Learning Theory*, 2017.

Ashok Cutkosky and Francesco Orabona. Black-Box Reductions for Parameter-free Online Learning in Banach Spaces. *Conference on Learning Theory*, 2018.

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.

R.A. Fisher. On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, 222(594-604):309–368, 1922.

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

Dylan J Foster, Satyen Kale, Mehryar Mohri, and Karthik Sridharan. Parameter-free online learning via model selection. In *Advances in Neural Information Processing Systems 30*, pages 6020–6030. 2017a.

Dylan J. Foster, Alexander Rakhlin, and Karthik Sridharan. Zigzag: A new approach to adaptive online learning. *The 30th Annual Conference on Learning Theory*, 2017b.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Elad Hazan, Satyen Kale, and Shai Shalev-Shwartz. Near-optimal algorithms for online matrix prediction. In *Conference on Learning Theory*, pages 38–1, 2012.

Tuomas Hytönen, Jan van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach spaces*. Springer, 2016.

Adrian Stephen Lewis. Convex analysis on the hermitian matrices. *SIAM Journal on Optimization*, 6 (1):164–177, 1996.

Elliott H Lieb. Convex trace functions and the wigner-yanase-dyson conjecture. *Advances in Mathematics*, 11(3):267–288, 1973.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Lester Mackey, Michael I Jordan, Richard Y Chen, Brendan Farrell, Joel A Tropp, et al. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3): 906–945, 2014.

Brendan McMahan and Jacob Abernethy. Minimax optimal algorithms for unconstrained linear optimization. In *Advances in Neural Information Processing Systems*, pages 2724–2732, 2013.

Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. In *Proceedings of The 27th Conference on Learning Theory*, pages 1020–1039, 2014.

Arkadi Nemirovski. Prox-method with rate of convergence O(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Arkadii Nemirovskii, David Borisovich Yudin, and Edgar Ronald Dawson. Problem complexity and method efficiency in optimization. 1983.

Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. 1998.

Francesco Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, pages 1116–1124, 2014.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 2016.

Adam Osękowski. Two inequalities for the first moments of a martingale, its square function and its maximal function. *Bulletin Polish Acad. Sci. Math.*, 53:441–449, 2005.

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

Adam Osękowski. Personal communication. 2017.

Gilles Pisier. Martingales with values in uniformly convex spaces. *Israel Journal of Mathematics*, 20:326–350, 1975. ISSN 0021-2172.

A. Rakhlin, K. Sridharan, and A. Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research*, 2015.

Alexander Rakhlin and Karthik Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. *Conference on Learning Theory*, pages 1704–1722, 2017.

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

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In *Advances in neural information processing systems*, pages 2645–2653, 2011.

Joel Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

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

Volodimir Vovk. Competitive on-line linear regression. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 364–370, Cambridge, MA, USA, 1998. MIT Press.

Volodimir G Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.

## Appendix A. Further Directions

The core techniques developed in this paper suggest a number of promising future directions and natural extensions.

**Finding sufficient statistics** This paper gives multiple examples of Burkholder function constructions and sufficient statistics. If one wishes to find sufficient statistics for an adaptive bound $\mathcal{A}$ of interest, a basic rule of thumb is to consider a single input instance (instead of all $n$ data points) and determine—say—a polynomial expansion or expansion in another basis for the terms in $\mathrm{Reg}_n - \mathcal{A}$ involving the instance. This gives a coarse sketch of which statistics are necessary.

As an example, take the standard square loss with linear predictors as the benchmark class and suppose we are interested in a non-adaptive bound. Following the heuristic above, we need to find an expansion for terms of the form "$(\hat{y} - y)^2 - (\langle w, x\rangle - y)^2 - \texttt{constant}$". Expanding this expression out, we find that $\hat{y}^2$, $y \cdot x$ and $xx^\top$ are all required to write the expression explicitly. In fact, for this square loss example, the weighted sum of the $x_t$s and the sum of the outer products $\sum_t x_t x_t^\top$ turn out to be sufficient statistics as well.

For the examples in this paper, we exclusively considered benchmark classes $\mathcal{F}$ that were linear, which appears to have made the search for sufficient statistics easier. However, even when one considers a class $\mathcal{F}$ of non-linear functions, the approach of trying to expand the desired regret inequality (which now involves nonlinear $f \in \mathcal{F}$) around a given instance $x$ in terms of some basis may still help to obtain an adequate sufficient statistics. Furthermore, one may enlarge the class $\mathcal{F}$ to make the sufficient statistic search easier. For instance, if we want to learn the class of boolean decision trees of depth $d$, we can exploit that the class can be represented by polynomials of degree $d$ by using the discrete Fourier coefficients of the input instances up to degree $d$ as a sufficient statistic. In summary, for non-linear classes one may still search for sufficient statistics and Burkholder functions by expressing nonlinearities (approximately) via linear combinations of higher-order terms.

**Toward plug-and-play online learning** A natural next step is to automatize the search for sufficient statistics and Burkholder functions. Suppose that the sufficient statistic pair $(\mathbf{T}, V)$ is fixed and all that remains is to find a Burkholder function $\mathbf{U}$. If $V$ can be written as a polynomial of degree over sufficient statistic space $\mathcal{T}$, a natural approach is to restrict the search to Burkholder functions $\mathbf{U}$ that are themselves polynomials and relax the inequalities $1^o$/$2^o$/$3^o$ to sum-of-squares constraints (Barak and Steurer, 2014). We can then jointly search for a function $\mathbf{U}$ and a degree-$d$ sum-of-squares proof that this function satisfies the three properties in polynomial time once the degree of $\mathbf{U}$ is fixed. As a specific example, the problem of finding the zig-zag concave Burkholder function for $\ell_p$ norms explored in Foster et al. (2017b) has a sufficient statistic $V$ that is a polynomial of degree $p$ when $p \geq 2$ is an integer.

This approach is sound in that it will never incorrectly return a function $\mathbf{U}$ that does not satisfy the three properties, but may not be complete a-priori. An interesting direction is therefore to explore whether there are conditions under which this system can indeed be made complete.

**Generalized/non-additive sufficient statistics** The restriction in Definition 1 that sufficient statistics combine additively can be relaxed. A more general form is as follows. First, define a *representation space* $\mathcal{T}$. The function $\mathbf{T}$ now takes the form:

$$\mathbf{T} : \mathcal{X} \times \mathcal{Y} \times [-L, L] \times \mathcal{T} \to \mathcal{T}.$$

The restricted concavity condition for $\mathbf{U}$ under this definition becomes

$$\forall z, \tau : \quad \sup_{\mathbb{E}[\alpha]=0} \mathbb{E}_{\alpha} \mathbf{U}(\mathbf{T}(z, \alpha, \tau)) \leq \mathbf{U}(\tau).$$

Properties $1^o$ and $2^o$ of Lemma 3 remain the same. This generalized notion of a sufficient statistic allows us to move beyond additive updates—$\mathbf{T}$ can multiply $z$ with elements of $\mathcal{T}$, for example—but still restricts storage to the space $\mathcal{T}$ and is fully compatible with the Burkholder method and general algorithm framework. The generalizations of the equivalence theorem (Lemma 3) and the Burkholder algorithm (Lemma 5) for this notion of sufficient statistic hold as well.

## Appendix B. Fast and Easy Parameter-Free Online Learning

So far all of our examples have concerned adaptive bounds $\mathcal{A}$ that adapt to the data sequence $x_1, \ldots, x_n$, not the comparator $f$. In this section we will show that the framework of Burkholder functions and sufficient statistics readily encompasses comparator-dependent norms by giving a new family of algorithms for the problem of *parameter-free online learning* (McMahan and Orabona, 2014). The setup is as follows: We equip the subset $\mathcal{X} \subseteq \mathbb{R}^d$ with a norm $\|\cdot\|$ and assume that $\|x_t\| \leq 1$ for all $t$.[5] Recall that $\|\cdot\|_\star$ will denote the dual norm. Rather than constraining the benchmark class to a compact set, we set $\mathcal{W} = \mathbb{R}^d$ and set $\mathcal{F} = \{x \mapsto \langle w, x\rangle \mid w \in \mathcal{W}\}$. We assume smoothness of the

[5] The result extends verbatim to the general Banach space case; this is only to simplify presentation.

norm: letting $\Psi(x) = \frac{1}{2}\|x\|^2$, it holds that[6]

$$\Psi(x+y) \le \Psi(x) + \langle \nabla\Psi(x), y\rangle + \frac{\beta}{2}\|y\|^2.$$

To ease notational burden, we will assume the loss is 1-Lipschitz in this section. We will efficiently obtain a regret bound of the form

$$\mathrm{Reg}_n(w) \le \mathcal{A}(w) := \|w\|_\star \sqrt{2\beta n \log\left(\sqrt{\beta n}\|w\|_\star + 1\right)} + 1 \quad \forall w \in \mathbb{R}^d \tag{16}$$

for any such smooth norm. We begin by stating a sufficient statistic representation for the problem. This is based on a familiar potential which has appeared in previous works on parameter-free online learning (e.g. (McMahan and Orabona, 2014)) in Hilbert spaces; we extend it to any smooth norm, then use it in the Burkholder method to provide *the first linear time/linear space algorithm for parameter-free learning with general smooth norms in online supervised learning.*[7]

**Proposition 14** *Suppose we are interested in an adaptive regret bound of*

$$\mathcal{A}(w) = \|w\|_\star \sqrt{2an \log\left(\frac{\sqrt{an}\|w\|_\star}{\gamma} + 1\right)} + c$$

*for constants $a, \gamma, c > 0$. Then $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = (\delta_t \cdot \widehat{y}_t, \delta_t \cdot x_t) \in \mathbb{R} \times \mathcal{X}$ and the function*

$$V(b, x) = b + \gamma \exp\left(\frac{\|x\|^2}{2an}\right) - c, \tag{17}$$

*yield a sufficient statistic pair for the regret bound $\mathcal{A}$.*

Because the regret bound we provide is not horizon independent unlike previous examples, it will be convenient to allow time-indexed Burkholder functions $(\mathbf{U}_t)_{t\le n}$. This indexing is of purely notational convenience, as time-dependent Burkholder functions fit squarely into the algorithmic framework of Lemma 5 by enlarging $\mathcal{X}$ to $\mathcal{X} \times [n]$. Nonetheless, we recap the analogous properties for time-dependent Burkholder functions in the proof of the following theorem.

**Theorem 15** *Suppose $c = 1$, $a = \beta$, and $\gamma = 1/\sqrt{n}$ in (17). Then*

$$\mathbf{U}_t(b, x) := b + \frac{1}{\sqrt{n}} \exp\left(\frac{\|x\|^2}{2\beta t} + \frac{1}{2}\sum_{s=t+1}^{n} \frac{1}{s}\right) - 1,$$

*is a family of time-varying Burkholder functions satisfying $1^o$, $2^o$, and $3'$.*

This Burkholder function immediately yields both a prediction strategy achieving (16) and a simple probabilistic martingale inequality. We will now state them both. Because $(\mathbf{U}_t)_{t\le n}$ satisfy additional convexity properties, the strategy is especially efficient (per Appendix E and Lemma 22).

[6] Our analysis extends to the general case where we instead have $\frac{1}{2}\|x\|^2 \le \Psi(x)$ for some $\Psi \ne \frac{1}{2}\|\cdot\|^2$ and the same smoothness inequality holds, which is needed for settings such as $\ell_1/\ell_\infty$.

