Adaptive Bound Optimization for Online Convex Optimization

H. Brendan McMahan
Google, Inc.
mcmahan@google.com

Matthew Streeter
Google, Inc.
mstreeter@google.com

July 8, 2010

Abstract

We introduce a new online convex optimization algorithm that adaptively chooses its regularization function based on the loss functions observed so far. This is in contrast to previous algorithms that use a fixed regularization function such as L_2-squared, and modify it only via a single time-dependent parameter. Our algorithm's regret bounds are worst-case optimal, and for certain realistic classes of loss functions they are much better than existing bounds. These bounds are problem-dependent, which means they can exploit the structure of the actual problem instance. Critically, however, our algorithm does not need to know this structure in advance. Rather, we prove competitive guarantees that show the algorithm provides a bound within a constant factor of the best possible bound (of a certain functional form) in hindsight.

1 Introduction

We consider online convex optimization in the full information feedback setting. A closed, bounded convex feasible set $\mathcal{F} \subseteq \mathbb{R}^n$ is given as input, and on each round $t = 1, \ldots, T$, we must pick a point $x_t \in \mathcal{F}$. A convex loss function f_t is then revealed, and we incur loss $f_t(x_t)$. Our regret at the end of T rounds is

$$\text{Regret} \equiv \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{F}} \sum_{t=1}^{T} f_t(x). \tag{1}$$

Existing algorithms for online convex optimization are worst-case optimal in terms of certain fundamental quantities. In particular, online gradient descent attains a bound of $\mathcal{O}(DM\sqrt{T})$ where D is the L_2 diameter of the feasible set and M is a bound on L_2-norm of the gradients of the loss functions. This bound is tight in the worst case, in that it is possible to construct problems where this much regret is inevitable. However, this does not mean that an algorithm that achieves this bound is optimal in a practical sense, as on easy problem instances such an algorithm is still allowed to incur the worst-case regret. In particular, although this bound is minimax optimal when the feasible set is a hypersphere [Abernethy et al., 2008], we will see that much better algorithms exist when the feasible set is the hypercube.

To improve over the existing worst-case guarantees, we introduce additional parameters that capture more of the problem's structure. These parameters depend on the loss functions, which are not known in advance. To address this, we first construct functional upper bounds on regret $B_R(\theta_1, \dots, \theta_T; f_1, \dots, f_T)$ that depend on both (properties of) the loss functions f_t and algorithm parameters θ_t. We then give algorithms for choosing the parameters θ_t adaptively (based only on $f_1, f_2, \dots, f_{t-1}$) and prove that these adaptive schemes provide a regret bound that is only a constant factor worse than the best possible regret bound of the form B_R. Formally, if for all possible function sequences $f_1, \dots f_T$ we have

$$B_R(\theta_1, \dots, \theta_T; f_1, \dots, f_T) \le \kappa \inf_{\theta'_1, \dots, \theta'_T \in \Theta^T} B_R(\theta'_1, \dots, \theta'_T; f_1, \dots, f_T)$$

for the adaptively-selected θ_t, we say the adaptive scheme is κ-competitive for the bound optimization problem. In Section 1.2, we provide realistic examples where known bounds are much worse than the problem-dependent bounds obtained by our algorithm.

1.1 Follow the proximally-regularized leader

We analyze a *follow the regularized leader* (FTRL) algorithm that adaptively selects regularization functions of the form

$$r_t(x) = \frac{1}{2} \|(Q_t^{\frac{1}{2}}(x - x_t)\|_2^2$$

where Q_t is a positive semidefinite matrix. Our algorithm plays $x_1 = 0$ on round 1 (we assume without loss of generality that $0 \in \mathcal{F}$), and on round $t+1$, selects the point

$$x_{t+1} = \arg\min_{x \in \mathcal{F}} \left(\sum_{\tau=1}^{t} \big(r_\tau(x) + f_\tau(x) \big) \right). \tag{2}$$

In contrast to other FTRL algorithms, such as the dual averaging method of Xiao [2009], we center the additional regularization at the current feasible point x_t rather than at the origin. Accordingly, we call this algorithm *follow the proximally-regularized leader* (FTPRL). This proximal centering of additional regularization is similar in spirit to the optimization solved by online gradient descent (and more generally, online mirror descent, [Cesa-Bianchi and Lugosi, 2006]). However, rather than considering only the current gradient, our algorithm considers the sum of all previous gradients, and so solves a global rather than local optimization on each round. We discuss related work in more detail in Section 4.

The FTPRL algorithm allows a clean analysis from first principles, which we present in Section 2. The proof techniques are rather different from those used for online gradient descent algorithms, and will likely be of independent interest.

We write $\vec{Q_T}$ as shorthand for $(Q_1, Q_2, \dots, Q_T)$, with $\vec{g_T}$ defined analogously. For a convex set $\mathcal{F}$, we define $\mathcal{F}_{\text{sym}} = \{x - x' \mid x, x' \in \mathcal{F}\}$. Using this notation, we can state our regret bound as

$$\text{Regret} \le B_R(\vec{Q_T}, \vec{g_T}) \equiv \frac{1}{2} \sum_{t=1}^{T} \max_{\hat{y} \in \mathcal{F}_{\text{sym}}} \left(\hat{y}^\top Q_t \hat{y} \right) + \sum_{t=1}^{T} g_t^\top Q_{1:t}^{-1} g_t \tag{3}$$

