Journal of Machine Learning Research 18 (2017) 1-31 Submitted 2/17; Revised 11/17; Published 12/17

Following the Leader and Fast Rates in Online Linear Prediction: Curved Constraint Sets and Other Regularities[*]

Ruitong Huang[†] RUITONG@UALBERTA.CA
Department of Computing Science
University of Alberta
Edmonton T6G 2E8, Canada

Tor Lattimore[‡] TOR.LATTIMORE@GMAIL.COM
Independent Researcher
Canberra, Australia

András György A.GYORGY@IMPERIAL.AC.UK
Department of Electrical and Electronic Engineering
Imperial College London
South Kensington Campus, London SW7 2BT, UK

Csaba Szepesvári SZEPESVA@CS.UALBERTA.CA
Department of Computing Science
University of Alberta
Edmonton T6G 2E8, Canada

Editor: Manfred Warmuth

Abstract

Follow the leader (FTL) is a simple online learning algorithm that is known to perform well when the loss functions are convex and positively curved. In this paper we ask whether there are other settings when FTL achieves low regret. In particular, we study the fundamental problem of linear prediction over a convex, compact domain with non-empty interior. Amongst other results, we prove that the curvature of the boundary of the domain can act as if the losses were curved: In this case, we prove that as long as the mean of the loss vectors have positive lengths bounded away from zero, FTL enjoys logarithmic regret, while for polytope domains and stochastic data it enjoys finite expected regret. The former result is also extended to strongly convex domains by establishing an equivalence between the strong convexity of sets and the minimum curvature of their boundary, which may be of independent interest. Building on a previously known meta-algorithm, we also get an algorithm that simultaneously enjoys the worst-case guarantees and the smaller regret of FTL when the data is 'easy'. Finally, we show that such guarantees are achievable directly (e.g., by the follow the regularized leader algorithm or by a shrinkage-based variant of FTL) when the constraint set is an ellipsoid.

Keywords: online linear optimization, follow the leader, logarithmic regret, strongly convex decision set, curvature

1. Introduction

Learning theory traditionally has been studied in a statistical framework, discussed at length, for example, by Shalev-Shwartz and Ben-David (2014). The issue with this approach is that the analysis of the performance of learning methods seems to critically depend on whether the data generating mechanism satisfies some probabilistic assumptions. Realizing that these assumptions

[*] An earlier version of this paper was published in NIPS 2016 (Huang et al., 2016).
[†] Ruitong Huang is now at Borealis AI, Edmonton, Canada.
[‡] Tor Lattimore is now at DeepMind, London, UK.

are not necessarily critical, much work has been devoted recently to studying learning algorithms in the so-called online learning framework (Cesa-Bianchi and Lugosi, 2006). The online learning framework makes minimal assumptions about the data generating mechanism, while allowing one to replicate results of the statistical framework through online-to-batch conversions (Cesa-Bianchi et al., 2004). By following a minimax approach, however, results proven in the online learning setting, at least initially, led to rather conservative results and algorithm designs, failing to capture how more regular, "easier" data, may give rise to faster learning. This is problematic as it may suggest overly conservative learning strategies, missing opportunities to extract more information when the data is nicer. Also, it is hard to argue that data resulting from passive data collection, such as weather data, would ever be adversarially generated (though it is equally hard to defend that such data satisfies precise stochastic assumptions). Realizing this issue, during recent years much work has been devoted to understanding what regularities and how can lead to faster learning speed. For example, much work has been devoted to showing that faster learning speed (smaller "regret") can be achieved in the online convex optimization setting when the loss functions are "curved", such as when the loss functions are strongly convex or exp-concave, or when the losses show small variations, or the best prediction in hindsight has a small total loss, and that these properties can be exploited in an adaptive manner (e.g., Merhav and Feder 1992, Freund and Schapire 1997, Gaivoronski and Stella 2000, Cesa-Bianchi and Lugosi 2006, Hazan et al. 2007, Bartlett et al. 2007, Kakade and Shalev-Shwartz 2009, Orabona et al. 2012, Rakhlin and Sridharan 2013, van Erven et al. 2015, Foster et al. 2015).

In this paper we contribute to this growing literature by studying online linear prediction and the follow the leader (FTL) algorithm. Online linear prediction is arguably the simplest yet fundamental of all the learning settings, and lies at the heart of online convex optimization, while it also serves as an abstraction of core learning problems such as prediction with expert advice. FTL, the online analogue of empirical risk minimization of statistical learning, is the simplest learning strategy, one can think of. Although the linear setting removes the possibility of exploiting the curvature of losses, there are multiple ways online learning problems can present data that allows for small regret, even for FTL. As is well known, in the worst case, FTL suffers a linear regret (e.g., Example 2.2 of Shalev-Shwartz 2012). However, for "curved" losses (e.g., exp-concave losses), FTL was shown to achieve small (logarithmic) regret (see, e.g., Merhav and Feder 1992; Cesa-Bianchi and Lugosi 2006; Gaivoronski and Stella 2000; Hazan et al. 2007).

We take a thorough look at FTL in the case when the losses are linear, but the problem perhaps exhibits other regularities. The motivation comes from the simple observation that, for prediction over the simplex, when the loss vectors are selected independently of each other from a distribution with a bounded support with a nonzero mean, FTL quickly locks onto selecting the loss-minimizing vertex of the simplex, achieving finite expected regret. In this case, FTL is an excellent algorithm. In fact, FTL is shown to be the minimax optimizer for the binary losses in the stochastic expert setting in the paper of Kotłowski (2016). Thus, we ask the question of whether there are other regularities that allow FTL to achieve nontrivial performance guarantees. Our main result shows that when the decision set (or constraint set) has a sufficiently "curved" boundary (or equivalently, if it is strongly convex) and the linear loss is bounded away from 0, FTL is able to achieve logarithmic regret even in the adversarial setting, thus opening up a new way to prove fast rates not based on the curvature of losses, but on that of the boundary of the constraint set and non-singularity of the linear loss. In a matching lower bound we show that this regret bound is essentially unimprovable. We also show an alternate bound for polytope constraint sets, which allows us to prove that (under certain technical conditions) for stochastic problems the expected regret of FTL will be finite. To finish, we use $(\mathcal{A}, \mathcal{B})$-prod of Sani et al. (2014) to design an algorithm that adaptively interpolates between the worst case $O(\sqrt{n \log n})$ regret and the smaller regret bounds, which we prove here for "easy data." We also show that if the constraint set is an ellipsoid, both the follow the regularized leader (FTRL) algorithm and a combination of FTL and shrinkage, which we call follow the shrunken leader

(FTSL), achieve logarithmic regret for easy data. Simulation results on artificial data complement the theoretical findings.

While we believe that we are the first to point out that the curvature of the constraint set $\mathcal{W}$ can help in speeding up learning, this effect is known in convex optimization since at least the work of Levitin and Polyak (1966), who showed that exponential rates are attainable for strongly convex constraint sets if the norm of the gradients of the objective function admit a uniform lower bound. More recently, Garber and Hazan (2015) proved an $O(1/n^2)$ optimization error bound (with problem-dependent constants) for the Frank-Wolfe algorithm for strongly convex and smooth objectives and strongly convex constraint sets. The effect of the shape of the constraint set was also discussed by Abbasi-Yadkori (2009) who demonstrated $O(\sqrt{n})$ regret in the linear bandit setting. Although at a high level these results are similar to ours, our proof technique is rather different.

2. Preliminaries, Online Learning and the Follow the Leader Algorithm

We consider the standard framework of online convex optimization, where a learner and an environment interact in a sequential manner over n rounds: In every round $t = 1, \ldots, n$, first the learner predicts $w_t \in \mathcal{W}$. Then the environment picks a loss function $\ell_t \in \mathcal{L}$, and the learner suffers loss $\ell_t(w_t)$ and observes ℓ_t. Here, $\mathcal{W}$ is a compact and convex subset of the d-dimensional Euclidean space $\mathbb{R}^d$ with non-empty interior, and $\mathcal{L}$ is a set of convex functions mapping $\mathcal{W}$ to the reals. The elements of $\mathcal{L}$ are called loss functions. The performance of the learner is measured in terms of its regret,

$$R_n = \sum_{t=1}^{n} \ell_t(w_t) - \min_{w\in\mathcal{W}} \sum_{t=1}^{n} \ell_t(w) \,.$$

The simplest possible case, which will be the focus of this paper, is when the losses are linear, that is, when $\ell_t(w) = \langle f_t, w\rangle$ for some $f_t \in \mathcal{F} \subset \mathbb{R}^d$. In fact, the linear case is not only simple, but is also fundamental since the case of nonlinear loss functions can be reduced to it: Indeed, even if the losses are nonlinear, defining $f_t \in \partial\ell_t(w_t)$ to be a subgradient[1] of ℓ_t at w_t and letting $\tilde{\ell}_t(u) = \langle f_t, u\rangle$, by the definition of subgradients, $\ell_t(w_t) - \ell_t(u) \le \ell_t(w_t) - (\ell_t(w_t) + \langle f_t, u - w_t\rangle) = \tilde{\ell}_t(w_t) - \tilde{\ell}_t(u)$, hence for any $u \in \mathcal{W}$,

$$\sum_t \ell_t(w_t) - \sum_t \ell_t(u) \le \sum_t \tilde{\ell}_t(w_t) - \sum_t \tilde{\ell}_t(u) \,.$$

In particular, if an algorithm keeps the regret small no matter how the linear losses are selected (even when allowing the environment to pick losses based on the choices of the learner), the algorithm can also be used to keep the regret small in the nonlinear case.

Hence, in what follows we will study the linear case $\ell_t(w) = \langle f_t, w\rangle$ and, in particular, we will study the regret of the so-called "Follow The Leader" (FTL) learner, which in round $t \ge 2$ picks

$$w_t = \operatorname*{argmin}_{w\in\mathcal{W}} \sum_{i=1}^{t-1} \ell_i(w) \,.$$

For the first round, $w_1 \in \mathcal{W}$ is picked in an arbitrary manner. When $\mathcal{W}$ is compact, the optimal w of $\min_{w\in\mathcal{W}} \sum_{i=1}^{t-1} \langle w, f_t\rangle$ is attainable, which we will assume henceforth. If multiple minimizers exist, we simply fix one of them as w_t. We will also assume that $\mathcal{F}$ is non-empty, compact and convex.

One problem of the linearization technique is that if some algorithm's performance depends on some additional properties of the linear loss function, linearization may not preserve these and could lead to suboptimal performance. For example, if the loss functions are strongly convex and the optimum in hindsight (in fact, w_{n+1}) is an inner point of $\mathcal{W}$, FTL has no chance to do well, since

[1] We let $\partial g(x)$ denote the subdifferential of a convex function $g : \mathrm{dom}(g) \to \mathbb{R}$ at x, that is, $\partial g(x) = \left\{\theta \in \mathbb{R}^d \,|\, g(x') \ge g(x) + \langle \theta, x' - x\rangle \; \forall x' \in \mathrm{dom}(g)\right\}$, where $\mathrm{dom}(g) \subset \mathbb{R}^d$ is the domain of g.

it will always predict points on the boundary. Thus, while our results extend from linear losses to arbitrary convex functions, some of the conditions of our regret bounds may be violated or the constants in the bounds might blow up, possibly leading to trivial or weak regret bounds. Thus, in practice, one should always check if the linearization step makes sense. On the positive side, no problem occurs if the optimum is outside of $\mathcal{W}$.

2.1 Support Functions

Let $\Theta_t = -\frac{1}{t}\sum_{i=1}^{t} f_i$ be the negative average of the first t vectors in $(f_t)_{t=1}^n$, $f_t \in \mathcal{F}$. For convenience, we define $\Theta_0 := 0$. Thus, for $t \geq 2$,

$$w_t = \operatorname*{argmin}_{w\in\mathcal{W}} \sum_{i=1}^{t-1} \langle w, f_i\rangle = \operatorname*{argmin}_{w\in\mathcal{W}} \langle w, -\Theta_{t-1}\rangle = \operatorname*{argmax}_{w\in\mathcal{W}} \langle w, \Theta_{t-1}\rangle\,.$$

Denote by $\Phi(\Theta) = \max_{w\in\mathcal{W}}\langle w, \Theta\rangle$ the so-called *support function* of $\mathcal{W}$. The support function, being the maximum of linear and hence convex functions, is itself convex. Further Φ is positive homogenous: for $a \geq 0$ and $\theta \in \mathbb{R}^d$, $\Phi(a\theta) = a\Phi(\theta)$. It follows then that the epigraph $\operatorname{epi}(\Phi) = \{(\theta, z)\,|\, z \geq \Phi(\theta), z \in \mathbb{R}, \theta \in \mathbb{R}^d\}$ of Φ is a cone, since for any $(\theta, z) \in \operatorname{epi}(\Phi)$ and $a \geq 0$, $az \geq a\Phi(\theta) = \Phi(a\theta)$, $(a\theta, az) \in \operatorname{epi}(\Phi)$ also holds.

The differentiability of the support function is closely tied to whether in the FTL algorithm the choice of w_t is uniquely determined:

Proposition 1 *Let $\mathcal{W} \neq \emptyset$ be convex and closed. Fix Θ and let $\mathcal{Z} := \{w \in \mathcal{W}\,|\,\langle w, \Theta\rangle = \Phi(\Theta)\}$. Then, $\partial\Phi(\Theta) = \mathcal{Z}$ and, in particular, $\Phi(\Theta)$ is differentiable at Θ if and only if $\max_{w\in\mathcal{W}}\langle w, \Theta\rangle$ has a unique optimizer. In this case, $\nabla\Phi(\Theta) = \operatorname{argmax}_{w\in\mathcal{W}}\langle w, \Theta\rangle$.*

The proposition follows from Danskin's theorem when $\mathcal{W}$ is compact (e.g., Proposition B.25 of Bertsekas 1999), but a simple direct argument, presented in Appendix A.1 for completeness, can also be used to show that it also remains true even when $\mathcal{W}$ is unbounded. By Proposition 1, when Φ is differentiable at Θ_{t-1}, $w_t = \nabla\Phi(\Theta_{t-1})$.

2.2 A Motivating Example

We close this section with an example demonstrating how fast rates can be achieved by the FTL algorithm. Consider the case when the losses are independent and identically distributed (i.i.d.), which means that (f_t) is an i.i.d. sequence with expectation $\mu \in \mathbb{R}^d$. Then $\mathbb{E}[\Theta_t] = -\mu$, and we have $\|\Theta_t + \mu\|_2 = O(1/\sqrt{t})$ with high probability. For $\mathcal{W}$ being the unit ball of $\mathbb{R}^d$ one has $w_t = \Theta_t / \|\Theta_t\|_2$ and therefore a crude bound suggests that $\|w_t - w^*\|_2 = O(1/\sqrt{t})$ where w^* is the optimal decision in hindsight, overall predicting that $\mathbb{E}[R_n] = O(\sqrt{n})$. On the other hand, in the rest of the paper we provide conditions when the expected regret can be much smaller than this. Below we give a simple geometric explanation how it can happen.

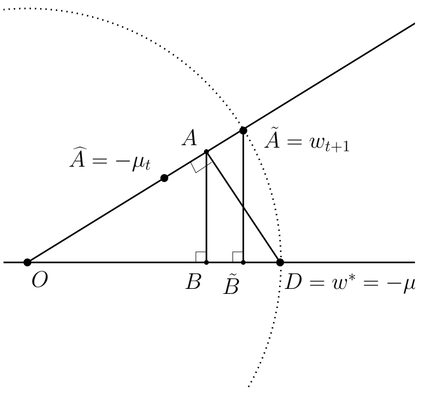

Figure 1: Illustration of how fast rates can be achieved by FTL.

Let $\mathcal{W} = \{w\,|\,\|w\|_2 \leq 1\}$ and consider a stochastic setting where the f_t are i.i.d. samples with expectation $\mathbb{E}[f_t] = \mu = (-1, 0, \ldots, 0)$ and $\|f_t\|_\infty \leq M$ almost surely. It is straightforward to see that $w^* = (1, 0, \ldots, 0)$, and thus $\langle w^*, \mu\rangle = -1$. Let $\mu_t = -\Theta_t$ denote our estimate of μ after t time steps; then $\|\mu_t - \mu\| = O(1/\sqrt{t})$ with high probability. Now consider Fig. 1: The origin is denoted by O, the optimal prediction $w^* = -\mu$ by D, and $-\mu_t$ by $\hat{A}$. Then the prediction of FTL at time t is $\tilde{A}$, the intersection of the line connecting O and $\hat{A}$ with the unit sphere, and its instantaneous excess loss is $\langle \overrightarrow{O\tilde{A}}, \overrightarrow{OD} \rangle - 1 = |\overline{\tilde{B}D}|$ where $\tilde{B}$ is the orthogonal projection of $\tilde{A}$ to $\overline{OD}$. Next we give a simple geometric argument showing that if $|\overline{\hat{A}D}| \le \epsilon$ then $|\overline{\tilde{B}D}| \le \epsilon^2$. Since $|\overline{\hat{A}D}| = \|\mu_t - \mu\| = O(1/\sqrt{t})$ with high probability, this means that the excess error at time t is $O(1/t)$, making the regret $O(\log n)$ in n time steps, much smaller than the previously anticipated $O(\sqrt{n})$ regret. To finish, let A denote the orthogonal projection of D to the line connecting O and $\hat{A}$; then the Pythagorean theorem implies that $|\overline{OA}| \le |\overline{OD}| = |\overline{O\tilde{A}}|$, and so $A \in \overline{O\tilde{A}}$. Therefore, the orthogonal projection of A to $\overline{OD}$, denoted by B, belongs to the segment $\overline{O\tilde{B}}$, and so $|\overline{BD}| \ge |\overline{\tilde{B}D}|$. Since the triangles OAD and ABD are similar, we have $\frac{|\overline{BD}|}{|\overline{AD}|} = \frac{|\overline{AD}|}{|\overline{OD}|}$. Therefore, $|\overline{BD}| \le |\overline{AD}|^2 \le |\overline{\hat{A}D}|^2$ (by the definition of A), implying $|\overline{\tilde{B}D}| \le |\overline{\hat{A}D}|^2$, which we wanted to prove.