[7] Since the original submission of this paper, the independent work of (Cutkosky and Orabona, 2018) has provided an algorithm with a similar regret guarantee and computational efficiency.

**Corollary 16** *Suppose that $\mathcal{Y} = [-B, B]$ for some $B > 0$. Then the deterministic prediction strategy*

$$\widehat{y}_t = \mathrm{proj}_{[-B,B]}\left(-\frac{1}{\sqrt{n}}\mathbb{E}_{\sigma\in\{\pm 1\}}\left[\sigma\cdot\exp\left(\frac{\left\|\sum_{s=1}^{t-1}\delta_s x_s + \sigma x_t\right\|^2}{2\beta t} + \frac{1}{2}\sum_{s=t+1}^{n}\frac{1}{s}\right)\right]\right)$$

*leads to a regret bound of*

$$\sum_{t=1}^{n}\ell(\widehat{y}_t, y_t) - \sum_{t=1}^{n}\ell(\langle w, x_t\rangle, y_t) \leq \|w\|_\star\sqrt{2\beta n\log\left(\sqrt{\beta}n\|w\|_\star + 1\right)} + 1 \quad \forall w\in\mathbb{R}^d.$$

The Burkholder function family stated above and Lemma 3 certify that $\sup\mathbb{E}[V] \leq 0$. One special case of this martingale inequality is the following mgf bound for vector-valued martingales under smooth norms.

**Corollary 17** *Let $\boldsymbol{x}_t(\epsilon) \coloneqq \boldsymbol{x}_t(\epsilon_1, \dots, \epsilon_{t-1})$ be adapted to the filtration $\mathcal{F}_{t-1} = \sigma(\epsilon_1, \dots, \epsilon_{t-1})$ for Rademacher random variables $\epsilon_1, \dots, \epsilon_n$, and let $\|\boldsymbol{x}_t\| \leq 1$ almost surely, where $\|\cdot\|$ is a $\beta$-smooth norm. Then it holds that*

$$\mathbb{E}_\epsilon\exp\left(\frac{\left\|\sum_{t=1}^{n}\epsilon_t\boldsymbol{x}_t(\epsilon)\right\|^2}{2\beta n}\right) \leq \sqrt{n}.$$

**Related work** Parameter-free online learning is a very active area of research, but essentially all results in this area that we are aware of (McMahan and Abernethy, 2013; McMahan and Orabona, 2014; Orabona, 2014; Orabona and Pál, 2016; Cutkosky and Boahen, 2016, 2017) only provide regret bounds of the form (16) in the special case where $\|\cdot\|$ is a Hilbert space. The only exception is (Foster et al., 2017a) which gives an algorithm for smooth norms $\|\cdot\|$, but has time $\mathrm{poly}(n)$ per step. Our Burkholder-based algorithm has running time $O(d)$ per step and only $O(d)$ memory.[8] The key ingredient to achieving this improvement was to examine a known potential through the lens of the Burkholder method. We hope that this approach can lead to similarly useful improvements by applying the Burkholder method to construct more sophisticated potentials as in, e.g. (Orabona and Pál, 2016; Cutkosky and Boahen, 2017), particularly to achieve regret bounds that adapt jointly to the model and to data.

## Appendix C. Necessary Conditions

We now state a simple, yet powerful result that characterizes when existence of a Burkholder function for a sufficient statistic representation pair $(\mathbf{T}, V)$ is not only sufficient, but *necessary* to obtain a particular regret bound.

**Proposition 18** *Let $\delta = (\delta_1, \dots, \delta_n)$ be a $[-L, L]$-valued martingale difference sequence over filtration $\mathcal{F}_{t-1} = \sigma(\delta_1, \dots, \delta_{t-1})$ and let $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)$ be a sequence of functions $\boldsymbol{z}_t : [-L, L]^{t-1} \to \mathcal{X}\times\mathcal{Y}$, each viewed as a predictable process with respect to $\mathcal{F}_{t-1}$. Suppose for every such $(\delta, \boldsymbol{z})$ pair there exists a randomized adversary strategy $(x_t, y_t)$ that guarantees, for every learner strategy $(\widehat{y}_t)_{t\leq n}$,*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\sum_{t=1}^{n}\ell(\widehat{y}_t, y_t) - \ell(f(x_t), y_t) - \mathcal{A}(f; x_1, \dots, x_n)\right] \geq \mathbb{E}\left[V\left(\sum_{t=1}^{n}\mathbf{T}(\boldsymbol{z}_t, \delta_t)\right)\right]. \quad (18)$$

[8] Technically our algorithm only applies to the online supervised learning setting, whereas the algorithm of Foster et al. (2017a) applies to the OCO setting.

*Then, if there exists a strategy $(\widehat{y}_t)_{t\le n}$ that achieves the regret bound $\mathcal{A}(f;x_{1:n})$, this implies that*

$$\sup_{\delta,\boldsymbol{z}}\mathbb{E}\left[V\left(\sum_{t=1}^{n}\mathbf{T}(\boldsymbol{z}_t,\delta_t)\right)\right]\le 0.^{9}$$

*Consequently, the regret bound $\mathcal{A}(f;x_{1:n})$ is achievable only if there exists a Burkholder function* $\mathbf{U}:\mathcal{T}\to\mathbb{R}$ *that satisfies properties $1^o/2^o/3^o$ of Lemma 3.*

*When $\alpha\mapsto V(\tau+\mathbf{T}(z,\alpha))$ is convex for any $z\in\mathcal{X}\times\mathcal{Y},\tau\in\mathcal{T}$, we only require the preceeding inequalities to hold for $\delta_t=\epsilon_t\cdot L$, $\forall t=1,\dots,n$, where $\epsilon_t$s are independent Rademacher random variables. In this case achievability of the regret bound $\mathcal{A}(f;x_{1:n})$ only implies existence of a Burkholder function* $\mathbf{U}$ *satisfying property $3'$, not $3^o$.*

**Linear Classes** At first glance the conditions of Proposition 18 may seem fairly restrictive, but it is fairly straightforward to instantiate for all the examples in this paper. Consider the following linear setting: Take $\mathcal{X}\subseteq\mathbb{R}^d$, $\mathcal{Y}$ arbitrary, and let $\mathcal{F}$ be a linear class of the form $\{x\mapsto\langle w,x\rangle\mid w\in\mathcal{W}\}$, where $\sup_{x\in\mathcal{X},w\in\mathcal{W}}\langle w,x\rangle\le 1$ and $\mathcal{W}$ is symmetric. Pick an arbitrary vector space $\overline{\mathcal{T}}$, let $\overline{\mathbf{T}}:\mathcal{X}\to\overline{\mathcal{T}}$ be an any featurization of the input space, and let $F:\overline{\mathcal{T}}\to\mathbb{R}$ be an arbitrary function. Our goal will be to achieve a regret bound of the form

$$\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)\le\mathcal{A}(x_{1:n}):=F\left(\sum_{t=1}^{n}\overline{\mathbf{T}}(x_t)\right).\tag{19}$$

Let us first consider a natural choice of $V$ for the upper bound in this setting. Linearizing and using symmetry of $\mathcal{W}$, we have

$$\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)-\mathcal{A}(x_{1:n})\le\sum_{t=1}^{n}\widehat{y}_t\cdot\delta_t+\sup_{w\in\mathcal{W}}\left\langle w,\sum_{t=1}^{n}\delta_t x_t\right\rangle-F\left(\sum_{t=1}^{n}\overline{\mathbf{T}}(x_t)\right).$$

This means that if we choose a sufficient statistic $\mathbf{T}:(x_t,\widehat{y}_t,\delta_t)\mapsto(\widehat{y}_t\delta_t,x_t\delta_t,\overline{\mathbf{T}}(x_t))\in\mathbb{R}\times\mathbb{R}^d\times\overline{\mathcal{T}}$ and choose $V(a,x,\overline{\tau})=a+\sup_{w\in\mathcal{W}}\langle w,x\rangle-F(\overline{\tau})$, then it holds that

$$\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)-\mathcal{A}(x_{1:n})\le V\left(\sum_{t=1}^{n}\mathbf{T}(x_t,\widehat{y}_t,\delta_t)\right).$$

Noting that $\alpha\mapsto V(\tau+\mathbf{T}(x,\widehat{y},\alpha))$ is convex, Lemma 2 implies that a sufficient condition to achieve the regret bound for any convex 1-Lipschitz loss is that

$$\sup_{\boldsymbol{z}}\mathbb{E}_{\epsilon}\left[V\left(\sum_{t=1}^{n}\mathbf{T}(\boldsymbol{z}_t,\epsilon_t)\right)\right]\le 0,\tag{20}$$

where $\boldsymbol{z}$ is any $\mathcal{X}\times\mathcal{Y}$-valued predictable process with respect to the Rademacher sequence $\epsilon_1,\dots,\epsilon_n$.

By specializing to the absolute loss $\ell(\widehat{y},y)=|\widehat{y}-y|$ and choosing an adversary that plays $y_t$ to be Rademacher random variables and $x_t$ to be any predictable sequence, it can be shown that (20) is also *necessary*; this is proven formally in the appendix. As a corollary, we derive the following result.

[9] In the more general case, if (18) holds up to additive slack $\Delta$, the corresponding condition is $\sup\mathbb{E}[V]\le\Delta$.

**Proposition 19** *There exists a Burkholder function* $\mathbf{U}$ *for the pair* $(\mathbf{T}, V)$ if and only if *the regret bound* (19) *is achievable.*

Consider the matrix prediction setting of Section 5 for the special case of $L = 1$ and $r = 1$. This setting fits into the linear class framework above by taking $\mathcal{W}$ to be the nuclear norm ball in $\mathbb{R}^{d_1 \times d_2}$ and setting $\overline{\mathbf{T}}(X) = \mathcal{M}(X)$ for any matrix $X \in \mathbb{R}^{d_1 \times d_2}$. For this setting Proposition 19 implies the following equivalence.

**Example 4 (Matrix Prediction)** *The following are equivalent:*

1. *The regret bound*

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{W : \|W\|_\Sigma \leq 1} \sum_{t=1}^{n} \ell(\langle W, X_t \rangle, y_t) \leq \frac{\eta}{2} \left\| \sum_{t=1}^{n} \mathcal{M}(X_t) \right\|_\sigma + \frac{c}{\eta}$$

   *is achievable.*

2. *The martingale inequality*

$$\mathbb{E}_\epsilon \left\| \sum_{t=1}^{n} \epsilon_t \boldsymbol{X}_t(\epsilon) \right\|_\sigma \leq \frac{\eta}{2} \mathbb{E}_\epsilon \left\| \sum_{t=1}^{n} \mathcal{M}(\boldsymbol{X}_t(\epsilon)) \right\|_\sigma + \frac{c}{\eta}$$

   *holds for all* $\mathbb{R}^{d_1 \times d_2}$*-valued predictable processes* $\boldsymbol{X}$.

3. *There exists a Burkholder function for the sufficient statistic pair* $(\mathbf{T}, V)$ *in* (20).