where g_t is a subgradient of f_t at x_t and $Q_{1:t} = \sum_{\tau=1}^{t} Q_\tau$. We prove competitive ratios with respect to this B_R for several adaptive schemes for selecting the Q_t matrices. In particular, when the FTPRL-Diag scheme is run on a hyperrectangle (a set of the form $\{x \mid x_i \in [a_i, b_i]\} \subseteq \mathbb{R}^n$), we achieve

$$\text{Regret} \le \sqrt{2} \inf_{\vec{Q} \in \mathcal{Q}_{\text{diag}}^T} B_R(\vec{Q_T}, \vec{g_T})$$

where $\mathcal{Q}_{\text{diag}} = \{\text{diag}(\lambda_1, \ldots, \lambda_n) \mid \lambda_i \geq 0\}$. When the FTPRL-Scale scheme is run on a feasible set of the form $\mathcal{F} = \{x \mid \|Ax\|_2 \leq 1\}$ for $A \in S^n_{++}$, it is competitive with arbitrary positive semidefinite matrices:

$$\text{Regret} \leq \sqrt{2} \inf_{\vec{Q} \in (S^n_+)^T} B_R(\vec{Q_T}, \vec{g_T}) \ .$$

Our analysis of FTPRL reveals a fundamental connection between the shape of the feasible set and the importance of choosing the regularization matrices adaptively. When the feasible set is a hyperrectangle, FTPRL-Diag has stronger bounds than known algorithms, except for degenerate cases where the bounds are identical. In contrast, when the feasible set is a hypersphere, $\{x \mid \|x\|_2 \leq 1\}$, the bound B_R is always optimized by choosing $Q_t = \lambda_t I$ for suitable $\lambda_t \in \mathbb{R}$. The FTPRL-Scale scheme extends this result to hyperellipsoids by applying a suitable transformation. These results are presented in detail in Section 3.

1.2 The practical importance of adaptive regularization

In the past few years, online algorithms have emerged as state-of-the-art techniques for solving large-scale machine learning problems [Bottou and Bousquet, 2008, Zhang, 2004]. Two canonical examples of such large-scale learning problems are text classification on large datasets and predicting click-through rates for ads on a search engine. For such problems, extremely large feature sets may be considered, but many features only occur rarely, while few occur very often. Our diagonal-adaptation algorithm offers improved bounds for problems such as these.

As an example, suppose $\mathcal{F} = [-\frac{1}{2}, \frac{1}{2}]^n$ (so $D = \sqrt{n}$). On each round t, the ith component of $\nabla f_t(x_t)$ (henceforth $g_{t,i}$) is 1 with probability $i^{-\alpha}$, and is 0 otherwise, for some $\alpha \in [1, 2)$. Such heavy-tailed distributions are common in text classification applications, where there is a feature for each word. In this case, gradient descent with a global learning rate[1] obtains an expected regret bound of $O(\sqrt{nT})$. In contrast, the algorithms presented in this paper will obtain expected regret on the order of

$$\mathbb{E}\left[\sum_{i=1}^{n} \sqrt{\sum_{t=1}^{T} g_{t,i}^2}\right] \leq \sum_{i=1}^{n} \sqrt{\sum_{t=1}^{T} \mathbb{E}\left[g_{t,i}^2\right]} = \sum_{i=1}^{n} \sqrt{T i^{-\alpha}} = O(\sqrt{T} \cdot n^{1-\frac{\alpha}{2}})$$

using Jensen's inequality. This bound is never worse than the $O(\sqrt{nT})$ bound achieved by ordinary gradient descent, and can be substantially better. For example, in problems where a constant fraction of examples contain a new feature, n is $\Omega(T)$ and the bound for ordinary gradient descent is vacuous. In contrast, the bound for our algorithm is $O(T^{\frac{3-\alpha}{2}})$, which is sublinear for $\alpha > 1$.

This performance difference is not merely a weakness in the regret bounds for ordinary gradient descent, but is a difference in actual regret. In concurrent work [Streeter and McMahan, 2010], we showed that for some problem families, a per-coordinate learning rate for online gradient descent provides asymptotically less regret than even the best non-increasing global learning rate (chosen in hindsight, given the observed loss functions). This construction can be adapted to FTPRL as:

Theorem 1. *There exists a family of online convex optimization problems, parametrized by the number of rounds T, where online subgradient descent with a non-increasing learning rate sequence (and FTPRL with non-decreasing coordinate-constant regularization) incurs regret at least $\Omega(T^{\frac{2}{3}})$, whereas FTPRL with appropriate diagonal regularization matrices Q_t has regret $O(\sqrt{T})$.*

[1] The $\mathcal{O}(DM\sqrt{T})$ bound (mentioned in the introduction) based on a $1/\sqrt{t}$ learning rate gives $\mathcal{O}(n\sqrt{T})$ here; to get $\mathcal{O}(\sqrt{nT})$ a global rate based on $\|g_t^2\|$ is needed, e.g., Corollary 1.

In fact, any online learning algorithm whose regret is $O(MD\sqrt{T})$ (where D is the L_2 diameter of the feasible region, and M is a bound on the L_2 norm of the gradients) will suffer regret $\Omega(T^{\frac{2}{3}})$ on this family of problems. Note that this does not contradict the $O(MD\sqrt{T})$ upper bound on the regret, because in this family of problems $D = T^{\frac{1}{6}}$ (and $M = 1$).