3. Non-Stochastic Analysis of FTL

We start by rewriting the regret of FTL in an equivalent form, which shows that we can expect FTL to enjoy a small regret when successive weight vectors move little.

Proposition 2 *The regret R_n of FTL satisfies the following identity:*

$$R_n = \sum_{t=1}^{n} t \, \langle w_{t+1} - w_t, \Theta_t \rangle \, .$$

The result is a direct corollary of Lemma 9 of McMahan (2010), which holds for any sequence of losses (even non-convex). It is also a tightening of the well-known inequality $R_n \le \sum_{t=1}^{n} \ell_t(w_t) - \ell_t(w_{t+1})$, which again holds for arbitrary loss sequences (e.g., Lemma 2.1 of Shalev-Shwartz, 2012). To keep the paper self-contained, we give a direct proof based on the summation by parts formula:

Proof The summation by parts formula states that for any $u_1, v_1, \ldots, u_{n+1}, v_{n+1}$ reals, $\sum_{t=1}^{n} u_t \, (v_{t+1} - v_t) = (u_{t+1} v_{t+1} - u_1 v_1) - \sum_{t=1}^{n} (u_{t+1} - u_t) \, v_{t+1}$. Applying this to the definition of regret with $u_t := w_{t,\cdot}$ and $v_{t+1} := t\Theta_t$, we get

$$\begin{aligned} R_n &= -\sum_{t=1}^{n} \langle w_t, t\Theta_t - (t-1)\Theta_{t-1} \rangle + \langle w_{n+1}, n\Theta_n \rangle \\ &= -\left\{ \cancel{\langle w_{n+1}, n\Theta_n \rangle} - 0 - \sum_{t=1}^{n} \langle w_{t+1} - w_t, t\Theta_t \rangle \right\} + \cancel{\langle w_{n+1}, n\Theta_n \rangle}. \end{aligned}$$

■

Our next proposition gives another identity for the regret. Although this formula is not directly needed for the rest of the paper, it provides interesting insights: as opposed to the previous result, it is independent of w_t, and directly connects the sequence $(\Theta_t)_t$ to the geometric properties of $\mathcal{W}$ through the support function Φ. A similar expression for a general "Follow the Regularized Leader" algorithm was also derived by Abernethy et al. (2014). For this proposition, we will momentarily assume that Φ is differentiable at $(\Theta_t)_{t \ge 1}$.

Proposition 3 *Assume Φ is differentiable at $\Theta_1, \ldots, \Theta_n$. Then*

$$R_n = \sum_{t=1}^{n} t \, D_\Phi(\Theta_t, \Theta_{t-1}) \, , \tag{1}$$

where $D_\Phi(\theta',\theta) = \Phi(\theta') - \Phi(\theta) - \langle \nabla\Phi(\theta), \theta' - \theta\rangle$ *is the Bregman divergence of* Φ *and we use the convention that* $\nabla\Phi(0) = w_1$*.*

Proof Let $v = \operatorname{argmax}_{w\in\mathcal{W}}\langle w,\theta\rangle$, $v' = \operatorname{argmax}_{w\in\mathcal{W}}\langle w,\theta'\rangle$. When Φ is differentiable at θ,

$$D_\Phi(\theta',\theta) = \Phi(\theta') - \Phi(\theta) - \langle \nabla\Phi(\theta), \theta' - \theta\rangle = \langle v',\theta'\rangle - \langle v,\theta\rangle - \langle v,\theta'-\theta\rangle = \langle v'-v,\theta'\rangle\,. \tag{2}$$

Therefore, by Proposition 2, $R_n = \sum_{t=1}^n t\langle w_{t+1} - w_t, \Theta_t\rangle = \sum_{t=1}^n t\, D_\Phi(\Theta_t, \Theta_{t-1})$. ■

When Φ is non-differentiable at some of the points $\Theta_1,\dots,\Theta_n$, the equality in the above proposition can be replaced with inequalities. Defining the upper Bregman divergence $\overline{D}_\Phi(\theta',\theta) = \sup_{w\in\partial\Phi(\theta)} \Phi(\theta') - \Phi(\theta) - \langle w, \theta'-\theta\rangle$ and the lower Bregman divergence $\underline{D}_\Phi(\theta',\theta)$ similarly with inf instead of sup, we can easily obtain an analogue of Proposition 3:

$$\sum_{t=1}^n t\,\underline{D}_\Phi(\Theta_t,\Theta_{t-1}) \le R_n \le \sum_{t=1}^n t\,\overline{D}_\Phi(\Theta_t,\Theta_{t-1})\,. \tag{3}$$

3.1 Constraint Sets with Positive Curvature

The previous results show in an implicit fashion that the curvature of $\mathcal{W}$ controls the regret. Before presenting our first main results, which make this connection explicit, we define some basic notions from differential geometry related to the curvature, while some extra details are presented in Appendix A.2 (all differential geometry concept and results that we need can be found in Section 2.5 of the book of Schneider, 2014).

3.1.1 Curvature and Strong Convexity

Given a twice continuously differentiable planar curve γ in $\mathbb{R}^2$, there exists a parametrization with respect to the curve length s, such that $\|\gamma'(s)\| = \|\left(x'(s), y'(s)\right)\| = \sqrt{x'(s)^2 + y'(s)^2} = 1$. Under the curve length parametrization, the curvature of γ at $\gamma(s)$ is $\|\gamma''(s)\|$. Define the unit normal vector $\mathbf{n}(s)$ as the unit vector that is perpendicular to $\gamma'(s)$.[2] Note that $\mathbf{n}(s)\cdot\gamma'(s) = 0$. Thus $0 = (\mathbf{n}(s)\cdot\gamma'(s))' = \mathbf{n}'(s)\cdot\gamma'(s) + \mathbf{n}(s)\cdot\gamma''(s)$, and $\|\gamma''(s)\| = \|\mathbf{n}(s)\cdot\gamma''(s)\| = \|\mathbf{n}'(s)\cdot\gamma'(s)\| = \|\mathbf{n}'(s)\|$. Therefore, the curvature of γ at point $\gamma(s)$ is the length of the differential of its unit normal vector.

Denote the boundary of $\mathcal{W}$ by $\mathrm{bd}(\mathcal{W})$ and a tangent plane of $\mathrm{bd}(\mathcal{W})$ at point w by $T_w\mathcal{W}$. We shall assume that $\mathcal{W}$ is twice continuously differentiable, that is, $\mathrm{bd}(\mathcal{W})$ is a twice continuously differentiable submanifold of $\mathbb{R}^d$. Then $T_w\mathcal{W}$ is unique, and there exists a unique unit vector at w that is perpendicular to $T_w\mathcal{W}$ and points outward of $\mathcal{W}$. In fact, one can define a continously differentiable normal unit vector field on $\mathrm{bd}(\mathcal{W})$, $u_\mathcal{W} : \mathrm{bd}(\mathcal{W}) \to \mathbb{S}^{d-1}$, the so-called Gauss map, which maps a boundary point $w \in \mathrm{bd}(\mathcal{W})$ to the unique outer normal vector to $\mathcal{W}$ at w, where $\mathbb{S}^{d-1} = \left\{x\in\mathbb{R}^d \mid \|x\|_2 = 1\right\}$ denotes the unit sphere in $\mathbb{R}^d$. Since $u_\mathcal{W}(w)$ maps $\mathrm{bd}(\mathcal{W})$ to unit vectors, the differential of the Gauss map, $\nabla u_\mathcal{W}(w)$, defines a linear endomorphism of $T_w\mathcal{W}$. Moreover, $\nabla u_\mathcal{W}(w)$ is a self-adjoint operator, with nonnegative eigenvalues. The differential of the Gauss map, $\nabla u_\mathcal{W}(w)$, describes the curvature of $\mathrm{bd}(\mathcal{W})$ via the second fundamental form. In particular, the *principal curvatures* of $\mathrm{bd}(\mathcal{W})$ at $w \in \mathrm{bd}(\mathcal{W})$ are defined as the eigenvalues of $\nabla u_\mathcal{W}(w)$. Perhaps a more intuitive, yet equivalent definition, is that the principal curvatures are the eigenvalues of the Hessian of $f = f_w$ in the parameterization $t \mapsto w + t - f_w(t)u_\mathcal{W}(w)$ of $\mathrm{bd}(\mathcal{W})$, which is valid in a small open neighborhood of w, where $f_w : T_w\mathcal{W} \to [0,\infty)$ is a suitable convex, nonnegative valued function that also satisfies $f_w(0) = 0$ (see Fig. 2). Thus, the principal curvatures at some point $w \in \mathrm{bd}(\mathcal{W})$ describe the local shape of $\mathrm{bd}(\mathcal{W})$ up to the second order. In this paper, we are interested in the minimum principal curvature at $w \in \mathrm{bd}(\mathcal{W})$, which can be interpreted as the minimum curvature at w over all the planar curves $\gamma \in \mathrm{bd}(\mathcal{W})$ that go through w.

[2] There exist two unit vectors that are perpendicular to $\gamma'(s)$ for each point on γ. Pick the ones that are consistently oriented.

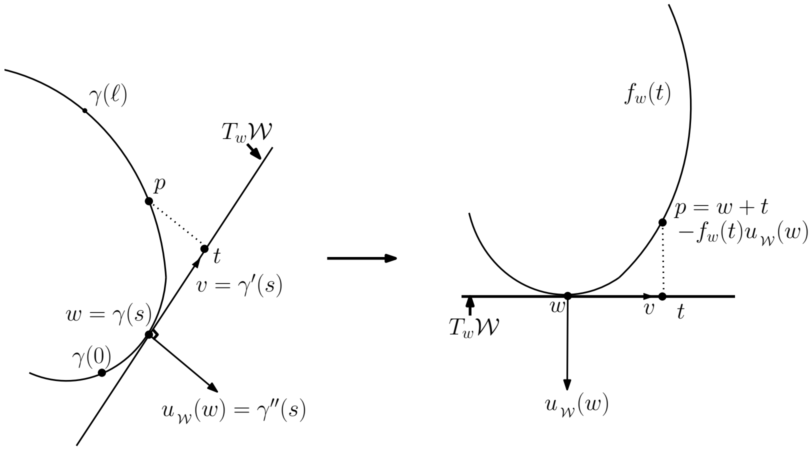

Figure 2: Some differential geometry notations.

A related concept that has been used in convex optimization to show fast rates is that of a strongly convex constraint set (Levitin and Polyak, 1966; Garber and Hazan, 2015): $\mathcal{W}$ is λ-strongly convex with respect to the norm $\|\cdot\|$ if, for any $x, y \in \mathcal{W}$ and $\gamma \in [0,1]$, the $\|\cdot\|$-ball with origin $\gamma x + (1-\gamma)y$ and radius $\gamma(1-\gamma)\lambda \|x-y\|^2/2$ is included in $\mathcal{W}$. That is, for any $z \in \mathbb{R}^d$ with $\|z\| = 1$, $\gamma x + (1-\gamma)y + \gamma(1-\gamma)\frac{\lambda}{2}\|x-y\|^2 z \in \mathcal{W}$. Next we show that a convex body $\mathcal{W}$ with twice continuously differentiable boundary is λ-strongly convex with respect to $\|\cdot\|_2$ if and only if the principal curvatures of the surface $\mathrm{bd}(\mathcal{W})$ are all at least λ.[3] In the rest of the paper, $B_r(x) = \left\{y \in \mathbb{R}^d \mid \|x-y\|_2 \le r\right\}$ will denote the Euclidean ball of radius r centered at x (in case x is the origin, it will often be omitted).

Proposition 4 *Let $\mathcal{W} \subset \mathbb{R}^d$ be a convex body with with twice continuously differentiable boundary and support function φ, and let λ be an arbitrary positive number. Then the following statements are equivalent:*

(i) The smallest principal curvature of $\mathcal{W}$ is at least λ.

(ii) $\mathcal{W} = \cap_{\theta \in \mathbb{S}^{d-1}} B_{1/\lambda}(w_\theta - \theta/\lambda)$ where $w_\theta \in \partial\varphi(\theta) \subset \mathrm{bd}(\mathcal{W})$.

(iii) $\mathcal{W}$ is λ-strongly convex.

Condition (ii), which is actually the definition of Polovinkin (1996) for strongly convex sets, means that $\mathcal{W}$ can be obtained as the intersection of closed balls of radius $1/\lambda$, such that there is one ball for every boundary point w and tangent hyperplane P where the ball touches P at w. Note that a ball with radius $1/\lambda$ satisfies all conditions: (i) and (ii) by definition, while (iii) holds, e.g., by Example 13 of Journée et al. (2010).

3.1.2 Regret Bounds

As promised, our next result connects the principal curvatures of $\mathrm{bd}(\mathcal{W})$ to the regret of FTL and shows that FTL enjoys logarithmic regret for highly curved surfaces, as long as $\|\Theta_t\|_2$ is bounded away from zero.

Theorem 5 *Assume $d \ge 2$ and let $\mathcal{W} \subset \mathbb{R}^d$ be a convex body with twice continuously differentiable boundary. Let $M = \max_{f \in \mathcal{F}} \|f\|_2$ and assume that Φ is differentiable at $(\Theta_t)_t$. Assume that the*

[3] Following Schneider (2014), a convex body in $\mathbb{R}^d$ is any compact, convex subset of $\mathbb{R}^d$ with non-empty interior.

principal curvatures of the surface $\mathrm{bd}(\mathcal{W})$ *are all at least* λ_0 *for some constant* $\lambda_0 > 0$ *(that is,* $\mathcal{W}$ *is* λ_0 *strongly convex) and* $L_n := \min_{1\le t\le n} \|\Theta_t\|_2 > 0$*. Choose* $w_1 \in \mathrm{bd}(\mathcal{W})$*. Then*

$$R_n \le \frac{2M^2}{\lambda_0 L_n}(1+\log n)\,.$$

Before presenting the proof of the theorem, we discuss some of its implications and refinements. After the proof we will provide some examples of constraint sets with positive minimum principal curvature.

Remark 6 As we will show later in an essentially matching lower bound, this bound is tight, showing that the forte of FTL is when L_n is bounded away from zero and λ_0 is large. Note that the bound is vacuous as soon as $L_n = O(\log n/n)$ and is worse than the minimax bound of $O(\sqrt{n})$ when $L_n = o(\log n/\sqrt{n})$. One possibility to reduce the bound's sensitivity to L_n is to use the trivial bound $\langle w_{t+1} - w_t, \Theta_t\rangle \le LW = L\sup_{w,w'\in\mathcal{W}} \|w - w'\|_2$ for indices t when $\|\Theta_t\|_2 \le L$ (with an arbitrary $L > 0$). Then, by optimizing the bound over L, one gets a data-dependent bound of the form

$$R_n \le \inf_{L>0}\left(\frac{2M^2}{\lambda_0 L}(1+\log n) + LW\sum_{t=1}^{n} t\,\mathbb{I}\left(\|\Theta_t\|_2 \le L\right)\right), \tag{4}$$

which is more complex, but is free of L_n and thus reflects the nature of FTL better. Note that in the case of stochastic problems, where $f_1,\dots,f_n$ are independent and identically distributed (i.i.d.) with $\mu := -\mathbb{E}\left[\Theta_t\right] \ne 0$, the probability that $\|\Theta_t\|_2 < \|\mu\|_2/2$ is exponentially small in t. Thus, selecting $L = \|\mu\|_2/2$ in the previous bound, the contribution of the expectation of the second term is $O(\|\mu\|_2 W)$, giving an overall bound of the form $O(\frac{M^2}{\lambda_0\|\mu\|_2}\log n + \|\mu\|_2 W)$. On the other hand, if $\|\Theta_t\|_2 = 1$ if t is odd and $\|\Theta_t\| = 0$ otherwise (such an example is trivial to construct), the optimal choice of L in the above bound is $L = \Theta\left(\frac{M}{n}\sqrt{\frac{\log n}{\lambda_0 W}}\right)$ leads to a vacuous $O\left(Mn\sqrt{\frac{W\log n}{\lambda_0}}\right)$ regret bound.