## Appendix D. Proofs

### D.1. Proofs from Section 3 and Section 4

**Proof** [**of Lemma 2**] We will use the notation $\langle\!\langle \ldots \rangle\!\rangle_{t=1}^{n}$ to denote the repeated application of operators, with the outer application corresponding to $t = 1$. Existence of a randomized strategy for (1) is equivalent to the following quantity being non-positive:

$$\left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta_\mathcal{Y}} \sup_{y_t \in \mathcal{Y}} \mathbb{E}_{\widehat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[ \sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \phi(x_1, y_1, \ldots, x_n, y_n) \right].$$

By the minimax theorem, this is equal to

$$\left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta_\mathcal{Y}} \inf_{\widehat{y}_t \in \mathcal{Y}} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[ \sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \phi(x_1, y_1, \ldots, x_n, y_n) \right].$$

See (Rakhlin et al., 2010, 2012; Foster et al., 2015) for detailed discussion of the technical conditions under which the minimax theorem can be applied in the online learning setting; briefly, our assumptions that $\mathcal{Y}$ is a compact subset of $\mathbb{R}$ and that $\ell$ and $\phi$ are bounded are sufficient. In view of (3), the above quantity is upper bounded by

$$\leq \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta_\mathcal{Y}} \inf_{\widehat{y}_t \in \mathcal{Y}} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[ V\left( \sum_{t=1}^{n} \mathbf{T}(x_t, \widehat{y}_t, \partial \ell(\widehat{y}_t, y_t)) \right) \right].$$

Now, for each time $t$, choose the dual strategy

$$\widehat{y}_t^* := \arg\min_{\widehat{y}\in\mathcal{Y}} \mathbb{E}_{y_t\sim p_t}\, \ell(\widehat{y}, y_t),$$

so that $0 \in \partial\, \mathbb{E}_{y_t\sim p_t}\, \ell(\widehat{y}_t^*, y_t)$; that this is possible is implied by the assumption on the loss $\ell$ stated in Section 2. This choice implies that $\partial\ell(\widehat{y}_t^*, y_t) = \delta_t$ is a zero mean real variable conditionally on the past, i.e. $\mathbb{E}[\delta_t \mid \mathcal{G}_t] = 0$, where $\mathcal{G}_t = \sigma(\widehat{y}_{1:t-1})$. This particular choice for the $\widehat{y}_t$ in the dual game leads to the upper bound

$$\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{p_t\in\Delta_{\mathcal{Y}}} \mathbb{E}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[ V\left( \sum_{t=1}^{n} \mathbf{T}(x_t, \widehat{y}_t^*, \delta_t) \right) \right],$$

which is, in turn, upper bounded by

$$\left\langle\!\!\left\langle \sup_{z_t\in\mathcal{X}\times\mathcal{Y}} \sup_{p_t\in\Delta_{[-L,L]}:\mathbb{E}[\delta_t]=0} \mathbb{E}_{\delta_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[ V\left( \sum_{t=1}^{n} \mathbf{T}(z_t, \delta_t) \right) \right].$$

The last expression can be written in the functional form as

$$\sup_{\boldsymbol{z},\boldsymbol{p}} \mathbb{E}_{\delta\sim\boldsymbol{p}} \left[ V\left( \sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t, \delta_t) \right) \right].$$

using the notation of the lemma, with the supremum over $\boldsymbol{p}$ ranging over all joint distributions on $\delta = (\delta_1, \ldots, \delta_n)$ satisfying $\mathbb{E}[\delta_t \mid \delta_{1:t-1}] = 0$ for all $t \in [n]$. The non-positivity of the latter quantity is therefore sufficient to ensure the existence of a prediction strategy satisfying (1). ■

**Proof** [**of Lemma 3**] We first establish existence of $\mathbf{U}$ under the premise of the lemma. The construction is given by

$$\mathbf{U}(\tau) = \sup_{\boldsymbol{z},\boldsymbol{p}} \mathbb{E}_{\delta\sim\boldsymbol{p}} \left[ V\left( \tau + \sum_{t\geq 1} \mathbf{T}(\boldsymbol{z}_t, \delta_t) \right) \right]. \tag{21}$$

Then under the probabilistic inequality that is the premise of the lemma, it holds that

$$\mathbf{U}(0) = \sup_{\boldsymbol{z},\boldsymbol{p}} \mathbb{E}_{\delta\sim\boldsymbol{p}} \left[ V\left( \sum_{t\geq 1} \mathbf{T}(\boldsymbol{z}_t, \delta_t) \right) \right] \leq 0.$$

Next, by our assumption, $\exists z^0$ s.t. $\mathbf{T}(z^0, 0) = 0$, we can lower bound the supremum in (21) by considering a particular $\boldsymbol{z}$ that is constant $\boldsymbol{z}_t := z^0$ for all $t$, and a distribution for $\delta_t$ that only places mass on the singleton $0$. This yields a lower bound

$$\mathbf{U}(\tau) \geq V(\tau).$$

To verify the third condition, observe that for any zero-mean random variable $\alpha$ with distribution $p$ supported on $[-L, L]$,

$$\begin{aligned}
\mathbb{E}_{\alpha}\left[\mathbf{U}(\tau + \mathbf{T}(z,\alpha))\right] &= \mathbb{E}_{\alpha}\left[ \sup_{\boldsymbol{z},\boldsymbol{p}} \mathbb{E}_{\delta\sim\boldsymbol{p}} \left[ V\left( \tau + \mathbf{T}(z,\alpha) + \sum_{t} \mathbf{T}(\mathbf{z}_t, \delta_t) \right) \right] \right] \\
&\leq \sup_{\boldsymbol{z},\boldsymbol{p}} \mathbb{E}_{\delta\sim\boldsymbol{p}} \left[ V\left( \tau + \sum_{t} \mathbf{T}(\boldsymbol{z}_t, \delta_t) \right) \right] \\
&= \mathbf{U}(\tau).
\end{aligned}$$

For the converse, assume we have a function $\mathbf{U}$ satisfying the three properties. Fix any $\boldsymbol{z}$ and $\boldsymbol{p}$ of length $n$. In this case, by property $2^o$, the following inequality holds deterministically:

$$V\left(\sum_{t=1}^{n}\mathbf{T}(\boldsymbol{z}_t,\delta_t)\right) \le \mathbf{U}\left(\sum_{t=1}^{n}\mathbf{T}(\boldsymbol{z}_t,\delta_t)\right).$$

By property $3^o$, we have that for any time $s$,

$$\mathbb{E}_{\delta_n}\mathbf{U}\left(\sum_{t=1}^{s}\mathbf{T}(\boldsymbol{z}_t,\delta_t)\right) \le \mathbf{U}\left(\sum_{t=1}^{s-1}\mathbf{T}(\boldsymbol{z}_t,\delta_t)\right).$$

Continuing this argument all the way to $t=0$ and using property $1^o$,

$$\sup_{\boldsymbol{z},\boldsymbol{p}}\mathbb{E}_{\delta\sim\boldsymbol{p}}\left[V\left(\sum_{t=1}^{n}\mathbf{T}(\boldsymbol{z}_t,\delta_t)\right)\right] \le \mathbf{U}(0) \le 0.$$

■

### D.2. Proofs from Section 5

**Proof** [**of Proposition 7**]

Recall that $\mathcal{A}_\eta(X_1,\ldots,X_n) = \frac{\eta r L^2}{2}\left\|\sum_{t=1}^{n}\mathcal{M}(X_t)\right\|_\sigma + \frac{c}{\eta}$. Linearizing the loss with the adaptive bound as in (2),

$$\begin{aligned}
&\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t) - \inf_{W\in\mathcal{W}}\ell(\langle W,X_t\rangle,y_t) - \mathcal{A}_\eta(X_1,\ldots,X_n)\\
&\le \sup_{W\in\mathcal{W}}\left\{\sum_{t=1}^{n}\partial\ell(\widehat{y}_t,y_t)(\widehat{y}_t - \langle W,X_t\rangle) - \mathcal{A}_\eta(X_1,\ldots,X_n)\right\}\\
&= \sum_{t=1}^{n}\partial\ell(\widehat{y}_t,y_t)\widehat{y}_t + r\left\|\sum_{t=1}^{n}\partial\ell(\widehat{y}_t,y_t)X_t\right\|_\sigma - \mathcal{A}_\eta(X_1,\ldots,X_n).
\end{aligned}$$

We now abbreviate $\partial\ell(\widehat{y}_t,y_t)=\delta_t$ and expand out $\mathcal{A}_\eta$, yielding

$$\sum_{t=1}^{n}\delta_t\cdot\widehat{y}_t + r\left\|\sum_{t=1}^{n}\delta_t X_t\right\|_\sigma - \frac{\eta r L^2}{2}\left\|\sum_{t=1}^{n}\mathcal{M}(X_t)\right\|_\sigma - \frac{c}{\eta}.$$

Using the fact that $\lambda_1(\mathcal{H}(X)) = \|X\|_\sigma$, linearity of $\mathcal{H}$, and that $\mathcal{M}(X_t)$ is positive semidefinite, we write this as

$$\sum_{t=1}^{n}\delta_t\cdot\widehat{y}_t + r\lambda_1\left(\sum_{t=1}^{n}\delta_t\mathcal{H}(X_t)\right) - r\lambda_1\left(\frac{\eta L^2}{2}\sum_{t=1}^{n}\mathcal{M}(X_t)\right) - \frac{c}{\eta}$$

Sub-additivity of $\lambda_1$ gives a further upper bound of

$$\sum_{t=1}^{n}\delta_t\cdot\widehat{y}_t + r\lambda_1\left(\sum_{t=1}^{n}\delta_t\mathcal{H}(X_t) - \frac{\eta L^2}{2}\sum_{t=1}^{n}\mathcal{M}(X_t)\right) - \frac{c}{\eta}$$

Then $\mathbf{T}(X_t, \widehat{y}_t, \delta_t) = (\delta_t \cdot \widehat{y}_t, \delta_t \cdot \mathcal{H}(X_t), \mathcal{M}(X_t)) \in \mathbb{R} \times \mathbb{S}^{d_1+d_2} \times \mathbb{S}_+^{d_1+d_2}$ is a sufficient statistic. Namely, writing

$$V(a, H, M) = a + r\lambda_1\left(H - \frac{\eta L^2}{2}M\right) - \frac{c}{\eta},$$

our calculation shows that

$$\sup_{W \in \mathcal{W}} \{\mathrm{Reg}_n(W) - \mathcal{A}(X_1, \ldots, X_n)\} \le V\left(\sum_{t=1}^n \mathbf{T}(X_t, \widehat{y}_t, \delta_t)\right).$$

■

**Proof** [**of Theorem 8**]

Recall that

$$\mathbf{U}(a, H, M) = a + \frac{r}{\eta}\log \operatorname{tr} \exp\left(\eta H - \frac{\eta^2 L^2}{2}M\right) - \frac{c}{\eta}$$

We will show that $\mathbf{U}$ satisfies the three properties of Lemma 3. For property $1^o$, we have

$$\mathbf{U}(0) = \frac{r}{\eta}\log(\operatorname{tr}(\exp(0))) - \frac{c}{\eta} = \frac{r\log(d_1 + d_2)}{\eta} - \frac{c}{\eta}.$$

Thus, $\mathbf{U}(0) \le 0$ as soon as $c \ge r\log(d_1 + d_2)$.