1.3 Adaptive algorithms and competitive ratios

In Section 3, we introduce specific schemes for selecting the regularization matrices Q_t for FTPRL, and show that for certain feasible sets, these algorithms provide bounds within a constant factor of those for the best post-hoc choice of matrices, namely

$$\inf_{\vec{Q_T} \in \mathcal{Q}^T} B_R(\vec{Q_T}, \vec{g_T}) \tag{4}$$

where $\mathcal{Q} \subseteq S_+^n$ is a set of allowed matrices; S_+^n is the set of symmetric positive semidefinite $n \times n$ matrices, with S_{++}^n the corresponding set of symmetric positive definite matrices. We consider three different choices for $\mathcal{Q}$: the set of coordinate-constant matrices $\mathcal{Q}_{\text{const}} = \{\alpha I \mid \alpha \geq 0\}$; the set of non-negative diagonal matrices,

$$\mathcal{Q}_{\text{diag}} = \{\text{diag}(\lambda_1, \ldots, \lambda_n) \mid \lambda_i \geq 0\}\,;$$

and, the full set of positive-semidefinite matrices, $\mathcal{Q}_{\text{full}} = S_+^n$.

We first consider the case where the feasible region is an L_p unit ball, namely $\mathcal{F} = \{x \mid \|x\|_p \leq 1\}$. For $p \in [1, 2]$, we show that a simple algorithm (an analogue of standard online gradient descent) that selects matrices from $\mathcal{Q}_{\text{const}}$ is $\sqrt{2}$-competitive with the best post-hoc choice of matrices from the full set of positive semidefinite matrices $\mathcal{Q}_{\text{full}} = S_+^n$. This algorithm is presented in Corollary 1, and the competitive ratio is proved in Theorem 6.

In contrast to the result for $p \in [1, 2]$, we show that for L_p balls with $p > 2$ a coordinate-independent choice of matrices ($Q_t \in \mathcal{Q}_{\text{const}}$) does not in general obtain the post-hoc optimal bound (see Section 3.3), and hence per-coordinate adaptation can help. The benefit of per-coordinate adaptation is most pronounced for the L_∞-ball, where the coordinates are essentially independent. In light of this, we develop an efficient algorithm (FTPRL-Diag, Algorithm 1) for adaptively selecting Q_t from $\mathcal{Q}_{\text{diag}}$, which uses scaling based on the width of $\mathcal{F}$ in the coordinate directions (Corollary 2). In this corollary, we also show that this algorithm $\sqrt{2}$-competitive with the best post-hoc choice of matrices from $\mathcal{Q}_{\text{diag}}$ when the feasible set is a hyperrectangle.

While per-coordinate adaptation does not help for the unit L_2-ball, it can help when the feasible set is a hyperellipsoid. In particular, in the case where $\mathcal{F} = \{x \mid \|Ax\|_2 \leq 1\}$ for $A \in S_{++}^n$, we show that an appropriate transformation of the problem can produce significantly better regret bounds. More generally, we show (see Theorem 5) that if one has a κ-competitive adaptive FTPRL scheme for the feasible set $\{x \mid \|x\| \leq 1\}$ for an arbitrary norm, it can be extended to provide a κ-competitive algorithm for feasible sets of the form $\{x \mid \|Ax\| \leq 1\}$. Using this result, we can show FTPRL-Scale is $\sqrt{2}$-competitive with the best post-hoc choice of matrices from S_+^n when $\mathcal{F} = \{x \mid \|Ax\|_2 \leq 1\}$ and $A \in S_{++}^n$; it is $\sqrt{2}$-competitive with $\mathcal{Q}_{\text{diag}}$ when $\mathcal{F} = \{x \mid \|Ax\|_p \leq 1\}$ for $p \in [1, 2)$.

Of course, in many practical applications the feasible set may not be so nicely characterized. We emphasize that our algorithms and analysis are applicable to arbitrary feasible sets, but the quality of the bounds and competitive ratios will depend on how tightly the feasible set can be approximated by a suitably chosen transformed norm ball. In Theorem 3, we show in particular that when FTPRL-Diag is applied to an arbitrary feasible set, it provides a competitive guarantee related to the ratio of the widths of the smallest hyperrectangle that contains $\mathcal{F}$ to the largest hyperrectangle contained in $\mathcal{F}$.

1.4 Notation and technical background

We use the notation $g_{1:t}$ as a shorthand for $\sum_{\tau=1}^{t} g_\tau$. Similarly we write $Q_{1:t}$ for a sum of matrices Q_t, and $f_{1:t}$ to denote the function $f_{1:t}(x) = \sum_{\tau=1}^{t} f_\tau(x)$. We write $x^\top y$ or $x \cdot y$ for the inner product between $x, y \in \mathbb{R}^n$. The ith entry in a vector x is denoted $x_i \in \mathbb{R}$; when we have a sequence of vectors $x_t \in \mathbb{R}^n$ indexed by time, the ith entry is $x_{t,i} \in \mathbb{R}$. We use $\partial f(x)$ to denote the set of subgradients of f evaluated at x.

Recall $A \in S^n_{++}$ means $\forall x \neq 0,\ x^\top A x > 0$. We use the generalized inequality $A \succ 0$ when $A \in S^n_{++}$, and similarly $A \prec B$ when $B - A \succ 0$, implying $x^\top A x < x^\top B x$. We define $A \preceq B$ analogously for symmetric positive semidefinite matrices S^n_+. For $B \in S^n_+$, we write $B^{1/2}$ for the square root of B, the unique $X \in S^n_+$ such that $XX = B$ (see, for example, Boyd and Vandenberghe [2004, A.5.2]). We also make use of the fact that any $A \in S^n_+$ can be factored as $A = PDP^\top$ where $P^\top P = I$ and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ where λ_i are the eigenvalues of A.