Remark 7 Now consider the case of i.i.d. losses in a bit more detail. Assume that $\mathcal{W} = \mathcal{F} = B_1$, the Euclidean unit ball centered at the origin, and assume $\mathbb{E}[f_t] = \mu \ne 0$. Then it is straightforward to derive a high probability lower bound for $\|\Theta_t\|$: Using that $\mathbb{E}\|\Theta_t\|_2^2 = \|\mu\|_2^2 + \frac{\sigma^2}{t}$ where $\sigma^2 = \mathbb{E}\|f_i\|_2^2 - \|\mu\|_2^2$, we get

$$\mathbb{P}\left[\|\Theta_t\|_2 \le \frac{\|\mu\|_2}{2}\right] = \mathbb{P}\left[\|\Theta_t\|_2^2 - \mathbb{E}\|\Theta_t\|_2^2 \le -\frac{3\|\mu\|_2^2}{4} - \frac{\sigma^2}{t}\right] \le e^{-\frac{t}{18}\left(\frac{3\|\mu\|_2^2}{4}+\frac{\sigma^2}{t}\right)^2} \le e^{-t\frac{\|\mu\|_2^4}{32}}\,,$$

where the first inequality is due to McDiarmid's inequality (Boucheron et al., 2013) after noticing that changing a single f_i to some $f_i' \in \mathcal{F}$ may change the value of $\|\Theta_t\|_2^2$ by at most $6/t$. Combining this with (4) for $L = \|\mu\|_2/2$, we get

$$\mathbb{E}R_n \le \frac{4}{\|\mu\|_2}(1+\log n) + \frac{\|\mu\|_2}{4\sinh\left(\frac{\|\mu\|_2^4}{32}\right)} \le \frac{4}{\|\mu\|_2}(1+\log n) + O(1/\|\mu\|_2^7)\,. \tag{5}$$

Koolen et al. (2016) also proved an $O(\log n)$ bound on the expected regret of the sophisticated algorithm MetaGrad for the above case (Theorem 3 and Lemma 5 of their paper): in particular, they showed that MetaGrad achieves $O(Bd\log n)$ regret where $B = \frac{2\lambda_{\max}}{\|\mu\|}$ with $\lambda_{\max}$ being the maximum eigenvalue of $\mathbb{E}[f_t f_t^\top]$. Since $\lambda_{\max}$ can be as large as 1 (if $\|f_t\|_2 = 1$), (5) can improve (asymptotically) a factor of d over this regret bound. On the other hand, if f_t is uniformly distributed on the half unit sphere (e.g., the first coordinate of f_t is nonnegative with probability 1), Koolen et al. (2016) shows that $B \le \frac{24}{\sqrt{d}}$, which leads to an $O(\sqrt{d}\log n)$ regret, essentially matching (5), as one can show that $\frac{c}{\sqrt{d}} \le \|\mu\|_2 \le \frac{1}{\sqrt{d}}$ for some constant c.

Proof of Theorem 5 Fix $\theta_1, \theta_2 \in \mathbb{R}^d$ and let $w^{(1)} = \operatorname{argmax}_{w \in \mathcal{W}} \langle w, \theta_1 \rangle$, $w^{(2)} = \operatorname{argmax}_{w \in \mathcal{W}} \langle w, \theta_2 \rangle$. Note that if $\theta_1, \theta_2 \neq 0$ then $w^{(1)}, w^{(2)} \in \operatorname{bd}(\mathcal{W})$. Below we will show that

$$\langle w^{(1)} - w^{(2)}, \theta_1 \rangle \leq \frac{1}{2\lambda_0} \frac{\|\theta_2 - \theta_1\|_2^2}{\|\theta_2\|_2} \,. \tag{6}$$

Proposition 2 shows that bounding the regret is equivalent to bounding $\langle w_{t+1} - w_t, \Theta_t \rangle$. We then apply (6), which shows that the regret can be bounded by controlling the stability of Θ_t. A straightforward calculation shows that Θ_t cannot move much: for any norm $\|\cdot\|$ on $\mathcal{F}$, we have

$$\begin{aligned} \|\Theta_t - \Theta_{t-1}\| &= \left\| \frac{1}{t-1} \sum_{i=1}^{t-1} f_i - \frac{1}{t} \sum_{i=1}^{t} f_i \right\| = \left\| \sum_{i=1}^{t-1} \left(\frac{1}{t-1} - \frac{1}{t} \right) f_i - \frac{1}{t} f_t \right\| \\ &\leq \left\| \sum_{i=1}^{t-1} \left(\frac{1}{t-1} - \frac{1}{t} \right) f_i \right\| + \left\| \frac{1}{t} f_t \right\| = \left\| \sum_{i=1}^{t-1} \frac{1}{t(t-1)} f_i \right\| + \left\| \frac{1}{t} f_t \right\| \\ &= \frac{1}{t} \left\| \frac{1}{t-1} \sum_{i=1}^{t-1} f_i \right\| + \frac{1}{t} \|f_t\| \leq \frac{2}{t} M \,. \end{aligned} \tag{7}$$

where $M = \max_{f \in \mathcal{F}} \|f\|$ is a constant that depends on $\mathcal{F}$ and the norm $\|\cdot\|$.

Combining inequality (6) with Proposition 2 and (7), we get

$$\begin{aligned} R_n = \sum_{t=1}^{n} t \langle w_{t+1} - w_t, \Theta_t \rangle &\leq \sum_{t=1}^{n} \frac{t}{2\lambda_0} \frac{\|\Theta_t - \Theta_{t-1}\|_2^2}{\|\Theta_{t-1}\|_2} \\ &\leq \frac{2M^2}{\lambda_0} \sum_{t=1}^{n} \frac{1}{t \|\Theta_{t-1}\|_2} \leq \frac{2M^2}{\lambda_0 L_n} \sum_{t=1}^{n} \frac{1}{t} \leq \frac{2M^2}{\lambda_0 L_n} (1 + \log n) \,. \end{aligned}$$

To finish the proof we need to show (6). Below we provide a derivation based on the definition of principal curvature. Using the equivalence between the principal curvature and the modulus of strong convexity (cf., Proposition 4), we also provide an alternative proof in Appendix A.3 based on strong convexity, which leads to the slightly weaker result (24), loosing a constant factor of 4.

The following elementary lemma relates the cosine of the angle between two vectors θ_1 and θ_2 to the squared normalized distance between the two vectors, thereby reducing our problem to bounding the cosine of this angle. For brevity, we denote by $\cos(\theta_1, \theta_2)$ the cosine of the angle between θ_1 and θ_2.

Lemma 8 *For any non-zero vectors $\theta_1, \theta_2 \in \mathbb{R}^d$,*

$$1 - \cos(\theta_1, \theta_2) \leq \frac{1}{2} \frac{\|\theta_1 - \theta_2\|_2^2}{\|\theta_1\|_2 \|\theta_2\|_2}. \tag{8}$$

Proof Note that $\|\theta_1\|_2 \|\theta_2\|_2 \cos(\theta_1, \theta_2) = \langle \theta_1, \theta_2 \rangle$. Therefore, (8) is equivalent to $2\|\theta_1\|_2 \|\theta_2\|_2 - 2\langle \theta_1, \theta_2 \rangle \leq \|\theta_1 - \theta_2\|_2^2$, which, by algebraic manipulations, is itself equivalent to $0 \leq (\|\theta_1\|_2 - \|\theta_2\|_2)^2$. ■

Given this result, it suffices to show that $\cos(\theta_1, \theta_2) \leq 1 - \lambda_0 \langle w^{(1)} - w^{(2)}, \frac{\theta_1}{\|\theta_1\|_2} \rangle$, which we prove using the tools from differential geometry introduced in Section 3.1.1. Let $\tilde{\theta}_i = \frac{\theta_i}{\|\theta_i\|_2}$ for $i = 1, 2$. The angle between θ_1 and θ_2 is the same as the angle between the normalized vectors $\tilde{\theta}_1$ and $\tilde{\theta}_2$. To calculate the cosine of the angle between $\tilde{\theta}_1$ and $\tilde{\theta}_2$, let P be a plane spanned by $\tilde{\theta}_1$ and $w^{(1)} - w^{(2)}$ and passing through $w^{(1)}$ (P is uniquely determined if $\tilde{\theta}_1$ is not parallel to $w^{(1)} - w^{(2)}$; if there are multiple planes, just pick any of them). Further, let $\hat{\theta}_2 \in \mathbb{S}^{d-1}$ be the unit vector along the projection of $\tilde{\theta}_2$ onto the plane P, as indicated in Fig. 3. Clearly, $\cos(\tilde{\theta}_1, \tilde{\theta}_2) \leq \cos(\tilde{\theta}_1, \hat{\theta}_2)$.

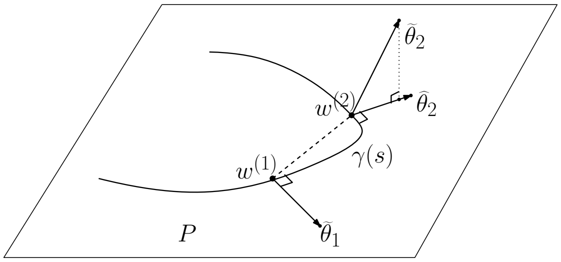

Figure 3: Illustration of the construction used in the proof of (6).

Consider a curve $\gamma(s)$ on $\mathrm{bd}(\mathcal{W})$ connecting $w^{(1)}$ and $w^{(2)}$ that is defined by the intersection of $\mathrm{bd}(\mathcal{W})$ and P and is parametrized by its curve length s so that $\gamma(0) = w^{(1)}$ and $\gamma(l) = w^{(2)}$, where l is the length of the curve γ between $w^{(1)}$ and $w^{(2)}$. Note that since γ is parametrized by its length, $\|\gamma'(s)\|_2 = 1$ for all $s \in [0, l]$. Let $u_{\mathcal{W}}(w)$ denote the outer normal vector to $\mathcal{W}$ at w as before, and let $u_\gamma : [0, l] \to \mathbb{S}^{d-1}$ denote the Gauss map of the planar curve γ, that is, $u_\gamma(s) = \hat{\theta}$ where $\hat{\theta}$ is the unit vector parallel to the projection of $u_{\mathcal{W}}(\gamma(s))$ on the plane P. Now, for any $\theta \in \mathbb{S}^{d-1}$, $w_\theta = \operatorname{argmax}_{w\in\mathcal{W}}\langle w, \theta\rangle$ is a point where a hyperplane with normal vector θ touches $\mathcal{W}$, thus, $u_{\mathcal{W}}(w_\theta) = \theta$. Therefore, $u_\gamma(0) = \tilde{\theta}_1$ and $u_\gamma(l) = \hat{\theta}_2$. In fact γ exists in two versions since $\mathcal{W}$ is a compact convex body, hence the intersection of P and $\mathrm{bd}(\mathcal{W})$ is a closed curve. Of these two versions we choose the one that satisfies that $\langle \gamma'(s), \tilde{\theta}_1\rangle \le 0$ for $s \in [0, l]$.[4] Given the above, we have

$$\cos(\tilde{\theta}_1, \hat{\theta}_2) = \langle \hat{\theta}_2, \tilde{\theta}_1\rangle = 1 + \langle \hat{\theta}_2 - \tilde{\theta}_1, \tilde{\theta}_1\rangle = 1 + \Big\langle \int_0^l u'_\gamma(s)\,\mathrm{d}s, \tilde{\theta}_1 \Big\rangle = 1 + \int_0^l \langle u'_\gamma(s), \tilde{\theta}_1\rangle\,\mathrm{d}s. \quad (9)$$

Note that γ is a planar curve on $\mathrm{bd}(\mathcal{W})$, thus its curvature $\lambda(s)$ satisfies $\lambda(s) \ge \lambda_0$ for any $s \in [0, l]$. Also, for all $s \in [0, l]$, $\gamma'(s)$ is a unit vector parallel to P (since γ is parametrized by its curve length). Moreover, $u'_\gamma(s)$ is parallel to $\gamma'(s)$ since $u_\gamma(s)$ is the Gauss map, and $\lambda(s) = \|u'_\gamma(s)\|_2$. Therefore,

$$\langle u'_\gamma(s), \tilde{\theta}_1\rangle = \|u'_\gamma(s)\|_2\langle \gamma'(s), \tilde{\theta}_1\rangle \le \lambda_0\langle \gamma'(s), \tilde{\theta}_1\rangle,$$

where the last inequality holds because $\langle \gamma'(s), \tilde{\theta}_1\rangle \le 0$. Plugging this into (9), we get the desired

$$\cos(\tilde{\theta}_1, \hat{\theta}_2) \le 1 + \lambda_0 \int_0^l \langle \gamma'(s), \tilde{\theta}_1\rangle\,\mathrm{d}s = 1 + \lambda_0\Big\langle \int_0^l \gamma'(s)\,\mathrm{d}s, \tilde{\theta}_1\Big\rangle = 1 - \lambda_0\langle w^{(1)} - w^{(2)}, \tilde{\theta}_1\rangle\,.$$

Reordering and combining with (8) we obtain

$$\langle w^{(1)} - w^{(2)}, \tilde{\theta}_1\rangle \le \frac{1}{\lambda_0}\Big(1 - \cos(\tilde{\theta}_1, \hat{\theta}_2)\Big) \le \frac{1}{\lambda_0}\left(1 - \cos(\theta_1, \theta_2)\right) \le \frac{1}{2\lambda_0}\frac{\|\theta_1 - \theta_2\|_2^2}{\|\theta_1\|_2\|\theta_2\|_2}\,.$$

Multiplying both sides by $\|\theta_1\|_2$ gives (6), thus, finishing the proof. ■

Next we present the smallest principal curvature of some common convex bodies (the proofs are relegated to the appendix), often used as constraint sets in machine learning.

Example 1 *(i) The smallest principal curvature λ_0 of the Euclidean ball $\mathcal{W} = \{w \mid \|w\|_2 \le r\}$ of radius r satisfies $\lambda_0 = \frac{1}{r}$.*

[4] γ' and u'_γ denote the derivatives of γ and u_γ, respectively, which exist since $\mathrm{bd}(\mathcal{W})$ is twice continuously differentiable. When $s = 0$ or $s = l$, it suffices to take the corresponding one-sided derivatives or, equivalently, extend the definitions of γ and u_γ to an interval $[-\epsilon, l + \epsilon]$ for some $\epsilon > 0$.

(ii) *Let Q be a positive definite matrix. If $\mathcal{W} = \left\{w \mid w^\top Q w \le 1\right\}$ then $\lambda_0 = \lambda_{\min}/\sqrt{\lambda_{\max}}$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimal, respectively, maximal eigenvalues of Q.*

(iii) *Let $p > 1$ and $\mathcal{W} = \{w \,|\|w\|_p \le 1\}$. If $p > 2$, then $\lambda_0 = 0$. Otherwise, if $1 < p \le 2$, then*

$$\lambda_0 = \min_{w \in \mathrm{bd}(\mathcal{W})} \min_{v \in \mathbb{S}^{d-1} : \langle w^{\odot(p-1)}, v\rangle = 0} (p-1) \frac{v^\top \operatorname{diag}\left(|w_1|^{p-2}, \cdots, |w_d|^{p-2}\right) v}{\|w^{\odot(p-1)}\|_2} \ge (p-1) d^{\frac{1}{2} - \frac{1}{p}} ,$$

where $w^{\odot(p-1)} = (|w_1|^{p-1}, \ldots, |w_d|^{p-1})$ and $\operatorname{diag}(a_1, \ldots, a_k)$ denotes a $k \times k$ diagonal matrix with diagonal entries $a_1, \ldots, a_k$.

(iv) *In general, let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a twice continuously differentiable convex function. Then, for $\mathcal{W} = \{w \mid \phi(w) \le 1\}$, $\lambda_0 = \min_{w \in \mathrm{bd}(\mathcal{W})} \min_{v \in \mathbb{S}^{d-1} : \langle \phi'(w), v\rangle = 0} \frac{v^\top \nabla^2 \phi(w) v}{\|\phi'(w)\|_2}$.*

Some of the results above have been derived in the literature based on seemingly different but equivalent assumptions explored in Proposition 4: (i) is a standard result in books on differential geometry; Polovinkin (1996) derived (ii) based on the strong convexity definition (ii) in Proposition 4, while (iii) was proved by Garber and Hazan (2015) based on the strong convexity definition (iii) in Proposition 4. Other examples of strongly sets convex sets, that is, sets with positive minimal principal curvature, can be found in the paper of Garber and Hazan (2015).

Our last result in this section is a lower bound for the linear game, showing that FTL achieves the optimal rate under the condition that $\min_t \|\Theta_t\|_2 \ge L > 0$.

Theorem 9 *Let $\lambda, L \in (0,1)$. Assume that $\{(1, -L), (-1, -L)\} \subset \mathcal{F}$ and let*

$$\mathcal{W} = \left\{(x, y) \in \mathbb{R}^2 : x^2 + \frac{y^2}{\lambda^2} \le 1\right\}$$

be an ellipsoid with principal curvature λ. Then, for any learning strategy, there exists a sequence of losses in $\mathcal{F}$ such that $\|\Theta_t\|_2 \ge L$ for all t and

$$R_n \ge \frac{1}{84\sqrt{2}} \frac{1}{\lambda L} \log n - \frac{1}{\lambda L}\left(\frac{2}{1 - e^{-\lambda^2 L^2}} + \frac{\pi^2}{108}\right) . \tag{10}$$

The theorem states that the regret of any learning strategy can be made at least as large as $\Omega\left(\log n/(L\lambda)\right)$. Note that by Example 1, the minimal principal curvature of $\mathcal{W}$ in the above theorem is λ. In fact, it is not too hard to extend the above argument for any set $\mathcal{W}$ such that there is $w \in \mathrm{bd}(\mathcal{W})$ where the curvature is λ, and the curvature is a continuous function in a neighborhood of w over the boundary $\mathrm{bd}(\mathcal{W})$. The constants in the bound then depend on how fast the curvature changes within this neighborhood. In the case above, for small λL, the n-independent term in (10) is of order $1/(\lambda L)^3$.