For property $2^o$, it suffices to show that $\lambda_1(H - \frac{\eta L^2}{2}M) \le \frac{1}{\eta}\log \operatorname{tr} \exp\left(\eta H - \frac{\eta^2 L^2}{2}M\right)$. To this end, we have

$$\lambda_1\left(H - \frac{\eta L^2}{2}M\right) = \frac{1}{\eta}\log \lambda_1\left(\exp\left(\eta H - \frac{\eta^2 L^2}{2}M\right)\right) \le \frac{1}{\eta}\log \operatorname{tr} \exp\left(\eta H - \frac{\eta^2 L^2}{2}M\right),$$

where the equality is well-defined because the matrix under consideration is symmetric and the inequality follows because $e^A$ is positive semidefinite for any symmetric matrix $A$.

For the third property, observe that the mapping $\alpha \mapsto V(\tau + \mathbf{T}(z, \alpha))$ is convex (e.g. (Lewis, 1996)). Consequently, by Lemma 3, it suffices only to prove property $3'$, i.e. that the restricted concavity condition holds only for Rademacher random variables.

Fix $\tau \in \mathcal{T}$ and $z = (X, \widehat{y}) \in \mathcal{X} \times \mathcal{Y}$, and let $\epsilon$ be a Rademacher random variable. Writing

$$\tau = (\tau_1, \tau_2, \tau_3) \in \mathbb{R} \times \mathbb{S}^{d_1+d_2} \times \mathbb{S}_+^{d_1+d_2},$$

we have

$$\begin{aligned}
&\mathbb{E}_\epsilon\left[\mathbf{U}(\tau + \mathbf{T}(z, \epsilon L))\right] \\
&= \mathbb{E}_\epsilon\left[\tau_1 + \widehat{y}\epsilon L + \frac{r}{\eta}\log \operatorname{tr} \exp\left(\eta\tau_2 - \frac{\eta^2 L^2}{2}\tau_3 + \eta\epsilon L\mathcal{H}(X) - \frac{\eta^2 L^2}{2}\mathcal{M}(X)\right)\right] - \frac{c}{\eta}. \\
&= \frac{r}{\eta}\mathbb{E}_\epsilon\left[\log \operatorname{tr} \exp\left(\eta\tau_2 - \frac{\eta^2 L^2}{2}\tau_3 + \eta\epsilon L\mathcal{H}(X) - \frac{\eta^2 L^2}{2}\mathcal{M}(X)\right)\right] + \tau_1 - \frac{c}{\eta}.
\end{aligned}$$

Focusing on the log-trace-exponential term, observe that

$$\begin{aligned}
&\mathbb{E}_\epsilon\left[\log \operatorname{tr} \exp\left(\eta\tau_2 - \frac{\eta^2 L^2}{2}\tau_3 + \eta\epsilon L\mathcal{H}(X) - \frac{\eta^2 L^2}{2}\mathcal{M}(X)\right)\right] \\
&= \mathbb{E}_\epsilon\left[\log \operatorname{tr} \exp\left(\eta\tau_2 - \frac{\eta^2 L^2}{2}\tau_3 + \log(\exp(\eta\epsilon L\mathcal{H}(X))) - \frac{\eta^2 L^2}{2}\mathcal{M}(X)\right)\right].
\end{aligned}$$

Since $\exp(\eta\epsilon L\mathcal{H}(X))$ is positive definite and $\eta\tau_2 - \frac{\eta^2 L^2}{2}\tau_3 - \frac{\eta^2 L^2}{2}\mathcal{M}(X)$ is symmetric (by assumption), we can apply Lieb's Concavity Theorem to upper bound this by

$$\log \operatorname{tr} \exp\left(\eta\tau_2 - \frac{\eta^2 L^2}{2}\tau_3 + \log(\mathbb{E}_\epsilon \exp(\eta\epsilon L\mathcal{H}(X))) - \frac{\eta^2 L^2}{2}\mathcal{M}(X)\right).$$

The Rademacher matrix mgf bound (Tropp, 2012) now yields

$$\log(\mathbb{E}_\epsilon \exp(\eta\epsilon L\mathcal{H}(X))) \leq \log\left(\exp\left(\eta^2 L^2\mathcal{M}(X)\right)/2\right) = \eta^2 L^2\mathcal{M}(X)/2.$$

Since $A \leq B$ implies $\operatorname{tr} e^A \leq \operatorname{tr} e^B$, this implies that

$$\mathbb{E}_\epsilon\left[\log \operatorname{tr} \exp\left(\eta\tau_2 - \frac{\eta^2 L^2}{2}\tau_3 + \eta\epsilon L\mathcal{H}(X) - \frac{\eta^2 L^2}{2}\mathcal{M}(X)\right)\right] \leq \log \operatorname{tr} \exp\left(\eta\tau_2 - \frac{\eta^2 L^2}{2}\tau_3\right)$$

Combining everything we proved so far, this implies

$$\mathbb{E}_\epsilon\left[U(\tau + \mathbf{T}(z,\epsilon L))\right] \leq \tau_1 + \frac{r}{\eta}\log \operatorname{tr} \exp\left(\eta\tau_2 - \frac{\eta^2 L^2}{2}\tau_3\right) - \frac{c}{\eta} = U(\tau).$$

■

**Proof** [**of Corollary 9**] The Burkholder function $\mathbf{U}$ satisfies the conditions of Lemma 22. Direct calculation shows that the strategy in Lemma 22 matches the strategy in the statement of the corollary. ■

**Proof** [**of Corollary 11**] We invoke the Burkholder function $\mathbf{U}$ from Theorem 8 for the special case $r = 1$ and $c = \log(d_1 + d_2)$, and $L = 1$. In particular, its existence per Lemma 3 implies (for the corresponding $V$, here denoted $V_\eta$ to refer to the $V$ given for a fixed value of $\eta$)

$$\inf_\eta \sup_{\boldsymbol{z},\boldsymbol{p},n} \mathbb{E}\left[V_\eta\left(\sum_{t=1}^n \mathbf{T}(\boldsymbol{z}_t,\delta_t)\right)\right] \leq 0$$

We use this inequality only for the special case where $\delta_t = \epsilon_t$ and $\boldsymbol{z}_t = (\boldsymbol{X}_t(\epsilon), 0)$. For this special case, the inequality implies

$$\inf_\eta \sup_{\boldsymbol{X},n} \mathbb{E}\left[\left\|\sum_{t=1}^n \epsilon_t \boldsymbol{X}_t(\epsilon)\right\| - \frac{\eta}{2}\left\|\sum_{t=1}^n \mathcal{M}(\boldsymbol{X}_t(\epsilon))\right\| - \frac{\log(d_1+d_2)}{\eta}\right] \leq 0.$$

For any fixed martingale $(\boldsymbol{X}_t(\epsilon))_{t\leq n}$, this implies

$$\begin{aligned}\mathbb{E}\left\|\sum_{t=1}^n \epsilon_t \boldsymbol{X}_t(\epsilon)\right\| &\leq \inf_{\eta>0}\left\{\frac{\eta}{2}\mathbb{E}\left\|\sum_{t=1}^n \mathcal{M}(\boldsymbol{X}_t(\epsilon))\right\| + \frac{\log(d_1+d_2)}{\eta}\right\} \\ &= \sqrt{2\,\mathbb{E}\left\|\sum_{t=1}^n \mathcal{M}(\boldsymbol{X}_t(\epsilon))\right\|\log(d_1+d_2)}.\end{aligned}$$

To conclude, observe that for any sequence $(X_t)$ we have

$$\left\|\sum_{t=1}^n \mathcal{M}(X_t)\right\|_\sigma \leq \max\left\{\left\|\sum_{t=1}^n X_t X_t^\top\right\|_\sigma, \left\|\sum_{t=1}^n X_t^\top X_t\right\|_\sigma\right\}.$$

Indeed, $\sum_{t=1}^n \mathcal{M}(X_t) = \begin{pmatrix} \sum_{t=1}^n X_t X_t^\top & 0 \\ 0 & \sum_{t=1}^n X_t^\top X_t \end{pmatrix}$ and the spectral norm of a block-diagonal matrix is always obtained by the spectral norm of one of its blocks.

■

### D.3. Proofs from Section 6

**Proof** [**Sketch of proofs for claims from Section 6.2**]

For the $\ell_2$ result we have

$$\begin{aligned}
&\sum_{t=1}^n \ell(\widehat{y}_t, y_t) - \min_{\|w\|_2 \le 1} \sum_{t=1}^n \ell(\langle w, x_t\rangle, y_t) - 2L\sqrt{\sum_{t=1}^n \|x_t\|_2^2} \\
&\le \sup_{\|w\|_2 \le 1} \left\{ \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)(\widehat{y}_t, -\langle w, x_t\rangle) \right\} - 2L\sqrt{\sum_{t=1}^n \|x_t\|_2^2} \\
&= \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)\widehat{y}_t + \left\| \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t) x_t \right\|_2 - 2L\sqrt{\sum_{t=1}^n \|x_t\|_2^2} \\
&\le \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)\widehat{y}_t + \mathbf{U}_{\text{square}}\left( \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t) x_t, L\sqrt{\sum_{t=1}^n \|x_t\|_2^2} \right).
\end{aligned}$$

The path from here to a Burkholder function in the sense of Lemma 3 is clear given the three properties of $\mathbf{U}_{\text{square}}$ stated in the main body.

For the $\ell_\infty$ result, the quantity

$$\sum_{t=1}^n \ell(\widehat{y}_t, y_t) - \min_{\|w\|_\infty \le 1} \sum_{t=1}^n \ell(\langle w, x_t\rangle, y_t) - 2L\left\| \left( \sum_{t=1}^n x_t^2 \right)^{1/2} \right\|_1$$

can be upper bounded by

$$\begin{aligned}
&\sup_{\|w\|_\infty \le 1} \left\{ \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)(\widehat{y}_t, -\langle w, x_t\rangle) \right\} - 2L\left\| \left( \sum_{t=1}^n x_t^2 \right)^{1/2} \right\|_1 \\
&= \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)\widehat{y}_t + \left\| \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t) x_t \right\|_1 - 2L\left\| \left( \sum_{t=1}^n x_t^2 \right)^{1/2} \right\|_1 \\
&\le \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)\widehat{y}_t + \sum_{i=1}^d \mathbf{U}_{\text{square}}\left( \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t) x_t[i], L\sqrt{\sum_{t=1}^n (x_t[i])_2^2} \right),
\end{aligned}$$

where $x_t[i]$ refers to the $i$th coordinate of $x_t$. Once again, the three properties of $\mathbf{U}_{\text{square}}$ directly lead to a valid Burkholder function $\mathbf{U}$. ■

**Proof** [**of Proposition 12**] Let $A_n = \rho \sum_{t=1}^n z_t z_t^\top + \lambda I$ and $A_0 = \lambda I$. Recall that $\Psi_A(w) = \frac{1}{2}\langle w, Aw\rangle$. We begin by rewriting the desired regret bound as

$$\mathcal{A}(w; z_1, \ldots, z_n) = \lambda\Phi((w, 1)) + c\log(\det(A_n)/\det(A_0))$$

for a constant $c > 0$ to be determined. With this definition, we have

$$\begin{aligned}
&\sup_{w\in\mathbb{R}^d}\{\mathrm{Reg}_n(w) - \mathcal{A}(w; z_1,\dots,z_n)\}\\
&= \sup_{w\in\mathbb{R}^d}\left\{\sum_{t=1}^n \ell(\widehat{y}_t, y_t) - \sum_{t=1}^n \ell(\langle w, x_t\rangle, y_t) - \lambda\Phi((w,1))\right\} - c\log(\det(A_n)/\det(A_0))
\end{aligned}$$