Following the arguments of Zinkevich [2003], for the remainder we restrict our attention to linear functions. Briefly, the convexity of f_t implies $f_t(x) \geq g_t^\top (x - x_t) + f_t(x_t)$, where $g_t \in \partial f(x_t)$. Because this inequality is tight for $x = x_t$, it follows that regret measured against the affine functions on the right hand side is an upper bound on true regret. Furthermore, regret is unchanged if we replace this affine function with the linear function $g_t^\top x$. Thus, so long as our algorithm only makes use of the subgradients g_t, we may assume without loss of generality that the loss functions are linear.

Taking into account this reduction and the functional form of the r_t, the update of FTPRL is

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathcal{F}} \left(\frac{1}{2} \sum_{\tau=1}^{t} (x - x_\tau)^\top Q_\tau (x - x_\tau) + g_{1:t} \cdot x \right). \tag{5}$$

2 Analysis of FTPRL

In this section, we prove the following bound on the regret of FTPRL for an arbitrary sequence of regularization matrices Q_t. In this section $\|\cdot\|$ always means the L_2 norm, $\|\cdot\|_2$.

Theorem 2. *Let $\mathcal{F} \subseteq \mathbb{R}^n$ be a closed, bounded convex set with $0 \in \mathcal{F}$. Let $Q_1 \in S^n_{++}$, and $Q_2, \ldots, Q_T \in S^n_+$. Define $r_t(x) = \frac{1}{2}\|Q_t^{\frac{1}{2}}(x - x_t)\|_2^2$, and $A_t = (Q_{1:t})^{\frac{1}{2}}$. Let f_t be a sequence of loss functions, with $g_t \in \partial f_t(x_t)$ a sub-gradient of f_t at x_t. Then, the FTPRL algorithm that that faces loss functions f, plays $x_1 = 0$, and uses the update of Equation* (5) *thereafter, has a regret bound*

$$\textit{Regret} \leq r_{1:T}(\mathring{x}) + \sum_{t=1}^{T} \|A_t^{-1} g_t\|^2$$

where $\mathring{x} = \arg\min_{x \in \mathcal{F}} f_{1:T}(x)$ is the post-hoc optimal feasible point.

To prove Theorem 2 we will make use of the following bound on the regret of FTRL, which holds for arbitrary (possibly non-convex) loss functions. This lemma can be proved along the lines of [Kalai and Vempala, 2005]; for completeness, a proof is included in Appendix A.

Lemma 1. *Let $r_1, r_2, \ldots, r_T$ be a sequence of non-negative functions. The regret of FTPRL (which plays x_t as defined by Equation* (2)*) is bounded by*

$$r_{1:T}(\mathring{x}) + \sum_{t=1}^{T} \left(f_t(x_t) - f_t(x_{t+1}) \right)$$

where $\mathring{x}$ is the post-hoc optimal feasible point.

Once Lemma 1 is established, to prove Theorem 2 it suffices to show that for all t,

$$f_t(x_t) - f_t(x_{t+1}) \le \|A_t^{-1} g_t\|^2. \tag{6}$$

To show this, we first establish an alternative characterization of our algorithm as solving an unconstrained optimization followed by a suitable projection onto the feasible set. Define the projection operator,

$$P_{\mathcal{F},A}(u) = \operatorname*{arg\,min}_{x\in\mathcal{F}} \|A(x-u)\|$$

We will show that the following is an equivalent formula for x_t:

$$\begin{aligned} u_{t+1} &= \operatorname*{arg\,min}_{u\in\mathbb{R}^n} \left(r_{1:t}(u) + g_{1:t}\cdot u\right) \\ x_{t+1} &= P_{\mathcal{F},A_t}\left(u_{t+1}\right). \end{aligned} \tag{7}$$

This characterization will be useful, because the unconstrained solutions depend only on the linear functions g_t, and the quadratic regularization, and hence are easy to manipulate in closed form.

To show this equivalence, first note that because $Q_t \in S_+^n$ is symmetric,

$$r_t(u) = \frac{1}{2}(u-x_t)^\top Q_t(u-x_t) = \frac{1}{2}u^\top Q_t u - x_t^\top Q_t u_t + \frac{1}{2}x_t^\top Q_t x_t.$$

Defining constants $q_t = Q_t x_t$ and $k_t = \frac{1}{2}x_t^\top Q_t x_t$, we can write

$$r_{1:t}(u) = \frac{1}{2}u^\top Q_{1:t}u - q_{1:t}u + k_{1:t}. \tag{8}$$

The equivalence is then a corollary of the following lemma, choosing $Q = Q_{1:t}$ and $h = g_{1:t} - q_{1:t}$ (note that the constant term $k_{1:t}$ does not influence the argmin).

Lemma 2. *Let $Q \in S_{++}^n$ and $h \in \mathbb{R}^n$, and consider the function*

$$f(x) = h^\top x + \frac{1}{2}x^\top Q x.$$

Let $\mathring{u} = \arg\min_{u\in\mathbb{R}^n} f(u)$. Then, letting $A = Q^{\frac{1}{2}}$, we have $P_{\mathcal{F},A}(\mathring{u}) = \arg\min_{x\in\mathcal{F}} f(x)$.