Proof We define a random loss sequence, and we will show that no algorithm on this sequence can achieve an $o(\log n/(\lambda_0 L)$ regret. Let P be a random variable with $\mathrm{Beta}(K, K)$ distribution for some $K > 0$, and, given P, assume that $X_t, t \ge 1$ are i.i.d. Bernoulli random variables with parameter P. Let $f_t = X_t(1, -L) + (1 - X_t)(-1, -L) = (2X_t - 1, -L)$. Thus, the second coordinate of f_t is always $-L$, and so $\|\Theta_t\|_2 = \left\|\frac{1}{t}\sum_{i=1}^t f_i\right\|_2 \ge L$. Furthermore, the conditional expectation of the loss vector is $f^p \triangleq \mathbb{E}\left[f_t \middle| P = p\right] = (2p - 1, -L)$.

Note that X_t is a function of f_t for all t; thus the conditional expectation of P, given $f_1, \ldots, f_{t-1}$, can be determined by the well-known formula $\hat{P}_{t-1} = \mathbb{E}\left[P \middle| f_1 \ldots f_{t-1}\right] = \frac{K + \sum_{i=1}^{t-1} X_i}{2K + t - 1}$. Given p, denote the optimizer of f^p by w^p, that is, $w^p = \operatorname{argmin}_{w \in \mathcal{W}} \langle w, f^p\rangle$. Then the Bayesian optimal choice in round t is

$$\begin{aligned}
\operatorname*{argmin}_{w\in\mathcal{W}} \mathbb{E}\left[[\left\langle w, f^P\right\rangle\middle|\, f_1\dots f_{t-1}\right] &= \operatorname*{argmin}_{w\in\mathcal{W}} \left\langle w, \mathbb{E}\left[f^P\middle|\, f_1\dots f_{t-1}\right]\right\rangle \\
&= \operatorname*{argmin}_{w\in\mathcal{W}} \left\langle w, f^{\hat{P}_{t-1}}\right\rangle \\
&= w^{\hat{P}_{t-1}}\,, \qquad (11)
\end{aligned}$$

where the first equality follows by linearity of the inner product, the second since f^p is a linear function of p and the third by the definition of w^p.

Thus, denoting by W_t the prediction of an arbitrary algorithm in round t, the expected regret can be bounded from below as

$$\begin{aligned}
\mathbb{E}\left[R_n\right] &= \mathbb{E}\left[\max_{w\in\mathcal{W}}\sum_{t=1}^n \langle W_t - w, f_t\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\max_{w\in\mathcal{W}}\sum_{t=1}^n \langle W_t - w, f_t\rangle\middle|\, P\right]\right] \\
&\ge \mathbb{E}\left[\mathbb{E}\left[\sum_{t=1}^n \left\langle W_t - w^P, f_t\right\rangle\middle|\, P\right]\right] = \mathbb{E}\left[\sum_{t=1}^n \mathbb{E}\left[\left\langle W_t - w^P, f_t\right\rangle\middle|\, P, f_1,\dots,f_{t-1}\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^n \mathbb{E}\left[\left\langle W_t - w^P, f^P\right\rangle\middle|\, f_1,\dots,f_{t-1}\right]\right] \qquad (12) \\
&\ge \mathbb{E}\left[\sum_{t=1}^n \min_{w\in\mathcal{W}} \mathbb{E}\left[\left\langle w - w^P, f^P\right\rangle\middle|\, f_1,\dots,f_{t-1}\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^n \mathbb{E}\left[\left\langle w^{\hat{P}_{t-1}} - w^P, f^P\right\rangle\middle|\, f_1,\dots,f_{t-1}\right]\right] \qquad (13) \\
&= \sum_{t=1}^n \mathbb{E}\left[\left\langle w^{\hat{P}_{t-1}} - w^P, f^P\right\rangle\right]\,,
\end{aligned}$$

where (12) holds because of the independence of the f_s given P and since W_t is chosen based on $f_1,\dots,f_{t-1}$ (but not on P), and (13) holds by (11).

By Lemma 17, given in Appendix A.4, we have

$$\begin{aligned}
\sum_{t=1}^n \mathbb{E}\left[\left\langle w^{\hat{P}_{t-1}} - w^P, f^P\right\rangle\right] &\ge \frac{\lambda L}{2}\sum_{t=1}^n \mathbb{E}\left[\frac{\left(\frac{2\hat{P}_{t-1}-2P}{\lambda L}\right)^2}{\sqrt{1+\left(\frac{1-2P}{\lambda L}\right)^2}\left(1+\left(\frac{1-2\hat{P}_{t-1}}{\lambda L}\right)^2\right)}\right] \qquad (14) \\
&= \frac{2}{\lambda L}\sum_{t=1}^n \mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{\lambda L}\right)^2}}\mathbb{E}\left[\frac{(\hat{P}_{t-1}-P)^2}{1+\left(\frac{1-2\hat{P}_{t-1}}{\lambda L}\right)^2}\middle|\, P\right]\right] \\
&\ge \frac{2}{\lambda L}\sum_{t=1}^n \mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{\lambda L}\right)^2}}\mathbb{E}\left[\frac{(\hat{P}_{t-1}-P)^2}{1+2\left(\frac{1-2P}{\lambda L}\right)^2+2\left(\frac{2P-2\hat{P}_{t-1}}{\lambda L}\right)^2}\middle|\, P\right]\right]\,, \qquad (15)
\end{aligned}$$

where in the last step we used $(a+b)^2 \le a^2+b^2$. Let $\mathcal{G}_t$ be the event that $|\hat{P}_t - P| \le \frac{K|1-2P|}{2K+t} + \frac{t\lambda L}{2K+t}$; note that $\mathcal{G}_t$ holds with high probability by Lemma 18 in Appendix A.4. Then, lower bounding the first term by 0, (15) can be lower bounded by

$$
\begin{aligned}
&\frac{2}{\lambda L}\sum_{t=1}^{n-1}\mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{\lambda L}\right)^2}}\mathbb{E}\left[\frac{(\hat{P}_t-P)^2}{1+2\left(\frac{1-2P}{\lambda L}\right)^2+2\left(\frac{2P-2\hat{P}_t}{\lambda L}\right)^2}\mathbb{I}(\mathcal{G}_t)\middle|\,P\right]\right]\\
&\geq \frac{2}{\lambda L}\sum_{t=1}^{n-1}\mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{\lambda L}\right)^2}}\frac{\mathbb{E}\left[(\hat{P}_t-P)^2\mathbb{I}(\mathcal{G}_t)\middle|\,P\right]}{\left(1+2\left(\frac{1-2P}{\lambda L}\right)^2+2\left(\frac{2K}{2K+t}\frac{|1-2P|}{\lambda L}+\frac{2t}{2K+t}\right)^2\right)}\right]\\
&\geq \frac{2}{\lambda L}\sum_{t=1}^{n-1}\mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{\lambda L}\right)^2}}\frac{\mathbb{E}\left[(\hat{P}_t-P)^2\mathbb{I}(\mathcal{G}_t)\middle|\,P\right]}{\left(9+4\left(\frac{1-2P}{\lambda L}\right)^2+8\frac{|1-2P|}{\lambda L}\right)}\right].
\end{aligned}
$$

Combining the above, and using $(\hat{P}_t-P)^2\leq 1$ together with the upper bound on the probability of the event $\mathcal{G}_t^c$, the complement of $\mathcal{G}_t$, given in Lemma 18, we get

$$
\begin{aligned}
\mathbb{E}\left[R_n\right] &\geq \frac{2}{\lambda L}\sum_{t=1}^{n-1}\mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{\lambda L}\right)^2}}\frac{\mathbb{E}\left[(\hat{P}_t-P)^2\middle|\,P\right]-\mathbb{P}\left[\mathcal{G}_t^c\right]}{\left(9+4\left(\frac{1-2P}{\lambda L}\right)^2+8\frac{|1-2P|}{\lambda L}\right)}\right]\\
&\geq \frac{2}{\lambda L}\sum_{t=1}^{n-1}\left(\mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{\lambda L}\right)^2}}\frac{\mathbb{E}\left[(\hat{P}_t-P)^2\middle|\,P\right]}{\left(9+4\left(\frac{1-2P}{\lambda L}\right)^2+8\frac{|1-2P|}{\lambda L}\right)}\right]-e^{-(t-1)\lambda^2L^2}\right)\\
&\geq \frac{2}{\lambda L}\left(\sum_{t=1}^{n-1}\mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{\lambda L}\right)^2}}\frac{\mathbb{E}\left[(\hat{P}_t-P)^2\middle|\,P\right]}{\left(9+4\left(\frac{1-2P}{\lambda L}\right)^2+8\frac{|1-2P|}{\lambda L}\right)}\right]-\frac{1}{1-e^{-\lambda^2L^2}}\right). \qquad (16)
\end{aligned}
$$

Now, by Lemma 19, given in Appendix A.4, we have

$$
\mathbb{E}\left[(\hat{P}_t-P)^2\middle|\,P\right]=\frac{K^2(1-2P)^2}{(2K+t)^2}+\frac{tP(1-P)}{(2K+t)^2}\geq P(1-P)\left(\frac{1}{t}-\frac{2}{t(2K+t)}\right).
$$

Combining this with (16) and introducing the constant

$$
C=\mathbb{E}\left[\frac{1}{\sqrt{1+\left(\frac{1-2P}{\lambda L}\right)^2}}\frac{P(1-P)}{\left(9+4\left(\frac{1-2P}{\lambda L}\right)^2+8\frac{|1-2P|}{\lambda L}\right)}\right]
$$

we obtain, for any $K>0$,

$$
\begin{aligned}
\mathbb{E}\left[R_n\right] &\geq \frac{2}{\lambda L}\left[-\frac{1}{1-e^{-\lambda^2L^2}}+\sum_{t=1}^{n-1}C\left(\frac{1}{t}-\frac{2}{t(2K+t)}\right)\right] &(17)\\
&\geq \frac{2C}{\lambda L}\log n-\frac{1}{\lambda L}\left(\frac{2}{1-e^{-\lambda^2L^2}}+\frac{C\pi^2}{3}\right). &(18)
\end{aligned}
$$

where we used $\sum_{t=1}^{n-1}\geq\int_1^n 1/t=\log n$ and $\sum_{t=1}^{n-1}1/(t(2K+t))\leq\sum_{t=1}^{\infty}1/t^2=\pi^2/6$. It remains to calculate a constant lower bound for C that is independent of λ and L. Denote $\frac{|1-2P|}{\lambda L}$ by Y; then $0\leq P(1-P)=\frac{1-Y^2\lambda^2L^2}{4}\leq 1/4$. Define $\widehat{\mathcal{G}}$ to be the event when $|Y|\leq 1$. Since P has $\mathrm{Beta}(K,K)$ distribution, $\mathbb{E}\left[P\right]=\frac{1}{2}$ and $\mathrm{Var}(P)=\frac{1}{8K}$. Therefore, by Chebyshev's inequality,

$$
\mathbb{P}\left[\widehat{\mathcal{G}}^c\right]=\mathbb{P}\left[\left|P-\frac{1}{2}\right|>\frac{\lambda L}{2}\right]\leq\frac{1}{2K\lambda^2L^2}.
$$

Therefore,

$$\begin{aligned}C &= \mathbb{E}\left[\frac{1}{\sqrt{1+Y^2}}\frac{1-Y^2\lambda^2L^2}{4(9+4Y^2+8Y)}\right] \geq \mathbb{E}\left[\frac{1}{\sqrt{1+Y^2}}\frac{1-Y^2\lambda^2L^2}{4(9+4Y^2+8Y)}\mathbb{I}(\widehat{\mathcal{G}})\right]\\ &\geq \frac{1}{84\sqrt{2}}\mathbb{E}\left[(1-Y^2\lambda^2L^2)\mathbb{I}(\widehat{\mathcal{G}})\right] \geq \frac{1}{84\sqrt{2}}\left(\mathbb{E}\left[1-Y^2\lambda^2L^2\right] - \mathbb{P}\left[\widehat{\mathcal{G}}^c\right]\right)\\ &\geq \frac{1}{84\sqrt{2}}\left(1-\mathbb{E}\left[(1-2P)^2\right]-\frac{1}{2K\lambda^2L^2}\right) = \frac{1}{84\sqrt{2}}\left(1-\frac{1}{2K}-\frac{1}{2K\lambda^2L^2}\right)\\ &\geq \frac{1}{84\sqrt{2}}\cdot\frac{1}{2}\end{aligned}$$

for any $K \geq 1 + \frac{1}{\lambda^2L^2}$. Hence,

$$\mathbb{E}\left[R_n\right] \geq \frac{1}{84\sqrt{2}}\frac{1}{\lambda L}\log n - \frac{1}{\lambda L}\left(\frac{2}{1-e^{-\lambda^2L^2}} + \frac{\pi^2}{108}\right),$$

where we used the trivial upper bound $C \leq 1/36$ (obtained by maximizing the argument of the expectation in P in the definition of C by selecting $P = 1/2$). The result is completed by noting that the worst-case regret is at least as big as the expected regret, thus, for every n, there exist a P and a sequence of loss vectors $f_1, \ldots, f_n$ such that the regret R_n satisfies (10). ■

3.2 Other Regularities

So far we have looked at the case when FTL achieves a low regret due to the curvature of $\mathrm{bd}(\mathcal{W})$. The next result characterizes the regret of FTL when $\mathcal{W}$ is a polytope, which has a flat, non-smooth boundary and thus Theorem 5 is not applicable. For this statement recall that given some norm $\|\cdot\|$, its dual norm is defined by $\|w\|_* = \sup_{\|v\|\leq 1}\langle v, w\rangle$.

Theorem 10 *Assume that $\mathcal{W}$ is a polytope and that Φ is differentiable at Θ_i, $i = 1, \ldots, n$. Let $w_t = \mathrm{argmax}_{w\in\mathcal{W}}\langle w, \Theta_{t-1}\rangle$, $W = \sup_{w_1,w_2\in\mathcal{W}}\|w_1 - w_2\|_*$ and $F = \sup_{f_1,f_2\in\mathcal{F}}\|f_1 - f_2\|$. Then the regret of FTL is*

$$R_n \leq W\sum_{t=1}^{n} t\,\mathbb{I}(w_{t+1}\neq w_t)\|\Theta_t - \Theta_{t-1}\| \leq FW\sum_{t=1}^{n}\mathbb{I}(w_{t+1}\neq w_t).$$

Note that when $\mathcal{W}$ is a polytope, w_t is expected to "snap" to some vertex of $\mathcal{W}$. Hence, we expect the regret bound to be non-vacuous, if, e.g., Θ_t "stabilizes" around some value. Some examples after the proof will illustrate this.

Proof Let $v = \mathrm{argmax}_{w\in\mathcal{W}}\langle w, \theta\rangle$, $v' = \mathrm{argmax}_{w\in\mathcal{W}}\langle w, \theta'\rangle$. Similarly to the proof of Theorem 5,

$$\begin{aligned}\langle v'-v, \theta'\rangle &= \langle v', \theta'\rangle - \langle v', \theta\rangle + \langle v', \theta\rangle - \langle v, \theta\rangle + \langle v, \theta\rangle - \langle v, \theta'\rangle\\ &\leq \langle v', \theta'\rangle - \langle v', \theta\rangle + \langle v, \theta\rangle - \langle v, \theta'\rangle = \langle v'-v, \theta'-\theta\rangle \leq W\,\mathbb{I}(v'\neq v)\|\theta'-\theta\|,\end{aligned}$$

where the first inequality holds because $\langle v', \theta\rangle \leq \langle v, \theta\rangle$. Therefore, by (7),

$$R_n = \sum_{t=1}^{n} t\,\langle w_{t+1}-w_t, \Theta_t\rangle \leq W\sum_{t=1}^{n} t\,\mathbb{I}(w_{t+1}\neq w_t)\|\Theta_t - \Theta_{t-1}\| \leq FW\sum_{t=1}^{n}\mathbb{I}(w_{t+1}\neq w_t).$$

■

As noted before, since $\mathcal{W}$ is a polytope, w_t is (generally) attained at the vertices. In this case, the epigraph of Φ is a polyhedral cone. Then, the event when $w_{t+1} \neq w_t$, that is, when the "leader" switches corresponds to when Θ_t and Θ_{t-1} belong to different linear regions corresponding to different linear pieces of the graph of Φ.