Using strong convexity of $\ell$:

$$\begin{aligned}
&= \sup_{w\in\mathbb{R}^d}\left\{\sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)(\widehat{y}_t - \langle w, x_t\rangle) - \frac{\rho}{2}(\widehat{y}_t - \langle w, x_t\rangle)^2 - \lambda\Phi((w,1))\right\} - c\log(\det(A_n)/\det(A_0))\\
&= \sup_{w\in\mathbb{R}^d}\left\{\sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)(-\langle (w,1), z_t\rangle) - \frac{\rho}{2}(\langle (w,1), z_t\rangle)^2 - \lambda\Phi((w,1))\right\} - c\log(\det(A_n)/\det(A_0))
\end{aligned}$$

We now move to an upper bound by allowing the final coordinate of $(w,1)$ to act as a free parameter.

$$\le \sup_{w\in\mathbb{R}^{d+1}}\left\{\sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)\langle w, z_t\rangle - \frac{\rho}{2}\langle w, z_t\rangle^2 - \lambda\Phi(w)\right\} - c\log(\det(A_n)/\det(A_0))$$

We can rewrite this as

$$\begin{aligned}
&\le \sup_{w\in\mathbb{R}^{d+1}}\left\{\left\langle w, \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)z_t\right\rangle - \Psi_{\rho\Sigma_n}(w) - \lambda\Phi(w)\right\} - c\log(\det(A_n)/\det(A_0))\\
&= \sup_{w\in\mathbb{R}^{d+1}}\left\{\left\langle w, \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)z_t\right\rangle - \Psi_{A_n}(w)\right\} - c\log(\det(A_n)/\det(A_0))\\
&= \Psi^\star_{A_n}\left(\sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)z_t\right) - c\log(\det(A_n)/\det(A_0)).
\end{aligned}$$

This establishes that $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = (\delta_t z_t, z_t z_t^\top) \in \mathbb{R}^{d+1}\times\mathbb{S}_+^{d+1}$ is a sufficient statistic. This is because we can write

$$V(x, A) = \Psi^\star_{\rho A+\lambda I}(x) - c\log(\det(\rho A + \lambda I)/\det(A_0)).$$

and we just proved that

$$\sup_{w\in\mathbb{R}^d}\{\mathrm{Reg}_n(w) - \mathcal{A}(x_1,\dots,x_n)\} \le V\left(\sum_{t=1}^n \mathbf{T}(x_t, \widehat{y}_t, \delta_t)\right).$$

■

**Proof** [**of Theorem 13**] Recall that we have defined

$$\mathbf{U}(x, A) = V(x, A) = \Psi^\star_A(x) - c\log(\det(A)/\det(A_0)).$$

We verify the properties from Lemma 3. Property $2^o$ is immediate, and for property $1^o$ we have

$$\mathbf{U}(0) = \Psi^\star_{0+\lambda I}(0) - c\log(\det(A_0)/\det(A_0)) = 0.$$

We proceed to prove property $3^o$. Fix $\tau = (\tau_1, \tau_2) \in \mathcal{T} = \mathbb{R}^{d+1} \times \mathbb{S}_+^{d+1}$ and a mean-zero distribution $p$ over $[-L, L]$. Then we have

$$\begin{aligned}\mathbb{E}_{\alpha\sim p}\,\mathbf{U}(\tau + \mathbf{T}(z,\alpha)) &= \mathbb{E}_{\alpha\sim p}\Big[\Psi^\star_{\rho(\tau_2+zz^\top)+\lambda I}(\tau_1 + \alpha z) - c\log(\det(\rho(\tau_2 + zz^\top) + \lambda I)/\det(A_0))\Big] \\ &= \mathbb{E}_{\alpha\sim p}\Big[\Psi^\star_{\rho(\tau_2+zz^\top)+\lambda I}(\tau_1 + \alpha z)\Big] - c\log(\det(\rho(\tau_2 + zz^\top) + \lambda I)/\det(A_0)).\end{aligned}$$

Let $A = \rho(\tau_2 + zz^\top) + \lambda I$ and $B = \rho\tau_2 + \lambda I$. Then since $\Psi^\star$ is a squared Euclidean norm and $\alpha$ is mean-zero:

$$\mathbb{E}_{\alpha\sim p}[\Psi^\star_A(\tau_1 + \alpha z)] \le \Psi^\star_A(\tau_1) + \mathbb{E}_{\alpha\sim p}\big[\alpha^2\langle z, A^{-1}z\rangle\big] \le \Psi^\star_A(\tau_1) + L^2\big[\alpha^2\langle z, A^{-1}z\rangle\big].$$

Also note that since $B \le A$, $\Psi^\star_A(\tau_1) \le \Psi^\star_B(\tau_1)$.

To conclude, observe that we just established

$$\mathbb{E}_{\alpha\sim p}\,\mathbf{U}(\tau + \mathbf{T}(z,\alpha)) \le \Psi^\star_B(\tau_1) + L^2\langle z, A^{-1}z\rangle - c\log(\det(A)/\det(A_0)).$$

Using a standard argument (e.g. from Cesa-Bianchi and Lugosi (2006)) and using that $A = B + \rho zz^\top$:

$$\le \Psi^\star_B(\tau_1) + \frac{L^2}{\rho}\log(\det(A)/\det(B)) - c\log(\det(A)/\det(A_0)).$$

For $c \ge L^2/\rho$, this is bounded by

$$\begin{aligned}&\le \Psi^\star_B(\tau_1) - c\log(\det(B)/\det(A_0)) \\ &= \mathbf{U}(\tau).\end{aligned}$$

■

### D.4. Proofs from Appendix B

**Proof** [**of Proposition 14**] We define a potential function that will eventually be used in the construction of the Burkholder function $\mathbf{U}$ we provide for $V$. As discussed in the main body, a variant of this potential was first introduced by McMahan and Orabona (2014) for the special case of Hilbert spaces. Let $\Psi(x) = \frac{1}{2}\|x\|^2$ (not necessarily a Hilbert space norm) and define

$$F_n(x) = \gamma\exp\left(\frac{\Psi(x)}{an}\right).$$

From (McMahan and Orabona, 2014, Lemma 14), along with the additional fact that $(f(\|\cdot\|))^\star = f^\star(\|\cdot\|_\star)$ for general dual norm pairs, it holds that

$$F_n^\star(w) \le \|w\|_\star\sqrt{2an\log\left(\frac{\sqrt{an}\|w\|_\star}{\gamma} + 1\right)}.$$

This is all we need to establish the result. We proceed as follows

$$\begin{aligned}
&\sup_{w\in\mathbb{R}^d}\{\mathrm{Reg}_n(w)-\mathcal{A}(w)\}\\
&=\sup_{w\in\mathbb{R}^d}\left\{\sum_{t=1}^n \ell(\widehat{y}_t,y_t)-\ell(\langle w,x_t\rangle,y_t)-\mathcal{A}(w)\right\}\\
&\le \sup_{w\in\mathbb{R}^d}\left\{\sum_{t=1}^n \partial\ell(\widehat{y}_t,y_t)(\widehat{y}_t-\langle w,x_t\rangle)-\mathcal{A}(w)\right\}\\
&=\sum_{t=1}^n \partial\ell(\widehat{y}_t,y_t)\cdot\widehat{y}_t+\sup_{w\in\mathbb{R}^d}\left\{\left\langle w,\sum_{t=1}^n \partial\ell(\widehat{y}_t,y_t)x_t\right\rangle-\mathcal{A}(w)\right\}
\end{aligned}$$

Using the inequality for the potential $F_n^\star$ stated above:

$$\le \sum_{t=1}^n \partial\ell(\widehat{y}_t,y_t)\cdot\widehat{y}_t+F_n^\star\left(\sum_{t=1}^n \partial\ell(\widehat{y}_t,y_t)x_t\right)-c$$

It follows that $\mathbf{T}(x_t,\widehat{y}_t,\delta_t)=(\delta_t\cdot\widehat{y}_t,\delta_t\cdot x_t)\in\mathbb{R}\times\mathcal{X}$ is a sufficient statistic. This is because we can write

$$V(b,x)=b+F_n^\star(x)-c.$$

and we have just shown that

$$\sup_w\{\mathrm{Reg}_n(w)-\mathcal{A}(x_1,\ldots,x_n)\}\le V\left(\sum_{t=1}^n \mathbf{T}(x_t,\widehat{y}_t,\delta_t)\right).$$

■

**Proof** [**of Theorem 15**] Since $\mathbf{U}$ depends on time, we generalize the properties of Lemma 3 to

$1^o$ $\mathbf{U}_0(0)\le 0$

$2^o$ For any $\tau\in\mathcal{T}$, $\mathbf{U}_n(\tau)\ge V(\tau)$

$3^o$ For any $\tau\in\mathcal{T}$, $z\in\mathcal{X}\times\mathcal{Y}$, and any mean-zero distribution $p$ on $[-L,L]$, and any $t\ge 1$

$$\mathbb{E}_{\alpha\sim p}\left[\mathbf{U}_t(\tau+\mathbf{T}(z,\alpha))\right]\le\mathbf{U}_{t-1}(\tau) \tag{22}$$

$3'$ For any $\tau\in\mathcal{T}$, $z\in\mathcal{X}\times\mathcal{Y}$, and any $t\ge 1$,

$$\forall\tau\in\mathcal{T},z\in\mathcal{X}\times\mathcal{Y},\quad \mathbb{E}_\epsilon\,\mathbf{U}_t(\tau+\mathbf{T}(z,\epsilon L))\le\mathbf{U}_{t-1}(\tau),$$

where $\epsilon$ is a Rademacher random variable.

Recall that for simplicity we assume $L=1$ and $\mathcal{X}$ is a unit ball: $\|x\|\le 1$. Let $\Psi(x)=\frac{1}{2}\|x\|^2$, where we have assumed that $\beta$-smoothness of $\Psi$:

$$\Psi(x+y)\le\Psi(x)+\langle\nabla\Psi(x),y\rangle+\frac{\beta}{2}\|y\|^2.$$

Define a family of potentials

$$F_t(x) = \gamma \exp\left(\frac{\Psi(x)}{at} + \frac{1}{2}\sum_{s=t+1}^{n}\frac{1}{s}\right)$$

and $F_0 = \gamma \exp\left(\frac{1}{2}\sum_{t=1}^{n}\frac{1}{t}\right)$. Note that $F_n$ here is the same as in the proof of Proposition 14.

Observe that

$$\mathbf{U}_t(b,x) = b + F_t^\star(x) - c,$$

where $F_t^\star$ is as defined as in the proof of Proposition 14. We proceed to establish the three properties of $\mathbf{U}$ from Lemma 3. Property $2^o$ holds since $V = \mathbf{U}_n$. We will show property $3'$ first, then conclude with property $1^o$. Note that $\alpha \mapsto \mathbf{U}_t(\tau + \mathbf{T}(z,\alpha))$ is convex with respect to $\alpha$, and so it indeed suffices to show property $3'$.