Proof. Note that $\nabla_u f(u) = h + Qu$, implying that $\mathring{u} = -Q^{-1}h$. Consider the function

$$f'(x) = \frac{1}{2}\|Q^{\frac{1}{2}}(x-\mathring{u})\|^2 = \frac{1}{2}(x-\mathring{u})^\top Q(x-\mathring{u}).$$

We have

$$\begin{aligned} f'(x) &= \frac{1}{2}\left(x^\top Qx - 2x^\top Q\mathring{u} + \mathring{u}^\top Q\mathring{u}\right) && \text{(because } Q \text{ is symmetric)} \\ &= \frac{1}{2}\left(x^\top Qx + 2x^\top Q(Q)^{-1}h + \mathring{u}^\top Q\mathring{u}\right) \\ &= \frac{1}{2}\left(x^\top Qx + 2x^\top h + \mathring{u}^\top Q\mathring{u}\right) \\ &= f(x) + \frac{1}{2}\mathring{u}^\top Q\mathring{u}\,. \end{aligned}$$

Because $\frac{1}{2}\mathring{u}^\top Q\mathring{u}$ is constant with respect to x, it follows that

$$\arg\min_{x\in\mathcal{F}} f(x) = \arg\min_{x\in\mathcal{F}} f'(x) = P_{\mathcal{F},A}(\mathring{u}),$$

where the last equality follows from the definition of the projection operator. $\square$

We now derive a closed-form solution to the unconstrained problem. It is easy to show $\nabla r_t(u) = Q_t u - Q_t x_t$, and so

$$\nabla r_{1:t}(u) = Q_{1:t}u - \sum_{\tau=1}^{t} Q_\tau x_\tau.$$

Because u_{t+1} is the optimum of the (strongly convex) unconstrained problem, and $r_{1:t}$ is differentiable, we must have $\nabla r_{1:t}(u_{t+1}) + g_{1:t} = 0$. Hence, we conclude $Q_{1:t}u_{t+1} - \sum_{\tau=1}^{t} Q_\tau x_\tau + g_{1:t} = 0$, or

$$u_{t+1} = Q_{1:t}^{-1}\left(\sum_{\tau=1}^{t} Q_\tau x_\tau - g_{1:t}\right). \tag{9}$$

This closed-form solution will let us bound the difference between u_t and u_{t+1} in terms of g_t. The next Lemma relates this distance to the difference between x_t and x_{t+1}, which determines our per round regret (Equation (6)). In particular, we show that the projection operator only makes u_t and u_{t+1} closer together, in terms of distance as measured by the norm $\|A_t \cdot\|$. We defer the proof to the end of the section.

Lemma 3. *Let $Q \in S^n_{++}$ with $A = Q^{\frac{1}{2}}$. Let $\mathcal{F}$ be a convex set, and let $u_1, u_2 \in \mathbb{R}^n$, with $x_1 = P_{\mathcal{F},A}(u_1)$ and $x_2 = P_{\mathcal{F},A}(u_2)$. Then,*

$$\|A(x_2 - x_1)\| \le \|A(u_1 - u_2)\|.$$

We now prove the following lemma, which will immediately yield the desired bound on $f_t(x_t) - f_t(x_{t+1})$.

Lemma 4. *Let $Q \in S^n_{++}$ with $A = Q^{\frac{1}{2}}$. Let $v, g \in \mathbb{R}^n$, and let $u_1 = -Q^{-1}v$ and $u_2 = -Q^{-1}(v+g)$. Then, letting $x_1 = P_{\mathcal{F},A}(u_1)$ and $x_2 = P_{\mathcal{F},A}(u_2)$,*

$$g^\top(x_1 - x_2) \le \|A^{-1}g\|^2.$$

Proof. The fact that $Q = A^\top A \succ 0$ implies that $\|A \cdot\|$ and $\|A^{-1}\cdot\|$ are dual norms (see for example [Boyd and Vandenberghe, 2004, Sec. 9.4.1, pg. 476]). Using this fact,

$$\begin{aligned}
g^\top(x_1 - x_2) &\le \|A^{-1}g\| \cdot \|A(x_1 - x_2)\| \\
&\le \|A^{-1}g\| \cdot \|A(u_1 - u_2)\| && \text{(Lemma 3)} \\
&= \|A^{-1}g\| \cdot \|A(Q^{-1}g)\| \\
&= \|A^{-1}g\| \cdot \|A(A^{-1}A^{-1})g)\| && \text{(Because } Q^{-1} = (AA)^{-1}\text{)} \\
&= \|A^{-1}g\| \cdot \|A^{-1}g\|.
\end{aligned}$$

$\square$

Proof of Theorem 2: First note that because $r_t(x_t) = 0$ and r_t is non-negative, $x_t = \arg\min_{x \in \mathcal{F}} r_t(x)$. For any functions f and g, if $x^* = \arg\min_{x\in\mathcal{F}} f(x)$ and $x^* = \arg\min_{x\in\mathcal{F}} g(x)$, then

$$x^* = \arg\min_{x\in\mathcal{F}} \left(f(x) + g(x)\right).$$

Thus we have

$$\begin{aligned} x_t &= \arg\min_{x\in\mathcal{F}} \left(g_{1:t-1}x + r_{1:t-1}(x)\right) \\ &= \arg\min_{x\in\mathcal{F}} \left(g_{1:t-1}x + r_{1:t}(x)\right) && \text{(Because } x_t = \arg\min_{x\in\mathcal{F}} r_t(x).\text{)} \\ &= \arg\min_{x\in\mathcal{F}} \left(hx + \frac{1}{2}x^\top Q_{1:t}x\right) \end{aligned}$$

where the last line follows from Equation (8), letting $h = g_{1:t-1} - q_{1:t} = g_{1:t-1} - \sum_{\tau=1}^{t} Q_\tau x_\tau$, and dropping the constant $k_{1:t}$. For x_{t+1}, we have directly from the definitions

$$x_{t+1} = \arg\min_{x\in\mathcal{F}} \left(g_{1:t}x + r_{1:t}(x)\right) = \arg\min_{x\in\mathcal{F}} \left((h + g_t)x + \frac{1}{2}x^\top Q_{1:t}x\right).$$