We now spell out a corollary for the stochastic setting. In particular, in this case FTL will often enjoy a constant regret:

Corollary 11 (Stochastic setting) *Assume that $\mathcal{W}$ is a polytope and that $(f_t)_{1\le t\le n}$ is an i.i.d. sequence of random variables such that $\mathbb{E}[f_i] = \mu$ and $\|f_i\|_\infty \le M$. Let $W = \sup_{w_1,w_2\in\mathcal{W}} \|w_1 - w_2\|_1$. Further assume that there exists a constant $r > 0$ such that Φ is differentiable for any ν such that $\|\nu - \mu\|_\infty \le r$. Then,*

$$\mathbb{E}[R_n] \le 2MW\left(3 + \frac{2M^2}{r^2}\log\left(\frac{2M^2 d}{r^2}\right)\right).$$

The existence of an r such that Φ is differentiable for any ν such that $\|\nu - \mu\|_\infty \le r$ is equivalent to that Φ is differentiable at μ. By Proposition 1, this condition requires that at μ, $\max_{w\in\mathcal{W}}\langle w, \theta\rangle$ has a unique optimizer (note that the volume of the set of vectors θ with multiple optimizers is zero). On the other hand, r should be selected to be the radius of the largest ball such that the optimal decisions for the expected losses μ and ν (i.e., the maximizers defining $\Phi(-\mu)$ and $\Phi(-\nu)$) belong to the same face of $\mathcal{W}$.

Proof Let $V = \{\nu \,|\, \|\nu - \mu\|_\infty \le r\}$. Note that the epigraph of the function Φ is a polyhedral cone. Since Φ is differentiable in the interior of V, $\{(\theta, \Phi(\theta)) \,|\, \theta \in V\}$ is a subset of a linear subspace. Therefore, for $-\Theta_t, -\Theta_{t-1} \in V$, $w_{t+1} = w_t$. Hence, by Theorem 10,

$$\mathbb{E}[R_n] \le 2MW \sum_{t=1}^{n} \mathbb{P}[-\Theta_t, -\Theta_{t-1} \notin V] \le 4MW\left(1 + \sum_{t=1}^{n}\mathbb{P}[-\Theta_t \notin V]\right). \qquad (19)$$

On the other hand, note that $\|f_i\|_\infty \le M$. Then

$$\mathbb{P}[-\Theta_t \notin V] = \mathbb{P}\left[\left\|\frac{1}{t}\sum_{i=1}^{t} f_i - \mu\right\|_\infty \ge r\right] \le \sum_{j=1}^{d}\mathbb{P}\left[\left|\frac{1}{t}\sum_{i=1}^{t} f_{i,j} - \mu_j\right| \ge r\right] \le 2de^{-\frac{tr^2}{2M^2}},$$

where the last inequality is due to Hoeffding's inequality. Now, using that for any $\alpha > 0$ and $\tau > 0$, $\sum_{t=\tau+1}^{n}\exp(-\alpha t) \le \int_\tau^n \exp(-\alpha t)dt \le \frac{1}{\alpha}\exp(-\alpha\tau)$, from (19) we obtain

$$\mathbb{E}[R_n] \le 2MW\left(1 + \tau + \frac{2d}{\alpha}e^{-\alpha\tau}\right).$$

Setting $\alpha = \frac{r^2}{2M^2}$ and $\tau = \frac{1}{\alpha}\log(d/\alpha)$ in the above bound finishes the proof. ■

4. Adaptive Algorithms

While FTL can exploit the curvature of the surface of the constraint set to achieve $O(\log n)$ regret, it requires the curvature condition and $\min_t \|\Theta_t\|_2 \ge L$ being bounded away from zero, or it may suffer linear regret. On the other hand, many algorithms such as the follow the regularized leader (FTRL) are known to achieve a regret guarantee of $O(\sqrt{n})$ even for the worst-case data in the linear setting (see, e.g., Shalev-Shwartz, 2012). This raises the question of whether one can have an algorithm that can achieve constant or $O(\log n)$ regret in the respective settings of Corollary 11 or Theorem 5, while it still maintains $O(\sqrt{n})$ regret for worst-case data. One way to design an adaptive algorithm is to use the $(\mathcal{A}, \mathcal{B})$-prod algorithm of Sani et al. (2014), trivially leading to the following result:

Algorithm 1 Follow The Shrunken Leader (FTSL)

```
Predict $w_1 = 0$;
**for** $t = 2, ..., n-1$ **do**
    FTL: Compute $\tilde{w}_t = \operatorname{argmin}_{w \in \mathcal{W}} \langle w, F_{t-1} \rangle$.
    Shrinkage: Predict $w_t = \frac{\|F_{t-1}\|_2}{\sqrt{\|F_{t-1}\|_2^2 + t + 2}} \tilde{w}_t$.
**end for**
FTL: Compute $\tilde{w}_n = \operatorname{argmin}_{w \in \mathcal{W}} \langle w, F_{n-1} \rangle$.
Shrinkage: Predict $w_n = \frac{\|F_{n-1}\|_2}{\sqrt{\|F_{n-1}\|_2^2 + n}} \tilde{w}_n$.
```

Proposition 12 *Consider $(\mathcal{A}, \mathcal{B})$-prod of Sani et al. (2014), where algorithm $\mathcal{A}$ is chosen to be FTRL with an appropriate regularization term, while $\mathcal{B}$ is chosen to be FTL. Then the regret of the resulting hybrid algorithm $\mathcal{H}$ enjoys the following guarantees:*

- *If FTL achieves constant regret as in the setting of Corollary 11, then the regret of $\mathcal{H}$ is also constant.*
- *If FTL achieves a regret of $O(\log n)$ as in the setting of Theorem 5, then the regret of $\mathcal{H}$ is also $O(\log n)$.*
- *Otherwise, the regret of $\mathcal{H}$ is at most $O(\sqrt{n \log n})$.*

In the next section we show that if the constraint set is an ellipsoid, it is possible to design adaptive algorithms directly.

4.1 Adaptive Algorithms for Ellipsoid Constraint Sets

In this section we provide some interesting results about adaptive algorithms for the case when $\mathcal{W}$ is an ellipsoid in $\mathbb{R}^d$. First, we show that a variant of FTL using shrinkage as regularization has $O(\log n)$ regret when $\|\Theta_t\|_2 \geq L > 0$ for all t, but it also has $O(\sqrt{n})$ worst case guarantee. Furthermore, we show that the standard FTRL algorithm is adaptive if the constraint set is an ellipsoid and the loss vectors are stochastic. Throughout the section we will use the notation $F_t = -t\,\Theta_t = \sum_{i=1}^{t} f_i$.

4.1.1 Follow the Shrunken Leader

In this section we are going to analyze a combination of the FTL algorithm and the idea of shrinkage often used for regularization purposes in statistics. We assume that $\mathcal{W}$ is the d-dimensional unit ball and, without loss of generality, we further assume that $\|f\|_2 \leq 1$ for all $f \in \mathcal{F}$. The Follow The Shrunken Leader (FTSL) algorithm is given in Algorithm 1. The main idea of the algorithm is to predict a shrunken version of the FTL prediction, in this way keeping it away from the boundary of $\mathcal{W}$. The next theorem shows that the right amount of shrinkage leads to a robust, adaptive algorithm.

Theorem 13 *Assume that $\mathcal{W} = \left\{ x \in \mathbb{R}^d \mid \|x\|_2 \leq 1 \right\}$ and $\|f\|_2 \leq 1$ for all $f \in \mathcal{F}$. Then the regret of FTSL is $O(\sqrt{n})$. If, in addition, there exists an $L > 0$ such that $\|\Theta_t\|_2 \geq L$ for $1 \leq t \leq n$, then the regret of is $O(\log n / L)$.*

Proof By the definition of F_t and $\mathcal{W}$, $\tilde{w}_t = -F_{t-1}/\|F_{t-1}\|_2$ for $t \geq 2$. Let $\sigma_n = \frac{\|F_{n-1}\|_2}{\sqrt{\|F_{n-1}\|_2^2 + n}}$. Our proof follows the idea of Abernethy et al. (2008). We compute the upper bound on the value of the game for each round backwards for $t = n, n-1, \ldots, 1$, by solving the optimal strategies for f_t. The value of the game using FTSL is defined as

$$\begin{aligned} V_n &= \max_{f_1,\dots,f_n} \sum_{t=1}^{n} \langle w_t, f_t\rangle - \min_{w\in\mathcal{W}} \langle w, F_n\rangle \\ &= \max_{f_1,\dots,f_{n-1}} \sum_{t=1}^{n-1} \langle w_t, f_t\rangle + \underbrace{\max_{f_n} \|F_{n-1}+f_n\|_2 + \langle f_n, w_n\rangle}_{=:U_n} \end{aligned}$$

We first prove that U_n, the second term above, is bounded from above by $\sqrt{\|F_{n-1}\|_2^2+n}$. To see this, let $f_n = a_n\tilde{F}_{n-1} + b_n\Omega_{n-1}$ where $\tilde{F}_{n-1}$ is the unit vector parallel to F_{n-1} and Ω_{n-1} is a unit vector orthogonal to F_{n-1}. Furthermore, since $\|f_n\|_2 \le 1$, we have $a_n^2+b_n^2 \le 1$. Thus,

$$\begin{aligned} U_n &= \max_{f_n} \sqrt{\|F_{n-1}\|_2^2 + 2a_n\|F_{n-1}\|_2 + a_n^2 + b_n^2} - a_n\sigma_n \\ &\le \max_a \sqrt{\|F_{n-1}\|_2^2 + 2a\|F_{n-1}\|_2 + n} - a\sigma_n \\ &= \sqrt{\|F_{n-1}\|_2^2 + n}, \end{aligned}$$

where the last equality follows since the maximum is attained at $a=0$. A similar statement holds for the other time indices: for any $t \ge 1$,

$$\max_{f_t} \sqrt{\|F_{t-1}+f_t\|_2^2 + t + 1} + \langle f_t, w_t\rangle \le \sqrt{\|F_{t-1}\|_2^2 + t} + \frac{1}{\sqrt{t}}\ . \tag{20}$$

Before proving this inequality, we show how it implies the first statement of the theorem:

$$\begin{aligned} V_n &\le \max_{f_1,\dots,f_{n-1}} \sum_{t=1}^{n-1} \langle w_t, f_t\rangle + \sqrt{\|F_{n-1}\|_2^2 + n} \\ &\le \max_{f_1,\dots,f_{n-2}} \sum_{t=1}^{n-2} \langle w_t, f_t\rangle + \sqrt{\|F_{n-2}\|_2^2 + n - 1} + \frac{1}{\sqrt{n-1}} \\ &\le \dots \\ &\le 1 + \sum_{t=1}^{n-1} \frac{1}{\sqrt{t}} \le 2(1+\sqrt{n-1}). \end{aligned}$$

Moreover, if $\|\Theta_t\|_2 \ge L > 0$ for $1 \le t \le n$, a stronger version of (20) also holds:

$$\max_{f_t} \sqrt{\|F_{t-1}+f_t\|_2^2 + t + 1} + \langle f_t, w_t\rangle \le \sqrt{\|F_{t-1}\|_2^2 + t} + \frac{1}{(t-1)L}. \tag{21}$$

This implies the second statement of the theorem, since

$$\begin{aligned} V_n &\le \max_{f_1,\dots,f_{n-1}} \sum_{t=1}^{n-1} \langle w_t, f_t\rangle + \sqrt{\|F_{n-1}\|_2^2 + n} \\ &\le \max_{f_1,\dots,f_{n-2}} \sum_{t=1}^{n-2} \langle w_t, f_t\rangle + \sqrt{\|F_{n-2}\|_2^2 + n - 1} + \frac{1}{(n-1)L} \\ &\le \dots \\ &\le 1 + \sum_{t=1}^{n-1} \frac{1}{tL} \le 1 + \frac{1}{L} + \frac{\log(n-1)}{L}\ . \end{aligned}$$

To finish the proof, it remains to show (20) and (21). Let $f_t = a_t\tilde{F}_{t-1} + b_t\Omega_{t-1}$ where $\tilde{F}_{t-1}$ is the unit vector parallel to F_{t-1} and Ω_{t-1} is a unit vector orthogonal to F_{t-1}. Since $\|f_t\|_2 \le 1$, observe that $a_t^2 + b_t^2 = \|f_t\|_2 \le 1$. Furthermore, let $\sigma_t = \frac{\|F_{t-1}\|_2}{\sqrt{\|F_{t-1}\|_2^2+t+2}}$. Then, for any $t \ge 1$,

$$\begin{aligned}
\Delta_t &= \max_{f_t} \sqrt{\|F_{t-1}\|_2^2 + 2a_t\|F_{t-1}\|_2 + a_t^2 + b_t^2 + t + 1} - a_t\sigma_t - \sqrt{\|F_{t-1}\|_2^2 + t} \\
&\le \max_{a_t} \sqrt{\|F_{t-1}\|_2^2 + 2a_t\|F_{t-1}\|_2 + t + 2} - a_t\sigma_t - \sqrt{\|F_{t-1}\|_2^2 + t} \\
&= \sqrt{\|F_{t-1}\|_2^2 + t + 2} - \sqrt{\|F_{t-1}\|_2^2 + t} \\
&= \frac{2}{\sqrt{\|F_{t-1}\|_2^2 + t + 2} + \sqrt{\|F_{t-1}\|_2^2 + t}} \qquad (22) \\
&\le \frac{1}{\sqrt{t}}.
\end{aligned}$$

This proves (20). Moreover, if $\|F_{t-1}\|_2 = \|(t-1)\Theta\|_2 \ge (t-1)L > 0$, by (22) we obtain

$$\Delta_t \le \frac{2}{\sqrt{\|F_{t-1}\|_2^2 + t + 2} + \sqrt{\|F_{t-1}\|_2^2 + t}} \le \frac{1}{\|F_{t-1}\|_2} \le \frac{1}{(t-1)L},$$

proving (21). ■

Remark 14 The above result can easily be extended to the case when $\mathcal{W}$ is an ellipsoid, that is, $\mathcal{W} = \left\{w \,|\, w^\top Q w \le 1\right\}$ for some positive-definite matrix Q. Transforming the predictions as $\hat{w}_t = Q^{1/2}w_t$ and the losses $\hat{f}_t = Q^{-1/2}f_t$, we see that, for all t, the new prediction $\hat{w}_t$ belongs to the unit ball (i.e., $\hat{w}_t \in B_1$) and $\langle w_t, f_t\rangle = \langle \hat{w}_t, \hat{f}_t\rangle$. Thus, the value of the game can be bounded as

$$\begin{aligned}
V_n &= \max_{f_1,\ldots,f_n \,:\, \|f_i\|_2 \le 1} \sum_{t=1}^n \langle w_t, f_t\rangle - \min_{w\in\mathcal{W}} \langle w, F_n\rangle \\
&= \max_{f_1,\ldots,f_n \,:\, \|f_i\|_2 \le 1} \sum_{t=1}^n \langle \hat{w}_t, Q^{-1/2}f_t\rangle - \min_{\hat{w}\in B_1} \langle \hat{w}, Q^{-1/2}F_n\rangle \\
&= \max_{\hat{f}_1,\ldots,\hat{f}_n \,:\, \|Q^{1/2}\hat{f}_i\|_2 \le 1} \sum_{t=1}^n \langle \hat{w}_t, \hat{f}_t\rangle - \min_{\hat{w}\in B_1} \langle \hat{w}, \widehat{F}_n\rangle \\
&\le \max_{\hat{f}_1,\ldots,\hat{f}_n \,:\, \|\hat{f}_i\|_2 \le \frac{1}{\sqrt{\lambda_{\min}}}} \sum_{t=1}^n \langle \hat{w}_t, \hat{f}_t\rangle - \min_{\hat{w}\in B_1} \langle \hat{w}, \widehat{F}_n\rangle \\
&= \frac{1}{\sqrt{\lambda_{\min}}} \max_{\tilde{f}_1,\ldots,\tilde{f}_n \,:\, \|\tilde{f}_i\|_2 \le 1} \sum_{t=1}^n \langle \hat{w}_t, \tilde{f}_t\rangle - \min_{\hat{w}\in B_1} \langle \hat{w}, \tilde{F}_n\rangle\,,
\end{aligned}$$

where $\lambda_{\min}$ is the minimal eigenvalue of the matrix Q, $\widehat{F}_n = \sum_{t=1}^n \hat{f}_t$, $\tilde{f}_t = \sqrt{\lambda_{\min}}\,\hat{f}_t$, and $\tilde{F}_n = \sqrt{\lambda_{\min}}\,\widehat{F}_n = \sum_{t=1}^n \tilde{f}_t$. Thus, playing over $\mathcal{W}$ with loss vectors from the unit ball is equivalent to playing over the unit ball B_1 against losses over $\mathcal{W}$ (note that $\|Q^{1/2}\hat{f}_i\|_2 \le 1$ is equivalent to $\hat{f}_i \in \mathcal{W}$), which can be reduced to playing against losses with maximum Euclidean norm 1. Thus, an algorithm for the ellipsoid constraint set $\mathcal{W}$ is to run FTSL over the unit ball B_1 with the transformed losses $\tilde{f}_t$, and predict $w_t = Q^{-1/2}\hat{w}_t$ where $\hat{w}_t$ is the prediction of FTSL. Assuming that $\|\Theta_t\|_2 \ge L$ for all t in the original problem, in the transformed problem we have $\|\sqrt{\lambda_{\min}}\,Q^{-1/2}\Theta_t\|_2 \ge L\sqrt{\lambda_{\min}/\lambda_{\max}}$ where $\lambda_{\max}$ is the largest eigenvalue of Q. Hence, the regret of the algorithm (in both the original and the transformed problems) is at most $O\big(\frac{\sqrt{\lambda_{\max}}}{L\lambda_{\min}}\log n\big)$. Note that this is exactly the same rate as we can obtain from Theorem 5 and Example 1 (ii) for the (non-adaptive) FTL algorithm for the ellipsoid constraint set $\mathcal{W}$ (a closer inspection of the constants shows that the leading constant for FTSL is actually a factor of 2 better).