Fix an element $\tau = (\tau_1, \tau_2) \in \mathbb{R} \times \mathcal{X} = \mathcal{T}$ of the sufficient statistic space. At time $n$ we have

$$\mathbb{E}_\epsilon\left[\mathbf{U}_n(\tau + \mathbf{T}(z,\epsilon))\right] = \mathbb{E}_\epsilon[\tau_1 + \epsilon \cdot \widehat{y} + F_n(\tau_2 + \epsilon x_n)] - c = \tau_1 + \mathbb{E}_\epsilon[F_n(\tau_2 + \epsilon x_n)] - c.$$

To handle $F_n$, begin by using smoothness of $\Psi$:

$$\mathbb{E}_\epsilon[F_n(\tau_2 + \epsilon x_n)] = \mathbb{E}_\epsilon \exp\left(\frac{\Psi(\tau_2 + \epsilon x)}{an}\right) \leq \mathbb{E}_\epsilon \exp\left(\frac{\Psi(\tau_2) + \epsilon\langle \nabla\Psi(\tau_2), x\rangle + \frac{\beta}{2}\|x\|^2}{an}\right)$$

Using the standard Rademacher mgf bound, $\mathbb{E}_\epsilon e^{\lambda\epsilon} \leq e^{\lambda^2/2}$, we upper bound the above quantity by

$$\exp\left(\frac{\Psi(\tau_2) + \frac{\beta}{2}\|x\|^2}{an} + \frac{\langle \nabla\Psi(\tau_2), x\rangle^2}{2(an)^2}\right) \leq \exp\left(\frac{\Psi(\tau_2) + \frac{\beta}{2}\|x\|^2}{an} + \frac{\|\nabla\Psi(\tau_2)\|_\star^2\|x\|^2}{2(an)^2}\right).$$

Using the assumption $\|x\| \leq 1$, we obtain an upper bound of

$$\exp\left(\frac{\Psi(\tau_2) + \frac{\beta}{2}}{an} + \frac{\|\nabla\Psi(\tau_2)\|_\star^2}{2(an)^2}\right).$$

We now use a basic fact from convex analysis, namely that any $\beta$-smooth convex function $f$, $\frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|_\star^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y\rangle$ . This yields an upper bound

$$\exp\left(\frac{\Psi(\tau_2) + \frac{\beta}{2}}{an} + \frac{\beta\Psi(\tau_2)}{(an)^2}\right)$$

Setting $a = \beta$, this is equal to

$$\exp\left(\frac{1}{\beta}\left(\frac{1}{n} + \frac{1}{n^2}\right)\Psi(\tau_2) + \frac{1}{2n}\right).$$

As a last step, observe that $\frac{1}{n} + \frac{1}{n^2} \leq \frac{1}{n-1}$. Indeed,

$$\frac{1}{n} + \frac{1}{n^2} = \frac{1}{n}\left(1 + \frac{1}{n}\right) = \frac{1}{n-1}\frac{n-1}{n}\left(1 + \frac{1}{n}\right) = \frac{1}{n-1}\left(1 - \frac{1}{n}\right)\left(1 + \frac{1}{n}\right) = \frac{1}{n-1}\left(1 - \frac{1}{n^2}\right) \leq \frac{1}{n-1}.$$

Therefore, we have established that

$$\mathbb{E}_\epsilon[F_n(\tau_2+\epsilon x_n)] \le \exp\left(\frac{\Psi(\tau_2)}{\beta(n-1)}+\frac{1}{2n}\right) = F_{n-1}(\tau_2),$$

and in particular $\mathbb{E}_\epsilon\, \mathbf{U}_n(\tau+\mathbf{T}(z,\epsilon)) \le \mathbf{U}_{n-1}(\tau)$. In fact, by folding the terms $\frac{1}{2}\sum_{s=t+1}^{n}\frac{1}{s}$—which do not depend on data—into a multiplicative constant, this argument yields, for any $t$ and any $\|x\| \le 1$,

$$\mathbb{E}_\epsilon[F_t(\tau+\epsilon x)] \le F_{t-1}(\tau).$$

Thus, for each $t \ge 2$ we have

$$\mathbb{E}_\epsilon\left[\mathbf{U}_t(\tau+\mathbf{T}(z,\epsilon))\right] = \mathbb{E}_\epsilon[\tau_1+\epsilon\cdot\widehat{y}+F_n(\tau_2+\epsilon x)] - c \le \mathbf{U}_{t-1}(\tau).$$

The argument also yields (by removing unnecessary steps):

$$\mathbb{E}_\epsilon[F_1(0+\epsilon x)] \le \gamma\exp\left(\frac{1}{2}\sum_{t=1}^{n}\frac{1}{t}\right) = F_0.$$

This means that

$$\mathbf{U}_0(0) = \gamma\exp\left(\frac{1}{2}\sum_{t=1}^{n}\frac{1}{t}\right) - c \le \gamma\exp(\log(n)/2) - c.$$

We will set $\gamma = \frac{1}{\sqrt{n}}$ and $c = 1$, which yields $\mathbf{U}_0(0) \le 0$. ■

### D.5. Proofs from Appendix C

**Proof** [**of Proposition 19**] Recall that the regret inequality of interest is

$$\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t) - F\left(\sum_{t=1}^{n}\overline{\mathbf{T}}(x_t)\right) \le 0.$$

As sketched in the Appendix C, Lemma 2 shows that this is implied by

$$\sup_{\boldsymbol{z}}\mathbb{E}_\epsilon\left[V\left(\sum_{t=1}^{n}\mathbf{T}(\boldsymbol{z}_t,\epsilon_t)\right)\right] \le 0, \tag{23}$$

so the remainder of this proof will focus on the opposite direction. Suppose that $\ell(\widehat{y},y) := |\widehat{y}-y|$ is the absolute loss. We fix a Rademacher sequence $\epsilon_1,\dots,\epsilon_n$ and a tree $\mathbf{x}$ with $\mathbf{x}_t(\epsilon) = \mathbf{x}_t(\epsilon_1,\dots,\epsilon_{t-1})$. As a lower bound, consider a randomized adversary that plays $y_t = \epsilon_t$ and $x_t = \mathbf{x}_t(\epsilon)$. In this case the expected value of the regret inequality is

$$\mathbb{E}_\epsilon\left[\sum_{t=1}^{n}\ell(\widehat{y}_t,\epsilon_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(\mathbf{x}_t(\epsilon)),\epsilon_t) - F\left(\sum_{t=1}^{n}\overline{\mathbf{T}}(\mathbf{x}_t(\epsilon))\right)\right].$$

Observe that for any $\epsilon \in \{\pm 1\}$ we have $\ell(\widehat{y}, \epsilon) = |1 - \widehat{y}\epsilon| \geq 1 - \widehat{y}\epsilon$. Since the range of each $f \in \mathcal{F}$ lies in $[-1, 1]$, we have $\ell(f(x), \epsilon) = 1 - f(x)\epsilon$ exactly. The expected value of the regret inequality is therefore lower bounded by

$$\begin{aligned}
&\mathbb{E}_\epsilon\left[\sum_{t=1}^n (1 - \widehat{y}_t\epsilon_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^n (1 - f(\mathbf{x}_t(\epsilon))\epsilon_t) - F\left(\sum_{t=1}^n \overline{\mathbf{T}}(\mathbf{x}_t(\epsilon))\right)\right] \\
&= \mathbb{E}_\epsilon\left[-\inf_{f \in \mathcal{F}} \sum_{t=1}^n (1 - f(\mathbf{x}_t(\epsilon))\epsilon_t) - F\left(\sum_{t=1}^n \overline{\mathbf{T}}(\mathbf{x}_t(\epsilon))\right)\right] \\
&= \mathbb{E}_\epsilon\left[\sup_{w \in \mathcal{W}} \left\langle w, \sum_{t=1}^n \epsilon_t \mathbf{x}_t(\epsilon)\right\rangle - F\left(\sum_{t=1}^n \overline{\mathbf{T}}(\mathbf{x}_t(\epsilon))\right)\right] \\
&= \mathbb{E}_\epsilon\left[V\left(\sum_{t=1}^n \mathbf{T}(\mathbf{x}_t(\epsilon), 0, \epsilon_t)\right)\right].
\end{aligned}$$

For the final step, let $\widetilde{\mathbf{y}}$ be an arbitrary $\mathcal{Y}$-valued tree $\widetilde{\mathbf{y}}_t(\epsilon) = \widetilde{\mathbf{y}}_t(\epsilon_1, \ldots, \epsilon_{t-1})$. Using the explicit form for $V$, we have

$$\begin{aligned}
\mathbb{E}_\epsilon\left[V\left(\sum_{t=1}^n \mathbf{T}(\mathbf{x}_t(\epsilon), \widetilde{\mathbf{y}}_t(\epsilon), \epsilon_t)\right)\right] &= \mathbb{E}_\epsilon\left[\sum_{t=1}^n \epsilon_t \widetilde{\mathbf{y}}_t(\epsilon) + \sup_{w \in \mathcal{W}} \left\langle w, \sum_{t=1}^n \epsilon_t \mathbf{x}_t(\epsilon)\right\rangle - F\left(\sum_{t=1}^n \overline{\mathbf{T}}(\mathbf{x}_t(\epsilon))\right)\right] \\
&= \mathbb{E}_\epsilon\left[0 + \sup_{w \in \mathcal{W}} \left\langle w, \sum_{t=1}^n \epsilon_t \mathbf{x}_t(\epsilon)\right\rangle - F\left(\sum_{t=1}^n \overline{\mathbf{T}}(\mathbf{x}_t(\epsilon))\right)\right] \\
&= \mathbb{E}_\epsilon\left[V\left(\sum_{t=1}^n \mathbf{T}(\mathbf{x}_t(\epsilon), 0, \epsilon_t)\right)\right].
\end{aligned}$$

Since the argument above holds for any trees $\mathbf{x}$ and $\widetilde{\mathbf{y}}$, we conclude that the regret inequality implies that

$$\sup_{\boldsymbol{z}} \mathbb{E}_\epsilon\left[V\left(\sum_{t=1}^n \mathbf{T}(\boldsymbol{z}_t, \epsilon_t)\right)\right] \leq 0.$$

for all $\mathcal{X} \times \mathcal{Y}$-valued trees.

■

## Appendix E. Burkholder Algorithm Implementation

### E.1. Generic Implementation

In this section we assume that $\mathcal{Y} = [-B, B]$ for $B > 0$ for simplicity. The only assumption we make on the form of $\mathbf{U}$ is Lipschitzness and boundedness.

**Assumption 1** *The are constants $K_t$ and $H_t$ such that the mapping*

$$\widehat{y} \mapsto \mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}, \partial\ell(\widehat{y}, y_t))\Big)$$

*is $K_t$-Lipschitz and bounded in magnitude by $H_t$ for any $y_t \in \mathcal{Y}$, $x_t \in \mathcal{X}$, and $\zeta_{t-1}$ of the form $\zeta_t = \sum_{s=1}^t \mathbf{T}(x_s, \widehat{y}_s, \partial(\widehat{y}_s, y_s))$.*

Consider the following strategy:

- Fix precision $\varepsilon_1 > 0$ and set $N = \lceil 2B/\varepsilon_1 \rceil$.
- Define control points $z_i = \min\{-B + \varepsilon_1 \cdot i, B\}$ for $0 \le i \le N$.
- Let $\widehat{\mu}_t$ be a solution to the convex program

$$\min_{\mu \in \Delta_N} \sup_{y \in \mathcal{Y}} \sum_{i=1}^{N} \mu_i \mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, z_i, \partial \ell(z_i, y))\Big) \tag{24}$$

  up to additive precision $\varepsilon_2$.
- Sample $\widehat{y}_t \sim \widehat{\mu}_t$.