Thus, Lemma 2 implies $x_t = P_{\mathcal{F},A_t}(-(Q_{1:t})^{-1}h)$ and similarly $x_{t+1} = P_{\mathcal{F},A_t}(-(Q_{1:t})^{-1}(h + g_t))$. Thus, by Lemma 4, $g_t(x_t - x_{t+1}) \le \|A_t^{-1}g_t\|^2$. The theorem then follows from Lemma 1. □

Proof of Lemma 3: Define

$$B(x, u) = \frac{1}{2}\|A(x-u)\|^2 = \frac{1}{2}(x-u)^\top Q(x-u),$$

so we can write equivalently

$$x_1 = \arg\min_{x\in\mathcal{F}} B(x, u_1).$$

Then, note that $\nabla_x B(x, u_1) = Qx - Qu_1$, and so we must have $(Qx_1 - Qu_1)^\top(x_2 - x_1) \ge 0$; otherwise for δ sufficiently small the point $x_1 + \delta(x_2 - x_1)$ would belong to $\mathcal{F}$ (by convexity) and would be closer to u_1 than x_1 is. Similarly, we must have $(Qx_2 - Qu_2)^\top(x_1 - x_2) \ge 0$. Combining these, we have the following equivalent inequalities:

$$\begin{gathered} (Qx_1 - Qu_1)^\top(x_2 - x_1) - (Qx_2 - Qu_2)^\top(x_2 - x_1) \ge 0 \\ (x_1 - u_1)^\top Q(x_2 - x_1) - (x_2 - u_2)^\top Q(x_2 - x_1) \ge 0 \\ -(x_2 - x_1)^\top Q(x_2 - x_1) + (u_2 - u_1)^\top Q(x_2 - x_1) \ge 0 \\ (u_2 - u_1)^\top Q(x_2 - x_1) \ge (x_2 - x_1)Q(x_2 - x_1). \end{gathered}$$

Letting $\hat{u} = u_2 - u_1$, and $\hat{x} = x_2 - x_1$, we have $\hat{x}^\top Q\hat{x} \le \hat{u}^\top Q\hat{x}$. Since Q is positive semidefinite, we have $(\hat{u} - \hat{x})^\top Q(\hat{u} - \hat{x}) \ge 0$, or equivalently $\hat{u}^\top Q\hat{u} + \hat{x}^\top Q\hat{x} - 2\hat{x}^\top Q\hat{u} \ge 0$ (using the fact Q is also symmetric). Thus,

$$\hat{u}^\top Q\hat{u} \;\ge\; -\hat{x}^\top Q\hat{x} + 2\hat{x}^\top Q\hat{u} \;\ge\; -\hat{x}^\top Q\hat{x} + 2\hat{x}^\top Q\hat{x} \;=\; \hat{x}^\top Q\hat{x},$$

and so

$$\|A(u_2 - u_1)\|^2 = \hat{u}^\top Q\hat{u} \ge \hat{x}^\top Q\hat{x} = \|A(x_2 - x_1)\|^2.$$

□

3 Specific Adaptive Algorithms and Competitive Ratios

Before proceeding to the specific results, we establish several results that will be useful in the subsequent arguments. In order to prove that adaptive schemes for selecting Q_t have good competitive ratios for the bound optimization problem, we will need to compare the bounds obtained by the adaptive scheme to the optimal post-hoc bound of Equation (4). Suppose the sequence $Q_1, \ldots, Q_T$ is optimal for Equation (4), and consider the alternative sequence $Q'_1 = Q_{1:T}$ and $Q'_t = 0$ for $t > 1$. Using the fact that $Q_{1:t} \succeq Q_{1:t-1}$ implies $Q_{1:t}^{-1} \preceq Q_{1:t-1}^{-1}$, it is easy to show the alternative sequence also achieves the minimum. It follows that a sequence with $Q_1 = Q$ on the first round, and $Q_t = 0$ thereafter is always optimal. Hence, to solve for the post-hoc bound we can solve an optimization of the form

$$\inf_{Q \in \mathcal{Q}} \quad \left(\max_{\hat{y} \in \mathcal{F}_{\text{sym}}} \left(\frac{1}{2} \hat{y}^\top Q \hat{y} \right) + \sum_{t=1}^{T} g_t^\top Q^{-1} g_t \right). \tag{10}$$

The diameter of $\mathcal{F}$ is $D \equiv \max_{y,y' \in \mathcal{F}} \|y - y'\|_2$, and so for $\hat{y} \in \mathcal{F}_{\text{sym}}$, $\|\hat{y}\|_2 \le D$. When $\mathcal{F}$ is symmetric ($x \in \mathcal{F}$ implies $-x \in \mathcal{F}$), we have $y \in \mathcal{F}$ if and only if $2y \in \mathcal{F}_{\text{sym}}$, so (10) is equivalent to:

$$\inf_{Q \in \mathcal{Q}} \quad \left(\max_{y \in \mathcal{F}} \left(2 y^\top Q y \right) + \sum_{t=1}^{T} g_t^\top Q^{-1} g_t \right). \tag{11}$$

For simplicity of exposition, we assume $g_{1,i} > 0$ for all i, which ensures that only positive definite matrices can be optimal.[2] This assumption also ensures $Q_1 \in S_{++}^n$ for the adaptive schemes discussed below, as required by Theorem 2. This is without loss of generality, as we can always hallucinate an initial loss function with arbitrarily small components, and this changes regret by an arbitrarily small amount. We will also use the following Lemma from Auer and Gentile [2000]. For completeness, a proof is included in Appendix B.