4.1.2 FTRL for Stochastic Losses

This section shows that when $\mathcal{W}$ is the unit ball B_1, FTRL with regularizer $R(w) = \frac{1}{2}\|w\|^2$ is an adaptive algorithm achieving logarithmic regret for stochastic losses. To fix the notation, in round t, FTLR predicts

$$w_t = \operatorname*{argmin}_{w\in\mathcal{W}} \eta_t\langle F_{t-1}, w\rangle + R(w),$$

if $t > 1$ and $w_1 = 0$. It has been well known that FTRL with $\eta_t = 1/\sqrt{t-1}$ is guaranteed to achieve $O(\sqrt{n})$ regret in the adversarial setting (see, e.g., Shalev-Shwartz, 2012). It remains to prove that FTRL indeed achieves a fast rate in the stochastic setting.

Theorem 15 *Assume that the sequence of loss vectors, $f_1, \ldots, f_n \in \mathbb{R}^d$ satisfies $\|f_t\|_2 \le 1$ almost surely and $\mathbb{E}[f_t] = \mu$ for all t with some $\|\mu\|_2 > 0$. Then FTRL with $\eta_t = 1/\sqrt{t-1}$ suffers $O(\log n)$ regret .*

Proof Using $R(w) = \frac{1}{2}\|w\|^2$ as its regularization, in round $t > 1$ FTRL predicts

$$w_t = \operatorname*{argmin}_{w\in\mathcal{W}} \eta_t\langle F_{t-1}, w\rangle + R(w) = \begin{cases} \frac{1}{\sqrt{t-1}}F_{t-1} & \text{if } \|F_{t-1}\| \le \sqrt{t-1}\,; \\ \frac{F_{t-1}}{\|F_{t-1}\|} & \text{otherwise.} \end{cases} \tag{23}$$

For any $1 \le t \le n$, denote the event $\|F_t\| \ge \sqrt{t}$ by $\mathcal{E}_t$. Note that if $\|F_{t-1}\| \ge \sqrt{t-1}$, FTRL predicts exactly the same w_t as FTL. Denote the accumulated loss of FTL in n rounds by $\mathcal{L}_n^{FTL}$. Thus, the regret of FTRL is

$$\begin{aligned}
\mathbb{E}[R_n] &= \mathbb{E}\left[\sum_{t=1}^n \langle f_t, w_t\rangle - \min_{w\in\mathcal{W}}\langle f_t, w\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^n \langle f_t, w_t\rangle - \mathcal{L}_n^{FTL}\right] + \mathbb{E}\left[\mathcal{L}_n^{FTL} - \min_{w\in\mathcal{W}}\langle f_t, w\rangle\right] \\
&\le 2\sum_{t=1}^n \mathbb{P}[\mathcal{E}_t^c] + O(\log n),
\end{aligned}$$

where to obtain the last inequality we applied (23) for the first term, while the second term is $O(\log n)$ by Remark 7. It remains to bound the first term, $2\sum_{t=1}^n \mathbb{P}[\mathcal{E}_t^c]$ in the above. For any $t > \frac{4}{\|\mu\|_2^2}$,

$$\begin{aligned}
\mathbb{P}\left[\|F_t\|_2 \le \sqrt{t}\right] &\le \mathbb{P}\left[\|F_t\|_2 < \frac{t}{2}\|\mu\|_2\right] \le \sum_{i=1}^d \mathbb{P}\left[|F_{t,i}| < \frac{t}{2}|\mu_i|\right] \\
&\le \sum_{i=1}^d \mathbb{P}\left[|F_{t,i} - t\mu_i| > \frac{t}{2}|\mu_i|\right] \le 2\sum_{i=1}^d e^{-\frac{\mu_i^2}{4}t}
\end{aligned}$$

Thus,

$$\begin{aligned}
\sum_{t=1}^{n} \mathbb{P}\left[\mathcal{E}_t^c\right] &= \sum_{t=1}^{4/\|\mu\|_2^2} \mathbb{P}\left[\mathcal{E}_t^c\right] + \sum_{t=4/\|\mu\|_2^2}^{n} \mathbb{P}\left[\mathcal{E}_t^c\right] \\
&\leq \frac{4}{\|\mu\|_2^2} + 2\sum_{i=1}^{d}\sum_{t=0}^{n} e^{-\frac{\mu_i^2}{4}t} \\
&\leq \frac{4}{\|\mu\|_2^2} + 2\sum_{i=1}^{d} \frac{1}{1-e^{-\frac{\mu_i^2}{4}}} \\
&\leq \frac{4}{\|\mu\|_2^2} + 2\sum_{i=1}^{d} \frac{\mu_i^2}{4} = \frac{4}{\|\mu\|_2^2} + \frac{\|\mu\|_2^2}{2}\ .
\end{aligned}$$

where in the last inequality we used $1/(1-e^{-a}) \leq a$. Therefore, if $\|\mu\| > 0$, the regret of FTRL satisfies

$$\mathbb{E}\left[R_n\right] \leq \frac{8}{\|\mu\|_2^2} + \|\mu\|_2^2 + O(\log n) = O(\log n).$$

■

Remark 16 Similarly to FTSL, the above result can be extended to ellipsoid constraint sets with an adequate choice of the regularizer $R(w)$. Assume that $\mathcal{W} = \left\{w \mid w^\top Q w \leq 1\right\}$ for some positive definite matrix Q, and let $R(w) = \frac{1}{2}w^\top Q w$. Then

$$w_t = \operatorname*{argmin}_{w\in\mathcal{W}} \eta_t\langle F_{t-1}, w\rangle + R(w) = Q^{-1/2} \operatorname*{argmin}_{\tilde{w}\in B_1} \eta_t\langle Q^{-1/2}F_{t-1}, \tilde{w}\rangle + \frac{1}{2}\|\tilde{w}\|_2^2,$$

and

$$R_n = \sum_{t=1}^{n}\langle f_t, w_t\rangle - \min_{w\in\mathcal{W}}\langle f_t, w\rangle = \sum_{t=1}^{n}\langle Q^{-1/2}f_t, \tilde{w}_t\rangle - \min_{w\in B_1}\langle Q^{-1/2}f_t, \tilde{w}\rangle.$$

Thus, the problem is equivalent to the case of a unit ball constraint set with the loss vector $Q^{-1/2}f_t$ for time t, and FTRL with the selected regularizer achieves $O(\sqrt{n})$ worst case regret and $O(\log n)$ regret in the case of an i.i.d. loss sequence. Whether FTRL with a constraint-set-independent regularizer $R(w) = \frac{1}{2}\|w\|_2^2$ achieves similar adaptivity, remains an open question.

5. Simulations

We performed three simulations to illustrate the differences between FTL, FTRL with the regularizer $R(w) = \frac{1}{2}\left\|w\right\|_2^2$ when $w_t = \operatorname{argmin}_{w\in\mathcal{W}} \sum_{i=1}^{t-1}\langle f_{i-1}, w\rangle + R(w)$, and the adaptive algorithm $(\mathcal{A},\mathcal{B})$-prod (AB) using FTL and FTRL as its candidates, which we shall call AB(FTL,FTRL).

For the experiments the constraint set $\mathcal{W}$ was chosen to be a slightly elongated ellipsoid in the 4-dimensional Euclidean space, with volume matching that of the 4-dimensional unit ball. The actual ellipsoid is given by $\mathcal{W} = \left\{w \in \mathbb{R}^4 \mid w^\top Q w \leq 1\right\}$ where Q is randomly generated as

$$Q = \begin{pmatrix} 4.3367 & 3.6346 & -2.2250 & 3.5628 \\ 3.6346 & 3.9966 & -2.3613 & 3.2817 \\ -2.2250 & -2.3613 & 2.0589 & -2.1295 \\ 3.5628 & 3.2817 & -2.1295 & 3.4206 \end{pmatrix}.$$

We experimented with three types of data to illustrate the behavior of the different algorithms: stochastic, “half-adversarial”, and “worst-case” data (worst-case for FTL), as will be explained below. The first two data sets are random, so the experiments were repeated 100 times, and we report the average regret with its standard deviation; the worst case data is deterministic, so there no repetition was needed. For each experiment, we set $n = 2500$. The regularization coefficient for the FTRL, and the learning rate for AB were chosen based on their theoretical bounds minimizing the worst-case regret.

5.1 Stochastic Data

In this setting we used the following model to generate f_t: Let $(\hat{f}_t)_t$ be an i.i.d. sequence drawn from the 4-dimensional standard normal distribution, and let $\tilde{f}_t = \hat{f}_t / \left\| \hat{f}_t \right\|_2$. Then, f_t is defined as $f_t = \tilde{f}_t + L e_1$ where $e_1 = (1, 0, \dots, 0)^\top$. Therefore, $\mathbb{E}\left[\left\| \frac{1}{t} \sum_{s=1}^{t} f_s \right\|_2 \right] \to L$ as $t \to \infty$. In the experiments we picked $L \in \{0, 0.1\}$.

The results are shown in Fig. 4. On the left-hand side we plotted the regret against the logarithm of the number of rounds, while on the right-hand side we plotted the regret against the square root of the number of rounds, together with the standard deviation of the results over the 100 independent runs. As can be seen from the figures, when $L = 0.1$, the growth-rate of the regret of FTL is indeed logarithmic, while when $L = 0$, the growth-rate is $\Theta(\sqrt{n})$. In particular, when $L = 0.1$, FTL enjoys a major advantage compared to FTRL, while for $L = 0$, FTL and FTRL perform essentially the same (in this special case, the regret of FTL will indeed be $O(\sqrt{n})$ as w_t will stay bounded, but $\|\Theta_t\| = O(1/\sqrt{t})$). As expected, AB(FTL,FTRL), gets the better of the two regrets with little to no extra penalty.

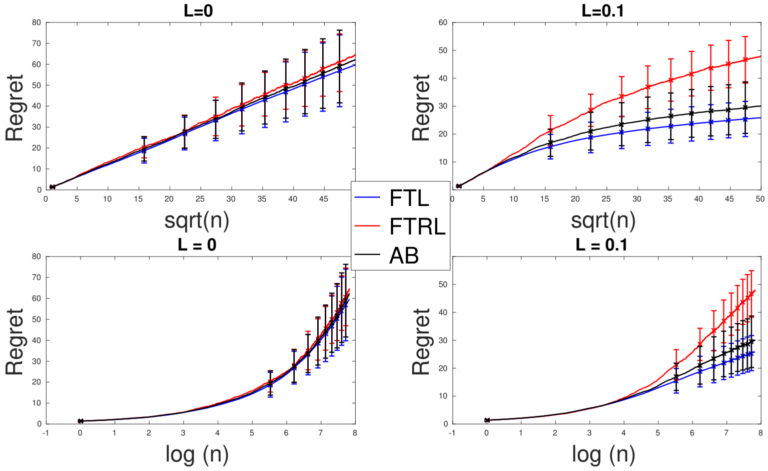

Figure 4: Regret of FTL, FTRL and AB(FTL,FTRL) against time for stochastic data.

5.2 “Half-Adversarial” Data

The half-adversarial data used in this experiment is the optimal solution for the adversary in the *linear game* when $\mathcal{W}$ is the unit ball (Abernethy et al., 2008). This data is generated as follows: The sequence $\hat{f}_t$ for $t = 1, \dots, n$ is generated randomly in the $(d-1)$-dimensional subspace $S = \operatorname{span}\{e_2, \dots, e_d\}$ (here e_i is the ith unit vector in $\mathbb{R}^d$) as follows: $\hat{f}_1$ is drawn from the uniform distribution on the unit sphere of S (actually $\mathbb{S}^{d-1}$. For $t = 2, \dots, n$, $\hat{f}_t$ is drawn from the uniform distribution on the unit sphere of the intersection of S and the hyperplane perpendicular to $\sum_{i=1}^{t-1} \hat{f}_i$ and going through the origin. Then, $f_t = Le_1 + \sqrt{1-L^2}\hat{f}_t$ for some $L \geq 0$.

The results are reported in Fig. 5. When $L = 0$, the regret of both FTL and FTRL grows as $O(\sqrt{n})$. When $L = 0.1$, FTL achieves $O(\log n)$ regret, while the regret of FTRL appears to be $O(\sqrt{n})$. AB(FTL,FTRL) closely matches the regret of FTL.

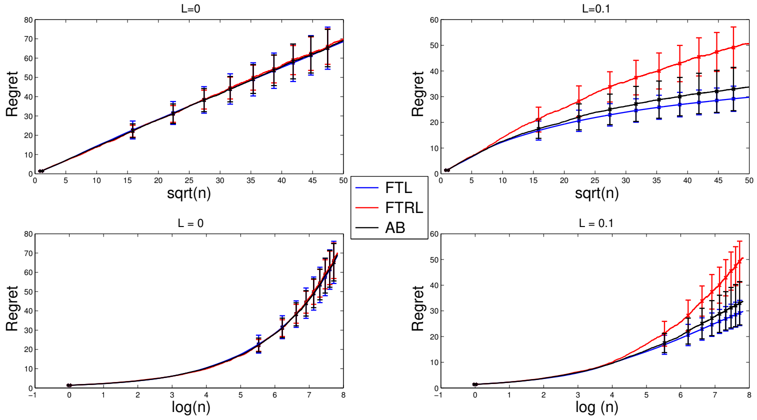

Figure 5: Experimental results for “half-adversarial” data.

5.3 Worst-Case Data

We also tested the algorithms on data where FTL is known to suffer linear regret, mainly to see how well AB(FTL,FTRL) is able to deal with this setting. In this case, we set $f_{t,i} = 0$ for all t and $i \geq 2$, while for the first coordinate, $f_{1,1} = 0.9$, and $f_{t,1} = 2(t \mod 2) - 1$ for $t \geq 2$.

The results are reported in Fig. 6. It can be seen that the regret of FTL is linear (as one can easily verify theoretically), and AB(FTL,FTRL) succeeds to adapt to FTRL, and they both achieve a much smaller $O(\sqrt{n})$ regret.

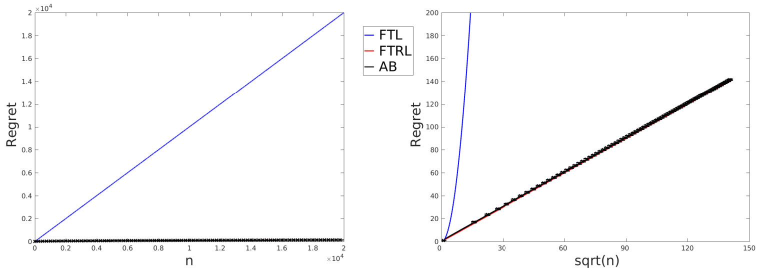

Figure 6: Experimental results for worst-case data.

5.4 The Unit Ball

We close this section by comparing the performance of our adaptive algorithms on the unit ball, namely, FTL, FTSL, FTLR, and AB(FTL,FTRL). All these algorithms are parametrized as above. The problem setup is similar to the stochastic data setting and the worst-case data setting. Again, we consider a 4-dimensional setting, that is, $\mathcal{W}$ is the unit ball in $\mathbb{R}^4$ centered at the origin. The worst-case data is generated exactly as above, while the generation process of the stochastic data is slightly modified to increase the difference between FTLR and FTL: we sample the i.i.d. vectors $\hat{f}_t$ from a zero-mean normal distribution with independent components whose variance is $1/16$, and let $\tilde{f}_t = \hat{f}_t$ if $\|\hat{f}_t\|_2 \le 1$ and $\tilde{f}_t = \hat{f}_t / \left\|\hat{f}_t\right\|_2$ when $\left\|\hat{f}_t\right\|_2 > 1$ (i.e., we only normalize if $\hat{f}_t$ falls outside of the unit ball). The reason of this modification is to encourage the occurrence of the event $\|F_{t-1}\|_2 < \sqrt{t-1}$. Recall that when $\|F_{t-1}\|_2 \ge \sqrt{t-1}$, the prediction of FTRL matches that of FTL, so we are trying to create some data where their behavior is actually different. As a result, we will be able to observe that the predictions of FTL and FTRL are different in the early rounds. Finally, as before, we let $f_t = \tilde{f}_t + Le_1$, and set the time horizon to $n = 20,000$.

The results of the simulation of the stochastic data setting are shown in Figure 7. In the case of $L = 0.1$, FTRL suffers more regret at the beginning for some rounds, but then succeeds to match the performance of FTL. The results of the simulation of the worst-case data setting are shown in Figure 8, where FTSL has similar performance as FTRL.

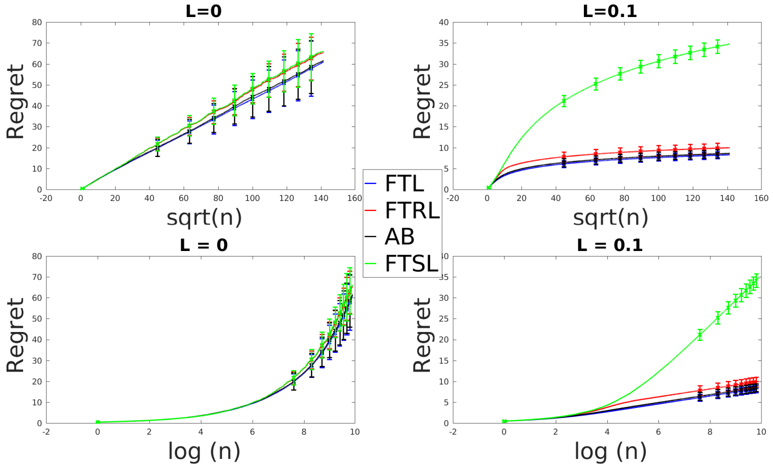

Figure 7: Experimental results for stochastic data when $\mathcal{W}$ is the unit ball.