**Proposition 20** *Given a Burkholder function* $\mathbf{U}$*, the strategy above guarantees*

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t)\right] - \phi(x_1, y_1, \ldots, x_n, y_n) \le \varepsilon_1 \sum_{t=1}^{n} K_t + \varepsilon_2 n.$$

*That is, the regret inequality (1) is obtained up to additive slack controlled by* $\varepsilon_1$ *and* $\varepsilon_2$*.*

Before proving the theorem, let us discuss the computational prospects of implementing this strategy. First, suppose $K_t = K$ and $H_t = H$ $\forall t \le n$. To obtain the regret inequality up to constant error it suffices to take $\varepsilon_1 = 1/Kn$ and $\varepsilon_2 = 1/n$. In this case, we have $N = O(BKn)$.

Now we must approximately solve (24), which is a standard finite-dimensional convex non-smooth optimization problem. There are many possible solvers; we will choose Mirror Descent (e.g. (Nemirovskii et al., 1983; Nesterov, 1998; Ben-Tal and Nemirovski, 2001)) for simplicity. Let $G(\mu) = \sup_{y \in \mathcal{Y}} \sum_{i=1}^{N} \mu_i \mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, z_i, \partial \ell(z_i, y))\Big)$. Our constraint set is $\ell_1$-bounded, and the boundedness assumption on $\mathbf{U}$ implies that $G$ is $H$-Lipschitz with respect to the $\ell_\infty$ norm. In this case, Mirror Descent with the entropic regularizer (a.k.a. multiplicative weights) guarantees an $\varepsilon$-approximate minimizer for $G(\mu)$ after $O\big(H \log(N)/\varepsilon^2\big)$ update steps, each of which requires one evaluation of the subgradient of this function.

Evaluating the subgradient of $G(\mu)$ requires computing a supremum over $y \in \mathcal{Y}$. If $\mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, z_i, \partial \ell(z_i, y))\Big)$ is convex with respect to $y$, then the supremum is obtained in $\{\pm B\}$ and so can be checked in time $O(N)$. In this case, since each Mirror Descent update takes time $O(N)$, the total complexity of the algorithm is $O(BHKn^3 \log(BKn))$.

If the supremum over $y \in \mathcal{Y}$ does not have a closed form, we can compute an approximate subgradient by taking a grid over the range $[-B, B]$ with spacing $\varepsilon'$ and computing the $\arg\max$ over this grid by brute force. If a $O(\varepsilon)$-precision solution to the convex program is required, then it suffices to set $\varepsilon' = \varepsilon/K$ and use the approximate subgradients in the Mirror Descent scheme above. The approximate subgradient computation time is $O(KN/\varepsilon)$ in this case, since we evaluate $\sum_{i=1}^{N} \mu_i \mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, z_i, \partial \ell(z_i, y))\Big)$ once per candiate $y$. The final time complexity is then $O(BHK^2 n^4 \log(BKn))$.

Lastly, we remark that if we replace Mirror Descent with Mirror Prox for saddle points (Nemirovski, 2004), the dependence on $n$ in running time for the two cases above can be improved to $O(n^2)$ and $O(n^3)$ respectively.

The runtime can improved further if a regret bound of order $O(\sqrt{n})$ is sufficient, as this requires less precision.

**Proof** [**of Proposition 20**]

To begin, observe that since $\widehat{\mu}_t$ is an approximate solution to (24), it holds that

$$\sup_{y\in\mathcal{Y}}\sum_{i=1}^{N}\widehat{\mu}_i\mathbf{U}\Big(\zeta_{t-1}+\mathbf{T}(x_t,z_i,\partial\ell(z_i,y_t))\Big)\le\inf_{\mu\in\Delta_N}\sup_{y\in\mathcal{Y}}\sum_{i=1}^{N}\mu_i\mathbf{U}\Big(\zeta_{t-1}+\mathbf{T}(x_t,z_i,\partial\ell(z_i,y_t))\Big)+\varepsilon_2.$$

The remainder of the proof will show that the right-hand-side above can be bounded as

$$\begin{aligned}&\inf_{\mu\in\Delta_N}\sup_{y\in\mathcal{Y}}\sum_{i=1}^{N}\mu_i\mathbf{U}\Big(\zeta_{t-1}+\mathbf{T}(x_t,z_i,\partial\ell(z_i,y_t))\Big)\\&\le\inf_{q\in\Delta_{\mathcal{Y}}}\sup_{y\in\mathcal{Y}}\mathbb{E}_{\widehat{y}\sim q}\,\mathbf{U}\Big(\zeta_{t-1}+\mathbf{T}(x_t,\widehat{y},\partial\ell(\widehat{y},y))\Big)+K_t\varepsilon_1\\&\le\mathbf{U}(\zeta_{t-1})+K_t\varepsilon_1,\end{aligned}$$

where the second inequality follows from property $3^o$ of $\mathbf{U}$ and was shown in the proof of Lemma 5.

The first inequality can be seen as follows. Let $q\in\Delta_{\mathcal{Y}}$ and $y\in\mathcal{Y}$ be fixed. Let $F(z):=\mathbf{U}(\zeta_{t-1},\mathbf{T}(x_t,z,\partial\ell(z,y)))$. Since $q$ is a Borel probability measure and $F$ is continuous and bounded, $F$ is integrable with respect to $q$:

$$\mathbb{E}_{\widehat{y}\sim q}\,\mathbf{U}(\zeta_{t-1},\mathbf{T}(x_t,\widehat{y},\partial\ell(\widehat{y},y)))=\int_{[-B,B]}F(z)dq(z).$$

Define $\mathcal{I}_1=[z_0,z_1]$ and $\mathcal{I}_i=(z_{i-1},z_i]$ for $2\le N$. Then $\{\mathcal{I}_i\}$ form a partition of $[-B,B]$ and the integral can be approximated as

$$\begin{aligned}\int_{[-B,B]}F(z)dq(z)&=\sum_{i=1}^{N}\int_{\mathcal{I}_i}F(z)dq(z)\\&\ge\sum_{i=1}^{N}\int_{\mathcal{I}_i}F(z_i)dq(z)-\sum_{i=1}^{N}\int_{\mathcal{I}_i}|F(z_i)-F(z)|dq(z)\\&=\sum_{i=1}^{N}q(\mathcal{I}_i)F(z_i)-\sum_{i=1}^{N}\int_{\mathcal{I}_i}|F(z_i)-F(z)|dq(z)\\&\ge\sum_{i=1}^{N}q(\mathcal{I}_i)F(z_i)-\sum_{i=1}^{N}\int_{\mathcal{I}_i}K_t\varepsilon_1dq(z)\\&=\sum_{i=1}^{N}q(\mathcal{I}_i)F(z_i)-K_t\varepsilon_1\sum_{i=1}^{N}q(\mathcal{I}_i)\\&=\sum_{i=1}^{N}q(\mathcal{I}_i)F(z_i)-K_t\varepsilon_1.\end{aligned}$$

Since this holds for any $q \in \Delta_{\mathcal{Y}}$ and $y \in \mathcal{Y}$, we have

$$\begin{aligned}
&\inf_{q\in\Delta_{\mathcal{Y}}}\sup_{y\in\mathcal{Y}}\mathbb{E}_{\widehat{y}\sim q}\,\mathbf{U}\Big(\zeta_{t-1}+\mathbf{T}(x_t,\widehat{y},\partial\ell(\widehat{y},y))\Big)\\
&\geq \inf_{q\in\Delta_{\mathcal{Y}}}\sup_{y\in\mathcal{Y}}\sum_{i=1}^{n}q(\mathcal{I}_i)\mathbf{U}\Big(\zeta_{t-1}+\mathbf{T}(x_t,z_i,\partial\ell(z_i,y))\Big)-K_t\varepsilon_1\\
&= \inf_{\mu\in\Delta_{N}}\sup_{y\in\mathcal{Y}}\sum_{i=1}^{n}\mu_i\mathbf{U}\Big(\zeta_{t-1}+\mathbf{T}(x_t,z_i,\partial\ell(z_i,y))\Big)-K_t\varepsilon_1.
\end{aligned}$$

■

### E.2. Faster Implementation under Specific Structure

In the remainder of this section of the appendix we show how to implement the Burkholder algorithm for certain special cases that enable admit especially simple strategies.

**Lemma 21** *Suppose that the map*

$$\widehat{y}\mapsto\mathbf{U}(\tau+\mathbf{T}((x,\widehat{y}),\partial(\widehat{y},y)))$$

*is convex for all $y$. Then the strategy*

$$\widehat{y}_t=\arg\min_{\widehat{y}\in\mathcal{Y}}\sup_{y\in\mathcal{Y}}\,\mathbf{U}\left(\sum_{j=1}^{t-1}\zeta_{t-1}+\mathbf{T}(x_t,\widehat{y},\partial\ell(\widehat{y},y))\right) \tag{25}$$

*achieves the value of the game in Lemma 5.*

**Proof** [**of Lemma 21**] This follows by reduction to the general case:

$$\begin{aligned}
\inf_{\widehat{y}\in\mathcal{Y}}\sup_{y\in\mathcal{Y}}\mathbf{U}\left(\zeta_{t-1}+\mathbf{T}(x_t,\widehat{y},\partial\ell(\widehat{y},y))\right)&=\inf_{q\in\Delta_{\mathcal{Y}}}\sup_{y\in\mathcal{Y}}\mathbf{U}\left(\zeta_{t-1}+\mathbf{T}(x_t,\mathbb{E}_{\widehat{y}\sim q}[\widehat{y}],\partial\ell(\mathbb{E}_{\widehat{y}\sim q}[\widehat{y}],y))\right)\\
&\leq\inf_{q\in\Delta_{\mathcal{Y}}}\sup_{y\in\mathcal{Y}}\mathbb{E}_{\widehat{y}\sim q}\,\mathbf{U}\left(\zeta_{t-1}+\mathbf{T}(x_t,\widehat{y},\partial\ell(\widehat{y},y))\right).
\end{aligned}$$

The strategy in (25) is the minimax strategy for second expression above. The final expression is precisely the value of the Burkholder algorithm, which is controlled when $\mathbf{U}$ is a Burkholder function via Lemma 5. ■

**Lemma 22** *Suppose that $\mathcal{Y}=[-B,B]$ for some $B>0$. Further suppose that we can write*

$$\mathbf{U}(\tau+\mathbf{T}((x,\widehat{y}),\delta))=\widehat{y}\cdot\delta+F(\tau,x,\delta),$$

*where $\delta\mapsto F(\tau,x,\delta)$ is convex for all $\tau,x$. Then the prediction strategy*

$$\widehat{y}_t=\mathrm{proj}_{[-B,B]}\Big(-\frac{1}{L}\,\mathbb{E}_{\sigma\in\{\pm1\}}[\sigma F(\zeta_{t-1},x_t,L\sigma)]\Big), \tag{26}$$