Lemma 5. *For any non-negative real numbers $x_1, x_2, \ldots, x_n$,*

$$\sum_{i=1}^{n} \frac{x_i}{\sqrt{\sum_{j=1}^{i} x_j}} \le 2 \sqrt{\sum_{i=1}^{n} x_i} \, .$$

3.1 Adaptive coordinate-constant regularization

We derive bounds where Q_t is chosen from the set $\mathcal{Q}_{\text{const}}$, and show that this algorithm comes within a factor of $\sqrt{2}$ of using the best constant regularization strength λI. This algorithm achieves a bound of $\mathcal{O}(DM\sqrt{T})$ where D is the diameter of the feasible region and M is a bound on $\|g_t\|_2$, matching the best possible bounds in terms of these parameters [Abernethy et al., 2008]. We will prove a much stronger competitive guarantee for this algorithm in Theorem 6.

Corollary 1. *Suppose $\mathcal{F}$ has L_2 diameter D. Then, if we run FTPRL with diagonal matrices such that*

$$(Q_{1:t})_{ii} = \bar{\alpha}_t = \frac{2\sqrt{G_t}}{D}$$

[2] In the case where $\mathcal{F}$ has 0 width in some direction, the infimum will not be attained by a finite Q, but by a sequence that assigns 0 penalty (on the right-hand side) to the components of the gradient in the direction of 0 width, requiring some entries in Q to go to ∞.

where $G_t = \sum_{\tau=1}^{t}\sum_{i=1}^{n} g_{\tau,i}^2$*, then*

$$\textit{Regret} \leq 2D\sqrt{G_T}.$$

If $\|g_t\|_2 \leq M$*, then* $G_T \leq M^2 T$*, and this translates to a bound of* $\mathcal{O}(DM\sqrt{T})$*. When* $\mathcal{F} = \{x \mid \|x\|_2 \leq D/2\}$*, this bound is* $\sqrt{2}$*-competitive for the bound optimization problem over* $\mathcal{Q}_{const}$*.*

Proof. Let the diagonal entries of Q_t all be $\alpha_t = \bar{\alpha}_t - \bar{\alpha}_{t-1}$ (with $\bar{\alpha}_0 = 0$), so $\alpha_{1:t} = \bar{\alpha}_t$. Note $\alpha_t \geq 0$, and so this choice is feasible. We consider the left and right-hand terms of Equation (3) separately. For the left-hand term, letting $\hat{y}_t$ be an arbitrary sequence of points from $\mathcal{F}_{\text{sym}}$, and noting $\hat{y}_t^\top \hat{y}_t \leq \|\hat{y}_t\|_2 \cdot \|\hat{y}_t\|_2 \leq D^2$,

$$\frac{1}{2}\sum_{t=1}^{T} \hat{y}_t^\top Q_t \hat{y}_t = \frac{1}{2}\sum_{t=1}^{T} \hat{y}_t^\top \hat{y}_t \alpha_t \leq \frac{1}{2} D^2 \sum_{t=1}^{T} \alpha_t = \frac{1}{2} D^2 \bar{\alpha}_T = D\sqrt{G_T}.$$

For the right-hand term, we have

$$\sum_{t=1}^{T} g_t^\top Q_{1:t}^{-1} g_t = \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{g_{t,i}^2}{\alpha_{1:t}} = \sum_{t=1}^{T} \frac{D}{2} \frac{\sum_{i=1}^{n} g_{t,i}^2}{\sqrt{G_t}} \leq D\sqrt{G_T},$$

where the last inequality follows from Lemma 5.

In order to make a competitive guarantee, we must prove a lower bound on the post-hoc optimal bound function B_R, Equation (10). This is in contrast to the upper bound that we must show for the regret of the algorithm. When $\mathcal{F} = \{x \mid \|x\|_2 \leq D/2\}$, Equation (10) simplifies to exactly

$$\min_{\alpha \geq 0} \left(\frac{1}{2}\alpha D^2 + \frac{1}{\alpha} G_T \right) = D\sqrt{2G_T} \tag{12}$$

and so we conclude the adaptive algorithm is $\sqrt{2}$-competitive for the bound optimization problem. □

3.2 Adaptive diagonal regularization

In this section, we introduce and analyze FTPRL-Diag, a specialization of FTPRL that uses regularization matrices from $\mathcal{Q}_{\text{diag}}$. Let $D_i = \max_{x,x' \in \mathcal{F}} |x_i - x'_i|$, the width of $\mathcal{F}$ along the ith coordinate. We construct a bound on the regret of FTPRL-Diag in terms of these D_i. The D_i implicitly define a hyperrectangle that contains $\mathcal{F}$. When $\mathcal{F}$ is in fact such a hyperrectangle, our bound is $\sqrt{2}$-competitive with the best post-hoc optimal bound using matrices from $\mathcal{Q}_{\text{diag}}$.