6. Conclusion

FTL is a simple method that is known to perform well in many settings, while existing worst-case results fail to explain its good performance. While taking a thorough look at why and when FTL can be expected to achieve small regret, we discovered that the curvature of the boundary of the constraint and having average loss vectors bounded away from zero help keep the regret of FTL small. These conditions are significantly different from previous conditions on the curvature of the loss functions which have been considered extensively in the literature. It would be interesting to further investigate this phenomenon for other algorithms or in other learning settings.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments that helped improving the presentation of the paper, in particular, for the alternative proof of (6) based on strong convexity. This work was supported in part by the Alberta Innovates Technology Futures through the Alberta Ingenuity Centre for Machine Learning and by NSERC. During part of this work, T. Lattimore was

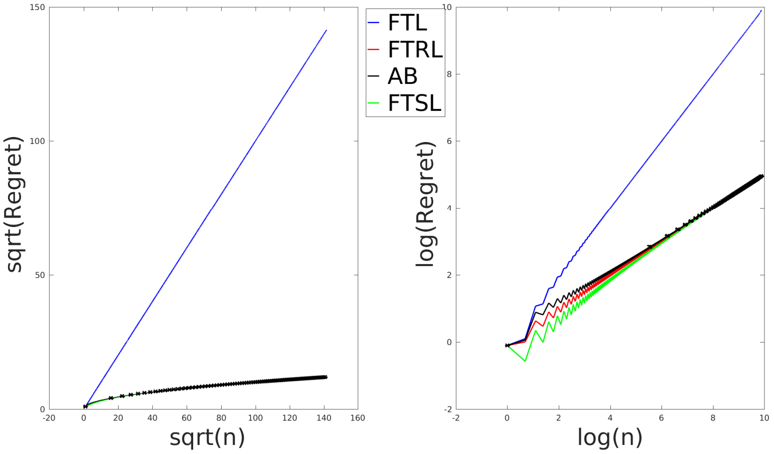

Figure 8: Experimental results for worst-case data when $\mathcal{W}$ is the unit ball.

with the Department of Computing Science, University of Alberta and with the School of Informatics and Computing, Indiana University Bloomington.

Appendix A. Technical Results

A.1 Proof of Proposition 1

Under the extra condition that $\mathcal{W}$ is compact the result follows from Danskin's theorem (e.g., Proposition B.25 of Bertsekas 1999). However, compactness is not required. For completeness, we provide a short, direct proof. We need to show that $\mathcal{Z} = \partial\varphi(\Theta)$ where recall that

$$\partial\varphi(\Theta) = \left\{u \in \mathbb{R}^d \,|\, \varphi(\Theta) + \langle u, \cdot - \Theta\rangle \leq \varphi(\cdot)\right\} = \left\{u \in \mathbb{R}^d \,|\, \varphi(\Theta) \leq \langle u, \Theta\rangle + \varphi(\cdot) - \langle u, \cdot\rangle\right\} .$$

Since $\mathcal{Z} \subset \mathcal{W}$, if $w \in \mathcal{Z}$, $\varphi(\Theta') \geq \langle w, \Theta'\rangle$ for any Θ' by the definition of φ. Hence, $\varphi(\Theta) = \langle w, \Theta\rangle \leq \langle w, \Theta\rangle + \varphi(\Theta') - \langle w, \Theta'\rangle$ for any Θ', implying that $w \in \partial\varphi(\Theta)$.

On the other hand, assume $w \in \partial\varphi(\Theta)$. Then $\varphi(\Theta) \leq \langle w, \Theta\rangle$ since $\varphi(0) = \langle w, 0\rangle = 0$. Since $\mathcal{W}$ is closed, $\mathcal{Z}$ is also closed. Therefore, if $w \notin \mathcal{Z}$, the strict separation theorem (applied to $\{w\}$, a convex compact set, and $\mathcal{Z}$, a convex closed set) implies that there exists $\rho \in \mathbb{R}^d$ such that $\langle z, \rho\rangle < \langle w, \rho\rangle$ for all $z \in \mathcal{Z}$. Let $\Theta' = \Theta + \rho$. Then, $\varphi(\Theta') = \max_{u\in\mathcal{W}}\langle u, \Theta\rangle + \langle u, \rho\rangle < \varphi(\Theta) + \langle w, \Theta' - \Theta\rangle \leq \langle w, \Theta'\rangle \leq \varphi(\Theta')$, a contradiction. Hence, $w \in \mathcal{Z}$.

A.2 Preliminaries in Differential Geometry

In this section we present some extra details on the tools we use from differential geometry. We focus on providing the intuitive picture and foundational results used in the paper, omitting formal definitions that do not directly contribute to this goal. The interested reader can find a detailed formal treatment, for example, in Section 2.5 of the book of Schneider (2014).

A.2.1 Planar Curves

We only consider twice continuously differentiable curves in this paper, defined as injective twice continuously differentiable functions $\gamma : [a, b] \to X$ from an interval $[a, b] \subset \mathbb{R}$ to a differentiable manifold $X \subset \mathbb{R}^d$ such that $\gamma'(u) \neq 0$. Given a differentiable curve γ, we define its length between $\gamma(u)$ and $\gamma(v)$ as $\ell([u, v]) = \int_u^v \|\gamma'(s)\| ds$ for $u, v \in [a, b]$; throughout this section $\|\cdot\|$ denotes the

Euclidean norm, and—with a slight abuse of notation—we will also use ℓ to denote the length of a curve or an interval. Given a twice continuously differentiable bijective mapping $r : [a,b] \to J$ (where $J \subset \mathbb{R}$ is an interval), γ can be reparametrized as a twice continuously differentiable function from J to X, by $\tilde{\gamma}(\ell) = \gamma(r^{-1}(\ell))$. In particular, when the mapping is $r(u) = \ell([a,u])$, that is, the curve is reparametrized by the curve length, we have $r^{-1}(\ell([a,u])) = u$ and $r'(u) = \|\gamma'(u)\| > 0$. Moreover, since $\ell = r(r^{-1}(\ell))$, we also have $1 = \frac{d\, r(r^{-1}(\ell))}{d\,\ell} = r'(u)\frac{d\, r^{-1}(\ell)}{d\,\ell}$ where $r(u) = \ell$. Thus, $\frac{d\, r^{-1}(\ell)}{d\,\ell} = \frac{1}{r'(u)} = \frac{1}{\|\gamma'(u)\|}$, and so

$$\|\tilde{\gamma}'(\ell)\| = \left\| \frac{d\,\gamma(r^{-1}(\ell))}{d\,\ell} \right\| = \left| \frac{d\, r^{-1}(\ell)}{d\,\ell} \right| \|\gamma'(u)\| = 1.$$

Thus, if the curve is parametrized by the its length, its gradient is always a unit vector. For the rest of this section, we always assume this parametrization.

A planar curve is a curve in a 2-dimensional plane. Given a point $\gamma(u)$ on the curve, one can compute its tangent vector in the plane by $\gamma'(u)$. Note that since γ is parametrized by its curve length, $\gamma'(u)$ is a unit vector. To measure how *curved* a planar curve is at point $\gamma(u)$, we define its curvature as

$$\kappa(\gamma(u)) = \left\| \frac{d\,\gamma'(u)}{d\,u} \right\| = \|\gamma''(u)\|_2.$$

Furthermore, since $\gamma'(u)$ is a unit vector,

$$0 = \frac{d\,1}{d\,u} = \frac{d\,\|\gamma'(u)\|^2}{d\,u} = \frac{d\,\langle \gamma'(u),\, \gamma'(u)\rangle}{d\,u} = 2\langle \gamma'(u),\, \gamma''(u)\rangle,$$

thus $\gamma'(u)$ is perpendicular to $\gamma''(u)$.

A.2.2 Manifolds, Tangent Plane, and Principal Curvature

A manifold M of dimension d is a Hausdorff topological space that is locally homeomorphic to $\mathbb{R}^d$. Given a convex body (a convex body is a compact, convex subset of $\mathbb{R}^d$ with non-empty interior) $\mathcal{W} \subset \mathbb{R}^d$, its boundary is a manifold $M = \mathrm{bd}(\mathcal{W})$ of dimension $d-1$. Assume that M is twice continuously differentiable, and let ζ denote the standard embedding map from M to $\mathcal{W}$.[5] Now let $\gamma_1, \gamma_2 : [-1,1] \to M$ be two curves (not necessarily parametrized by curve length), such that $\gamma_1(0) = \gamma_2(0) = w$. The two curves are equivalent at the point w if and only if their derivatives are equal at the point u, that is, $(\zeta \circ \gamma_1)'(0) = (\zeta \circ \gamma_2)'(0)$ and the tangent vector embedded in $\mathbb{R}^d$ associated with this equivalence class is $(\zeta \circ \gamma_1)'(0)$. The set of all the tangent vectors form the tangent space, denoted by $T_w M$ or $T_w \mathcal{W}$. One can verify that $T_w M$ is a $d-1$ dimensional hyperplane in $\mathbb{R}^d$. Note that this definition of tangent space is consistent with the 'natural' tangent plane in the Euclidean space $\mathbb{R}^3$.[6]

Since $T_w M$ is a $d-1$ dimensional hyperplane in $\mathbb{R}^d$, there exists a unique vector that is perpendicular to $T_w M$, is of length 1 and points outward of $\mathcal{W}$ (note that in this sense $M = \mathrm{bd}(\mathcal{W})$ is oriented): this vector is called the Gauss vector at point w for $\mathcal{W}$. The mapping $u_{\mathcal{W}} : \mathrm{bd}(\mathcal{W}) \to \mathbb{S}^{d-1}$ that maps every $w \in \mathrm{bd}(\mathcal{W})$ to the corresponding Gauss vector is called the Gauss map. Since M is twice continuously differentiable, the Gauss map $u_{\mathcal{W}}$ is continuously differentiable. One can actually show that $\nabla u_{\mathcal{W}}(w)$, the so-called Weingarten map, is a self-adjoint operator with nonnegative eigenvalues, which can be represented as a $(d-1) \times (d-1)$ positive semidefinite matrix. The eigenvalues of this matrix (or the self-adjoint operator) are called the principal curvatures of M at

[5] A detailed discussion of manifolds including local charts and atlases can be found in Section 2.5 of the book of Schneider (2014), together with a formal definition of differentiability. To build the intuition required to follow the arguments in the paper it is sufficient to think of a manifold as the boundary of a convex body.

[6] More generally, one can define the tangent space without embedding it to $\mathbb{R}^d$, but, for simplicity, we only consider here the definitions through the embedding ζ, which is always possible and allows to perform calculations in $\mathbb{R}^d$.

point w. Intuitively, how fast the Gauss map $u_{\mathcal{W}}$ changes characterizes the curvature of the manifold M. An interesting property of the operator $\nabla u_{\mathcal{W}}(w)$ is that it maps $T_w M$ to itself: for any unit vector $v \in T_w M$, $0 = \lim_{\epsilon \to 0} \frac{\partial 1}{\partial \epsilon} = \lim_{\epsilon \to 0} \frac{\partial \|u_{\mathcal{W}}(w+\epsilon v)\|_2^2}{\partial \epsilon} = 2 \lim_{\epsilon \to 0} \langle \nabla u_{\mathcal{W}}(w+\epsilon v)v,\, u_{\mathcal{W}}(w+\epsilon v)\rangle = 2\langle \nabla u_{\mathcal{W}}(w)v,\, u_{\mathcal{W}}(w)\rangle$, thus $\nabla u_{\mathcal{W}}(w)v$ is perpendicular to $u_{\mathcal{W}}(w)$, thus belongs to $T_w M$.

A.3 Technical Proofs Related to Strongly Convex Sets and Principal Curvatures

Proof of Proposition 4 We show that (i) implies (ii), (ii) implies (iii), and (iii) implies (i). We start with showing that (i) implies (ii). First note that all principal curvatures of the d-dimensional ball $B = B_{1/\lambda}$ with radius $1/\lambda$ (centered at the origin) are λ. Therefore, (i) and Theorem 3.2.9 of Schneider (2014) implies that there is a convex body $\mathcal{M}$ such that $\mathcal{W} + \mathcal{M} = B$, where for two sets, $S_1, S_2 \subset \mathbb{R}^d$, $S_1 + S_2$ is defined as $\{s_1 + s_2 \,|\, s_1 \in S_1, s_2 \in S_2\}$. For any $\theta \in \mathbb{S}^{d-1}$, let $m_\theta \in \operatorname{argmax}_{m \in \mathcal{M}} \langle m, \theta \rangle$. Then clearly $w_\theta + m_\theta$ maximizes $\langle b, \theta \rangle$ for $b \in \mathcal{W} + \mathcal{M}$. Therefore, $\mathcal{W} + m_\theta$ is a subset of B and touches it at $w_\theta + m_\theta$, or equivalently $\mathcal{W} \subset B - m_\theta$ and they touch each other, and a tangent hyperplane with normal vector θ, in w_θ. This proves that (i) implies (ii).

Next we prove that (ii) implies (iii). Assuming (ii) holds, let $w \in \mathcal{W}$ be any point in the interior of $\mathcal{W}$, and let $p \in \operatorname{bd}(\mathcal{W})$ be the closest boundary point to w, and recall that $T_p\mathcal{W}$ is the tangent space of $\mathcal{W}$ at p. By construction, $B_{\|w-p\|_2}(w)$ touches the boundary of $\mathcal{W}$ at p (in the sense that they do not intersect, but they can have multiple common points), and so $w - p$ is orthogonal to $T_p\mathcal{W}$. Therefore, $B_{\|w-p\|_2}(w)$ also touches the boundary of the ball $B = B_{1/\lambda}(p + \frac{w-p}{\lambda\|w-p\|_2})$, which contains $\mathcal{W}$ by assumption (ii). Now consider any two points $x, y \in \mathcal{W}$ and $\gamma \in [0, 1]$ such that $w = \gamma x + (1-\gamma)y$. Then the ball with radius $\lambda\gamma(1-\gamma)\|x-y\|_2^2/2$ centered at w is contained in B, since B is λ-strongly convex. But then its radius is at most $\|p - w\|_2$, and so it is also contained in $\mathcal{W}$. This shows that $\mathcal{W}$ is λ-strongly convex, thus (iii) holds.

To finish the proof of the proposition, assume (iii). To prove that (i) holds, we have to show that for any point w on $\operatorname{bd}(\mathcal{W})$ and for any unit vector $v \in T_w\mathcal{W}$, the curvature of the boundary along v is at least λ. Using the same notations as in Fig. 2 in Section 3.1, let P be the hyperplane spanned by v and the outer normal vector $u_{\mathcal{W}}(w)$ of $\mathcal{W}$ at point w, and consider the planar curve γ defined by $\operatorname{bd}(\mathcal{W}) \cap P$. Using v as the axis of a local 2-dimensional coordinate system, a point $\gamma(s)$ on the curve γ in the neighborhood of w can be expressed as $\gamma(s) = w + sv - f(s)u_{\mathcal{W}}(w)$, where w serves as the origin in the local coordinate system, f is the restriction of the function $f_w(sv)$ (see Section 3.1) to P (to simplify the notation, we denote it by $f(s)$, omitting v and w), and the curve γ is the epigraph of the function f, as in Fig. 9.

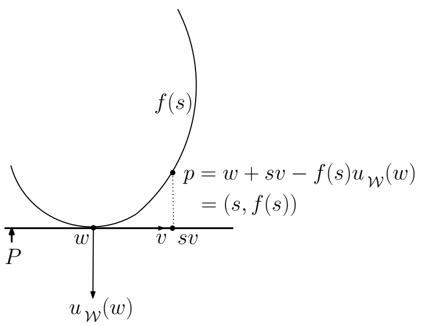

Figure 9: The local coordinate system at w.