*achieves the value of the game in Lemma 5.*

**Proof** [**of Lemma 22**] Let $\widetilde{y}_t$ denote the unprojected version of $\widehat{y}_t$:

$$\widetilde{y}_t = -\frac{1}{L}\,\mathbb{E}_{\sigma\in\{\pm 1\}}[\sigma F(\zeta_{t-1}, x_t, L\sigma)].$$

We prove the lemma by inducting backwards. Let $t \in [n]$ be fixed. We first claim that

$$\begin{aligned}\sup_{y\in\mathcal{Y}} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}_t, \partial\ell(\widehat{y}_t, y))\right) &= \sup_{y\in\mathcal{Y}}[\widehat{y}_t \cdot \partial\ell(\widehat{y}_t, y) + F(\zeta_{t-1}, x_t, \partial\ell(\widehat{y}_t, y))] \\ &\leq \sup_{y\in\mathcal{Y}}[\widetilde{y}_t \cdot \partial\ell(\widehat{y}_t, y) + F(\zeta_{t-1}, x_t, \partial\ell(\widehat{y}_t, y))].\end{aligned}$$

This holds by the assumption that $\arg\min_{\widehat{y}\in\mathbb{R}} \ell(\widehat{y}, y)$ is obtained in $[-B, B]$ for any $y$. The assumption implies that for any $y$, $\partial\ell(\widehat{y}, y) \geq 0$ for $\widehat{y} \geq B$ and $\partial\ell(\widehat{y}, y) \leq 0$ for $\widehat{y} \leq -B$. If $\widehat{y}_t \neq \widetilde{y}_t$, then either $\widehat{y}_t = B$ and $\widetilde{y}_t > B$, so that $\partial\ell(\widehat{y}_t, y)\widehat{y}_t \leq \partial\ell(\widehat{y}_t, y)\widetilde{y}_t$, or similarly $\widehat{y}_t = -B$ and $\widetilde{y}_t < -B$, which also implies $\partial\ell(\widehat{y}_t, y)\widehat{y}_t \leq \partial\ell(\widehat{y}_t, y)\widetilde{y}_t$.

Now, by the convexity assumption of the lemma, it holds that

$$\begin{aligned}\sup_{y\in\mathcal{Y}}[\widetilde{y}_t \cdot \partial\ell(\widehat{y}_t, y) + F(\zeta_{t-1}, x_t, \partial\ell(\widehat{y}_t, y))] &\leq \sup_{\delta\in[-L,L]}[\widetilde{y}_t \cdot \delta + F(\zeta_{t-1}, x_t, \delta)] \\ &= \max_{\sigma\in\{\pm1\}}[\widetilde{y}_t \cdot L\sigma + F(\zeta_{t-1}, x_t, L\sigma)].\end{aligned}$$

The choice of $\widetilde{y}_t$ guarantees that $\widetilde{y}_t \cdot L \cdot (1) + F(\zeta_{t-1}, x_t, L \cdot (1)) = \widetilde{y}_t \cdot L \cdot (-1) + F(\zeta_{t-1}, x_t, L \cdot (-1))$; this can be seen by rearranging this equality and solving for $\widetilde{y}_t$. This means that we can take $\sigma = 1$ to obtain the maximum in the expression above. Substituting in the value of $\widetilde{y}_t$ then yields

$$\max_{\sigma\in\{\pm1\}}[\widetilde{y}_t \cdot L\sigma + F(\zeta_{t-1}, x_t, L\sigma)] = \widetilde{y}_t \cdot L \cdot (1) + F(\zeta_{t-1}, x_t, L \cdot (1)) = \mathbb{E}_{\sigma\in\{\pm1\}}[F(\zeta_{t-1}, x_t, \sigma L)].$$

Finally, we use property $3'$ of $\mathbf{U}$ and the explicit form for $\mathbf{U}$ assumed in the lemma statement to proceed back to time $t-1$:

$$\begin{aligned}\mathbb{E}_{\sigma\in\{\pm1\}}[F(\zeta_{t-1}, x_t, \sigma L)] &= \mathbb{E}_{\sigma\in\{\pm1\}}[\widehat{y}_t \sigma L + F(\zeta_{t-1}, x_t, \sigma L)] \\ &= \mathbb{E}_{\sigma\in\{\pm1\}}\,\mathbf{U}(\zeta_{t-1} + \mathbf{T}((x_t, \widehat{y}_t), \sigma L)) \\ &\leq \mathbf{U}(\zeta_{t-1}).\end{aligned}$$

■

## Appendix F. Algebra of Burkholder Functions

This appendix contains some additional structural results about Burkholder functions which may be useful for algorithm designers.

**Proposition 23** *The following statements are true:*

1. *Given a Burkholder function* $\mathbf{U}$, *if we define the* $X_t = \mathbf{U}(\sum_{j=1}^{t} \mathbf{T}(z_j, \delta_j))$, *then for any real-valued martingale difference sequence* $\delta_t$*s and predictable* $z_t$*s,* $(X_t)_{t\geq0}$ *is a supermartingale with* $\mathbb{E}[X_0] \leq 0$.

2. *Any convex combination of Burkholder functions is a Burkholder function.*
3. *The minimum of a family of Burkholder functions is a Burkholder function.*
4. *Suppose we have a finite set $A$ that indexes a family of functions $V_a : \mathcal{T} \to \mathbb{R}$, each of which belongs to a sufficient statistic pair $(\mathbf{T}, V_a)$ for some regret inequality of interest, and suppose each $V_a$ has a corresponding Burkholder function $\mathbf{U}_a$. Then the following probabilistic inequality is true:*

$$\mathbb{E}\left[\max_{a\in A}\left\{V_a\left(\sum_{t=1}^{n}\mathbf{T}(z_t,\delta_t)\right) - \eta n C[a]\right\}\right] \le \frac{1}{\eta}\log|A|,$$

*where $C[a] = \sup_{\tau,z,\alpha}(\mathbf{U}_a(\tau+\mathbf{T}(z,\alpha)) - \mathbf{U}_a(\tau))^2$. Note that $C \in \mathbb{R}^A$ may be thought as a sufficient statistic, though it is fixed and does not depend on instances. Furthermore, a Burkholder function $\mathbf{U} : \mathcal{T}\times\mathbb{R}^A \to \mathbb{R}$ that certifies this inequality is:*

$$\mathbf{U}(\tau,\gamma) = \frac{1}{\eta}\log\left(\sum_{a\in A}\exp\left(\eta\mathbf{U}_a(\tau) - \eta^2\gamma[a]\right)\right) - \frac{\log|A|}{\eta} \tag{27}$$

**Proof** [**of Proposition 23**] The first statement follows from property $3^o$ of the Burkholder function $\mathbf{U}$, which immediately implies that it is a supermartingale. The second statement is trivial. To prove the third statement it suffices to verify property $3^o$, which holds due to concavity of the minimum.

We now prove the fourth statement. Given a family of Burkholder functions $\{\mathbf{U}_a\}_{a\in A}$, define a new Burkholder function $\mathbf{U} : \mathcal{T}\times\mathbb{R}^A \to \mathbb{R}$ as:

$$\mathbf{U}(\tau,\gamma) = \frac{1}{\eta}\log\left(\sum_{a\in A}\exp\left(\eta\mathbf{U}_a(\tau) - \eta^2\gamma[a]\right)\right) - \frac{\log|A|}{\eta}.$$

whose sufficient statistics are the original sufficient statistic of the family of $V_a$s along with an additional $|A|$-dimensional real vector, for which one coordinate per $a \in A$ will be used to represent $C[a] = \sup_{\tau,z,\alpha}(\mathbf{U}_a(\tau+\mathbf{T}(z,\alpha)) - \mathbf{U}_a(\tau))^2$ (note that this is a vacuous statistic as it is constant for each instance). Property $3^o$ for $\mathbf{U}$ holds as follows:

$$\begin{aligned}
&\mathbb{E}_\alpha\mathbf{U}\left((\tau,\gamma) + (\mathbf{T}(z,\alpha), C)\right)\\
&\quad= \frac{1}{\eta}\mathbb{E}_\alpha\log\left(\sum_{a\in A}\exp\left(\eta\mathbf{U}_a(\tau+\mathbf{T}(z,\alpha)) - \eta^2\gamma[a] - \eta^2 C[a]\right)\right) - \frac{\log|A|}{\eta}\\
&\quad\le \frac{1}{\eta}\log\left(\sum_{a\in A}\mathbb{E}_\alpha\exp\left(\eta\mathbf{U}_a(\tau+\mathbf{T}(z,\alpha)) - \eta^2\gamma[a] - \eta^2 C[a]\right)\right) - \frac{\log|A|}{\eta}\\
&\quad= \frac{1}{\eta}\log\left(\sum_{a\in A}\mathbb{E}_\alpha\exp\left(\eta\left(\mathbf{U}_a(\tau+\mathbf{T}(z,\alpha)) - \mathbf{U}_a(\tau)\right) + \eta\mathbf{U}_a(\tau) - \eta^2\gamma[a] - \eta^2 C[a]\right)\right) - \frac{\log|A|}{\eta}.
\end{aligned}$$

Now note that by property $3^o$ of the Burkholder functions $\{\mathbf{U}_a\}_{a\in A}$, the random variable $X_a = (\mathbf{U}_a(\tau+\mathbf{T}(z,\alpha)) - \mathbf{U}_a(\tau))$ is such that $\mathbb{E}_\alpha[X_a] \le 0$. Further from our assumption we have that $|X_a|^2 \le C[a]$. Hence, the standard mgf bound implies $\mathbb{E}_\alpha[\exp(\eta X_a)] \le \exp(\eta^2 C[a]/2)$.

$$\begin{aligned}
&\le \frac{1}{\eta}\log\left(\sum_{a\in A}\exp\left(\eta\mathbf{U}_a(\tau) + \frac{\eta^2}{2}C[a] - \eta^2\gamma[a] - \eta^2 C[a]\right)\right) - \frac{\log|A|}{\eta}\\
&\le \frac{1}{\eta}\log\left(\sum_{a\in A}\exp\left(\eta\mathbf{U}_a(\tau) - \eta^2\gamma[a]\right)\right) - \frac{\log|A|}{\eta}.
\end{aligned}$$

For property $1^o$ it can be seen immediately that $\mathbf{U}(0) \le 0$. Property $2^o$ holds via

$$\begin{aligned}
\mathbf{U}(\tau, \gamma) &= \frac{1}{\eta} \log\left( \sum_{a \in A} \exp\left( \eta \mathbf{U}_a(\tau) - \eta^2 \gamma[a] \right) \right) - \frac{\log|A|}{\eta} \\
&\ge \max_{a \in A} \{ \mathbf{U}_a(\tau) - \eta \gamma[a] \} - \frac{\log|A|}{\eta} \qquad \text{(softmax upper bounds max)} \\
&\ge \max_{a \in A} \{ V_a(\tau) - \eta \gamma[a] \} - \frac{\log|A|}{\eta}.
\end{aligned}$$

■

We remark that one uses non-additive sufficient statistics as discussed in Appendix A, then one can make the bound implied by the Burkholder function $\mathbf{U}$ above more data-dependent by replacing $C[a]$ with $\sup_\delta \left( \mathbf{U}_a(\tau + \mathbf{T}(z, \delta)) - \mathbf{U}_a(\tau) \right)^2$ for each $a$.