Note that $f'(0) = 0$, and by Proposition 2.1 of Pressley (2010), the curvature of γ at p can be obtained as

$$\left. \frac{f''(s)}{\sqrt{1 + f'(s)^2}^3} \right|_{s=0} = f''(0) \; .$$

Now since $w(s), w(-s) \in \mathcal{W}$ for a sufficiently small s, the strong convexity of $\mathcal{W}$ applied to $w(s)$ and $w(-s)$ with $\gamma = 1/2$ implies that $q = \frac{w(s)+w(-s)}{2} + \frac{\lambda}{8}\|w(s) - w(-s)\|_2^2 u \in \mathcal{W}$. Substituting the definition of $w(s)$ and $w(-s)$, we get

$$q = p - u\left[\frac{f(s) + f(-s)}{2} - \frac{\lambda}{8}\Big(4s^2 + (f(s) - f(-s))^2\Big)\right] .$$

Therefore, $q \in \mathcal{W}$ implies $f(s) + f(-s) \geq \lambda s^2$, and so

$$f''(0) = \lim_{s \to 0} \frac{\frac{f(s)-f(0)}{s} - \frac{f(0)-f(-s)}{s}}{s} = \frac{f(s)+f(-s)}{s^2} \geq \lambda.$$

Thus (i) holds, finishing the proof of the proposition. ■

Proof of a weakened variant of (6) **based on strong convexity**
Given θ_1 and θ_2, for any $0 < \gamma < 1$, define $\theta_\gamma = \gamma\theta_1 + (1-\gamma)\theta_2$ and

$$w_\gamma = \gamma w^{(1)} + (1-\gamma)w^{(2)} + \frac{\lambda_0}{2}\gamma(1-\gamma)\|w^{(1)} - w^{(2)}\|_2^2 \frac{\theta_\gamma}{\|\theta_\gamma\|_2} \ .$$

By the strong convexity of $\mathcal{W}$, $w_\gamma \in \mathcal{W}$, and so by the definition and convexity of the support function Φ, we have

$$\langle w_\gamma, \theta_\gamma \rangle \leq \Phi(\theta_\gamma) \leq \gamma\Phi(\theta_1) + (1-\gamma)\Phi(\theta_2) = \gamma\langle w^{(1)}, \theta_1 \rangle + (1-\gamma)\langle w^{(2)}, \theta_2 \rangle.$$

Plugging in the definitions of w_γ and θ_γ, and applying the Cauchy-Schwarz inequality, we obtain

$$\frac{\lambda_0}{2}\gamma(1-\gamma)\|w^{(1)} - w^{(2)}\|_2^2\|\theta_\gamma\|_2 \leq \gamma(1-\gamma)\langle \theta_1 - \theta_2, w^{(1)} - w^{(2)} \rangle \leq \gamma(1-\gamma)\|\theta_1 - \theta_2\|_2\|w^{(1)} - w^{(2)}\|_2 \ .$$

Rearranging and letting $\gamma \to 0$ implies

$$\|w^{(1)} - w^{(2)}\|_2 \leq \frac{2}{\lambda_0}\frac{\|\theta_1 - \theta_2\|_2}{\|\theta_2\|_2} \ .$$

Finally, the definition of $w^{(2)}$ and the Cauchy-Schwarz inequality yields

$$\begin{aligned}\langle w^{(1)} - w^{(2)}, \theta_1 \rangle &\leq \langle w^{(1)} - w^{(2)}, \theta_1 \rangle + \underbrace{\langle w^{(2)} - w^{(1)}, \theta_2 \rangle}_{\geq 0} = \langle w^{(1)} - w^{(2)}, \theta_1 - \theta_2 \rangle \\ &\leq \|w^{(1)} - w^{(2)}\|_2\|\theta_1 - \theta_2\|_2 \leq \frac{2}{\lambda_0}\frac{\|\theta_1 - \theta_2\|_2^2}{\|\theta_2\|_2} \ , \qquad (24)\end{aligned}$$

finishing the proof. ■

Proof of Example 1 We start with proving the last statement, part (iv), which implies the rest. Fix $w \in \mathrm{bd}(\mathcal{W})$. Note that $\phi'(w)$ is a normal vector at w for $\mathrm{bd}(\mathcal{W})$, thus $T_w\mathcal{W} = \{v : \langle v, \phi'(w) \rangle\}$. Then the Gauss map $u_\mathcal{W}$ of $\mathcal{W}$ satisfies $u_\mathcal{W}(w) = \frac{\phi'(w)}{\|\phi'(w)\|_2}$ for $w \in \mathrm{bd}(\mathcal{W})$. According to Schneider (2014, page 105), the principal curvatures of $\mathcal{W}$ at w are the eigenvalues of the Weingarten map $W_w(v)$, which is a linear map from $T_w\mathcal{W}$ to itself defined through the derivative of $u_\mathcal{W}$: $W_w(v) = \langle \frac{du_\mathcal{W}}{dw}, v \rangle$. In our case,

$$W_w(v) = \left\langle \frac{du_\mathcal{W}}{dw}, v \right\rangle = \frac{\nabla^2\phi(w)v}{\|\phi'(w)\|_2} - \frac{\phi'(w)\nabla^2\phi(w)\phi'(w)^\top v}{\|\phi'(w)\|_2^3} = \frac{\nabla^2\phi(w)v}{\|\phi'(w)\|_2} \ ,$$

where in the last step we used that $\phi'(w)$ is orthogonal to the tangent space $T_w\mathcal{W}$ (since it is parallel to the normal vector $u_\mathcal{W}(w)$), and $v \in T_w\mathcal{W}$. Therefore, the smallest principal curvature at w is the smallest eigenvalue $\min_{v \in \mathbb{S}^{d-1}:\langle \phi'(w), v \rangle = 0} \frac{v^\top\nabla^2\phi(w)v}{\|\phi'(w)\|_2}$. Taking minimum over all $w \in \mathrm{bd}(\mathcal{W})$ finishes the proof.

Now part (ii) follows for $\phi(w) = w^\top Qw$, as we need to minimize $v^\top Qv/\|w^\top Q\|_2$. It is easy to see that the denominator is maximized when $w \in \mathrm{bd}(\mathcal{W})$ is an eigenvector of Q corresponding to

$\lambda_{\max}$ (with length $1/\sqrt{\lambda_{\max}}$), and the numerator is minimized (for arbitrary $v \in \mathbb{S}^{d-1}$) when v is an eigenvector of Q corresponding to $\lambda_{\min}$. Since the two eigenvectors, w and v are orthogonal (or can be chosen to be orthogonal if they are not unique), v is orthogonal to $\phi'(w) = Qw = \lambda_{\max} w$, and hence it is a valid minimizer. This completes the proof of part (ii), and part (i) follows as a special case.

Part (iii) follows similarly: Due to symmetry, it is enough to consider w in the nonnegative quadrant (i.e., $w_i \geq 0$ for all i). Calculating the first and second derivatives of $\phi(w) = \|w\|_p = \left(\sum_{i=1}^d w_i^p\right)^{1/p}$, for any $w \in \mathrm{bd}(\mathcal{W})$ (i.e., $\|w\|_p = 1$), we obtain

$$\phi'(w) = \left(\sum_{i=1}^d w_i^p\right)^{1/p-1} w^{\odot(p-1)} = w^{\odot(p-1)} \quad \text{and} \quad \nabla^2\phi(w) = (p-1)\,\mathrm{diag}\left(w_1^{p-2}, \cdots, w_d^{p-2}\right).$$

When $p > 2$, picking $w = (1, 0, 0, \cdots, 0)$, one can easily verify that $\lambda_0 = 0$. For $1 < p \leq 2$, $|w_i|^{p-2} \geq 1$ since $|w_i| \leq 1$ by the assumption that $\|w\|_p = 1$. Thus, the minimum eigenvalue of $\mathrm{diag}\left(w_1^{p-2}, \cdots, w_d^{p-2}\right)$ is at least 1, and so $\lambda_0 \geq (p-1)/\|w^{\odot(p-1)}\|_2$ since $v^\top \mathrm{diag}\left(w_1^{p-2}, \cdots, w_d^{p-2}\right) v \geq 1$. Defining q via $1/p + 1/q = 1$, for any $w \in \mathrm{bd}(\mathcal{W})$ (i.e., with $\|w_p\| = 1$), Hölder's inequality implies

$$\|w^{\odot(p-1)}\|_2 \leq d^{\frac{1}{2}-\frac{1}{q}} \|w^{\odot(p-1)}\|_q = d^{\frac{1}{2}-\frac{1}{q}} \left(\sum_{i=1}^d w_i^{(p-1)q}\right)^{\frac{1}{q}} = d^{\frac{1}{2}-\frac{1}{q}} \left(\sum_{i=1}^d w_i^p\right)^{\frac{1}{q}} = d^{\frac{1}{2}-\frac{1}{q}} = d^{\frac{1}{p}-\frac{1}{2}}.$$

Thus, $\lambda_0 \geq (p-1)d^{\frac{1}{2}-\frac{1}{p}}$, as desired. ■

A.4 Technical Lemmas for the Lower Bound, Theorem 9

Lemma 17 *Under the assumptions of Theorem 9, for any* $0 < P_1, P_2 < 1$,

$$\left\langle w^{P_2} - w^{P_1}, f^{P_1} \right\rangle \geq \frac{\lambda L}{2} \frac{\left(\frac{2P_2 - 2P_1}{\lambda L}\right)^2}{\sqrt{1 + \left(\frac{1-2P_1}{\lambda L}\right)^2}\left(1 + \left(\frac{1-2P_2}{\lambda L}\right)^2\right)}.$$

Proof It is easy to see that for any p, w^p is on the boundary of $\mathcal{W}$, that is, $w^p = \operatorname{argmin}_{w \in \mathcal{W}} \langle w, f^p \rangle = (\cos(\varphi^p), \lambda \sin(\varphi^p))$ for some φ^p. Then $\langle w^p, f^p \rangle = (2p-1)\cos(\varphi^p) - \lambda L \sin(\varphi^p)$, and so taking the derivative it is easy to verify that $\tan(\varphi^p) = \frac{\lambda L}{1-2p}$ and $\sin(\varphi^p) = \frac{\lambda L}{\sqrt{(\lambda L)^2 + (1-2p)^2}} > 0$. Thus, $1 - 2P_1 = \frac{\lambda L \cos(\varphi^{P_1})}{\sin(\varphi^{P_1})}$. To simplify notation, let $\varphi_1 = \varphi^{P_1}$ and $\varphi_2 = \varphi^{P_2}$. Then,

$$\begin{aligned}
\langle w^{P_2} - w^{P_1}, f^{P_1} \rangle &= \left\langle \begin{pmatrix} \cos\varphi_2 - \cos\varphi_1 \\ \lambda(\sin\varphi_2 - \sin\varphi_1) \end{pmatrix}, \begin{pmatrix} \frac{-\lambda L \cos\varphi_1}{\sin\varphi_1} \\ -L \end{pmatrix} \right\rangle \\
&= -\lambda L \left((\cos(\varphi_2) - \cos(\varphi_1)) \frac{\cos(\varphi_1)}{\sin(\varphi_1)} + (\sin(\varphi_2) - \sin(\varphi_1)) \right) \\
&= \frac{-\lambda L}{\sin(\varphi_1)} \left(\cos(\varphi_2)\cos(\varphi_1) - \cos^2(\varphi_1) + \sin(\varphi_1)\sin(\varphi_2) - \sin^2(\varphi_1) \right) \\
&= \frac{\lambda L}{\sin(\varphi_1)} (1 - \cos(\varphi_2)\cos(\varphi_1) - \sin(\varphi_1)\sin(\varphi_2)) \\
&= \frac{\lambda L}{\sin(\varphi_1)} (1 - \cos(\varphi_1 - \varphi_2)) \\
&= \frac{\lambda L}{\sin(\varphi_1)} \left(\frac{1}{2}(\cos(\varphi_1 - \varphi_2) - 1)^2 + \frac{1}{2}\sin^2(\varphi_1 - \varphi_2) \right)
\end{aligned}$$

$$\begin{aligned} &\geq \frac{\lambda L}{2\sin(\varphi_1)}\sin^2(\varphi_1-\varphi_2) \\ &= \frac{\lambda L}{2}\sin(\varphi_1)\sin^2\varphi_2\left(\cot(\varphi_1)-\cot(\varphi_2)\right)^2 \ . \end{aligned}$$

The proof is finished by substituting $\cot(\varphi_i) = \frac{1-2P_i}{\lambda L}$, $\sin(\varphi_1) = \frac{1}{\sqrt{1+(\frac{1-2P_1}{\lambda L})^2}}$ and $\sin^2(\varphi_2) = \frac{1}{1+(\frac{1-2P_2}{\lambda L})^2}$. ■

Lemma 18 (Concentration of $\hat{P}_t$) *For any $u > 0$,*

$$\mathbb{P}\left[|\hat{P}_t - P| > \frac{K}{2K+t}|1-2P| + \frac{t}{2K+t}u \,\middle|\, P\right] \leq 2\exp(-tu^2) \ .$$

Proof Recall that $\hat{P}_t = \frac{K+\sum_{i=1}^t X_i}{2K+t}$. Thus,

$$\begin{aligned} \mathbb{P}\left[|\hat{P}_t - P| > u \,\middle|\, P\right] &= \mathbb{P}\left[\left|\frac{K+\sum_{i=1}^t X_i}{2K+t} - P\right| > \frac{K}{2K+t}|1-2P| + \frac{t}{2K+t}u \,\middle|\, P\right] \\ &= \mathbb{P}\left[\left|\sum_{i=1}^t X_i - Pt + K(1-2P)\right| > K|1-2P| + tu \,\middle|\, P\right] \\ &\leq \mathbb{P}\left[\left|\sum_{i=1}^t X_i - Pt\right| > tu \,\middle|\, P\right], \end{aligned} \tag{25}$$

where the last inequality is due to $\mathbb{P}[|A+b| > c] \leq \mathbb{P}[|A| > c - |b|]$. Note that conditioned on P, $X_1, \ldots, X_t$ are independent Bernoulli random variables with expectation P, thus (25) holds by Hoeffding's inequality (see, e.g., (Cesa-Bianchi and Lugosi, 2006, Corollary A.1)). ■

Lemma 19

$$\mathbb{E}\left[(P-\hat{P}_t)^2 \,\middle|\, P\right] = \frac{K^2(1-2P)^2}{(2K+t)^2} + \frac{tP(1-P)}{(2K+t)^2} \ .$$

Proof Recall that $\hat{P}_t = \frac{K+\sum_{i=1}^t X_i}{2K+t}$.Thus,

$$\begin{aligned} \mathbb{E}\left[(P-\hat{P}_t)^2 \,\middle|\, P\right] &= \mathbb{E}\left[\left(\frac{K(1-2P)}{2K+t} + \frac{\sum_{i=1}^t X_i - Pt}{2K+t}\right)^2 \,\middle|\, P\right] \\ &= \frac{K^2(1-2P)^2}{(2K+t)^2} + \frac{1}{(2K+t)^2}\mathbb{E}\left[\left(\sum_{i=1}^t X_i - tP\right)^2 \,\middle|\, P\right] \\ &= \frac{K^2(1-2P)^2}{(2K+t)^2} + \frac{tP(1-P)}{(2K+t)^2} \ , \end{aligned}$$

where the second equality is due to $\mathbb{E}\left[\sum_{i=1}^t X_i - Pt \,\middle|\, P\right] = 0$, and the last equality is due to that conditioned on P, $\sum_{i=1}^t X_i$ has a Binomial distribution with parameters t and P. ■

References

Y. Abbasi-Yadkori. Forced-exploration based algorithms for playing in bandits with large action sets. Master's thesis, University of Alberta, 2009.

J. Abernethy, P.L. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Conference on Learning Theory (COLT)*, pages 415–424, 2008.

J. Abernethy, C. Lee, A. Sinha, and A. Tewari. Online linear optimization via smoothing. In *Conference on Learning Theory (COLT)*, pages 807–823, 2014.

P.L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pages 65–72, 2007.

D. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 1999.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford University Press, 2013.

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

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. Information Theory*, 50(9):2050–2057, 2004.

D.J. Foster, A. Rakhlin, and K. Sridharan. Adaptive online learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3357–3365, 2015.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

A.A. Gaivoronski and F. Stella. Stochastic nonstationary optimization for finding universal portfolios. *Annals of Operations Research*, 100(1–4):165–188, 2000.

D. Garber and E. Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *International Conference on Machine Learning (ICML)*, pages 541–549, 2015.

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

R. Huang, T. Lattimore, A. György, and Cs. Szepesvári. Following the leader and fast rates in linear prediction: curved constraint sets and other regularities. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4970–4978. 2016.

M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11:517–553, 2010.

S. M. Kakade and S. Shalev-Shwartz. Mind the duality gap: Logarithmic regret algorithms for online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1457–1464, 2009.

Wouter M. Koolen, Peter Grünwald, and Tim van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. *CoRR*, abs/1605.06439, 2016. URL `http://arxiv.org/abs/1605.06439`.

W. Kotłowski. Minimax strategy for prediction with expert advice under stochastic assumptions. *Algorithmic Learning Theory (ALT)*, 2016.

E.S. Levitin and B.T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.

H.B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and implicit updates. arXiv, 2010. URL `http://arxiv.org/abs/1009.3240`.

N. Merhav and M. Feder. Universal sequential learning and decision from individual data sequences. In *ACM Workshop on Computational Learning Theory (COLT)*, pages 413—427, 1992.

F. Orabona, N. Cesa-Bianchi, and C. Gentile. Beyond logarithmic bounds in online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 823–831, 2012.

E. S. Polovinkin. Strongly convex analysis. *Sbornik: Mathematics*, 187(2):259, 1996.

A. N. Pressley. *Elementary differential geometry*. Springer Science & Business Media, 2010.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory (COLT)*, pages 993–1019, 2013.

A. Sani, G. Neu, and A. Lazaric. Exploiting easy data in online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 810–818, 2014.

R. Schneider. *Convex Bodies: The Brunn–Minkowski Theory*. Encyclopedia of Mathematics and its Applications. Cambridge Univ. Press, 2nd edition, 2014.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2012.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

T. van Erven, P. Grünwald, N. Mehta, M. Reid, and R. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research (JMLR)*, 16:1793–1861, 2015. Special issue in Memory of Alexey Chervonenkis.