Corollary 2. *Let* $\mathcal{F}$ *be a convex feasible set of width* D_i *in coordinate* i*. We can construct diagonal matrices* Q_t *such that the* i*th entry on the diagonal of* $Q_{1:t}$ *is given by:*

$$\bar{\lambda}_{t,i} = \frac{2}{D_i}\sqrt{\sum_{\tau=1}^{t} g_{\tau,i}^2}.$$

Then the regret of FTPRL satisfies

$$\textit{Regret} \leq 2\sum_{i=1}^{n} D_i \sqrt{\sum_{t=1}^{T} g_{t,i}^2}.$$

Algorithm 1 FTPRL-Diag

Input: feasible set $\mathcal{F} \subseteq \times_{i=1}^{n}[a_i, b_i]$
Initialize $x_1 = 0 \in \mathcal{F}$
$(\forall i),\ G_i = 0, q_i = 0, \lambda_{0,i} = 0, D_i = b_i - a_i$
for $t = 1$ **to** T **do**
  Play the point x_t, incur loss $f_t(x_t)$
  Let $g_t \in \partial f_t(x_t)$
  for $i = 1$ **to** n **do**
    $G_i = G_i + g_{t,i}^2$
    $\lambda_{t,i} = \frac{2}{D_i}\sqrt{G_i} - \lambda_{1:t-1,i}$
    $q_i = q_i + x_{t,i}\lambda_{t,i}$
    $u_{t+1,i} = (g_{1:t,i} - q_i)/\lambda_{1:t,i}$
  end for
  $A_t = \text{diag}(\sqrt{\lambda_{1:t,1}}, \ldots, \sqrt{\lambda_{1:t,n}})$
  $x_{t+1} = \text{Project}_{\mathcal{F},A_t}(u_{t+1})$
end for

Algorithm 2 FTPRL-Scale

Input: feasible set $\mathcal{F} \subseteq \{x \mid \|Ax\| \leq 1\}$,
 with $A \in S_{++}^n$
Let $\hat{\mathcal{F}} = \{x \mid \|x\| \leq 1\}$
Initialize $x_1 = 0$, $(\forall i)\ D_i = b_i - a_i$
for $t = 1$ **to** T **do**
  Play the point x_t, incur loss $f_t(x_t)$
  Let $g_t \in \partial f_t(x_t)$
  $\hat{g}_t = (A^{-1})^\top g_t$
  $\bar{\alpha} = \sqrt{\sum_{\tau=1}^{t}\sum_{i=1}^{n}\hat{g}_{\tau,i}^2}$
  $\alpha_t = \bar{\alpha} - \alpha_{1:t-1}$
  $q_t = \alpha_t x_t$
  $\hat{u}_{t+1} = (1/\bar{\alpha})(q_{1:t} - g_{1:t})$
  $A_t = (\bar{\alpha} I)^{\frac{1}{2}}$
  $\hat{x}_{t+1} = \text{Project}_{\hat{\mathcal{F}},A_t}(\hat{u}_{t+1})$
  $x_{t+1} = A^{-1}\hat{x}$
end for

When $\mathcal{F}$ is a hyperrectangle, then this algorithm is $\sqrt{2}$-competitive with the post-hoc optimal choice of Q_t from the $\mathcal{Q}_{diag}$. That is,

$$\textit{Regret} \leq \sqrt{2} \inf_{Q \in \mathcal{Q}_{diag}} \left(\max_{\hat{y} \in \mathcal{F}_{sym}} \left(\frac{1}{2} \hat{y}^\top Q \hat{y} \right) + \sum_{t=1}^{T} g_t^\top Q^{-1} g_t \right).$$

Proof. The construction of $Q_{1:t}$ in the theorem statement implies $(Q_t)_{ii} = \lambda_{t,i} \equiv \bar{\lambda}_{t,i} - \bar{\lambda}_{t-1,i}$. These entries are guaranteed to be non-negative as $\bar{\lambda}_{t,i}$ is a non-decreasing function of t.

We begin from Equation (3), letting $\hat{y}_t$ be an arbitrary sequence of points from $\mathcal{F}_{\text{sym}}$. For the left-hand term,

$$\frac{1}{2}\sum_{t=1}^{T} \hat{y}_t^\top Q_t \hat{y}_t = \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{n} \hat{y}_{t,i}^2 \lambda_{t,i} \leq \frac{1}{2}\sum_{i=1}^{n} D_i^2 \sum_{t=1}^{T} \lambda_{t,i} = \frac{1}{2}\sum_{i=1}^{n} D_i^2 \bar{\lambda}_{T,i} = \sum_{i=1}^{n} D_i \sqrt{\sum_{t=1}^{T} g_{t,i}^2}.$$

For the right-hand term, we have

$$\sum_{t=1}^{T} g_t^\top Q_{1:t}^{-1} g_t = \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{g_{t,i}^2}{\bar{\lambda}_{t,i}} = \sum_{i=1}^{n} \frac{D_i}{2} \sum_{t=1}^{T} \frac{g_{t,i}^2}{\sqrt{\sum_{\tau=1}^{t} g_{\tau,i}^2}} \leq \sum_{i=1}^{n} D_i \sqrt{\sum_{t=1}^{T} g_{t,i}^2},$$

where the last inequality follows from Lemma 5. Summing these bounds on the two terms of Equation (3) yields the stated bound on regret.

Now, we consider the case where the feasible set is exactly a hyperrectangle, that is, $\mathcal{F} = \{x \mid x_i \in [a_i, b_i]\}$ where $D_i = b_i - a_i$. Then, the optimization of Equation (10) decomposes on a per-coordinate basis, and in particular there exists a $\hat{y} \in \mathcal{F}_{\text{sym}}$ so that $\hat{y}_i^2 = D_i^2$ in each coordinate. Thus, for $Q =$

$\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, the bound function is exactly

$$\sum_{i=1}^{n} \frac{1}{2}\lambda_i D_i^2 + \frac{1}{\lambda_i}\sum_{t=1}^{T} g_{t,i}^2.$$

Choosing $\lambda_i = \frac{1}{D_i}\sqrt{2\sum_{t=1}^{T} g_{t,i}^2}$ minimizes this quantity, producing a post-hoc bound of

$$\sqrt{2}\sum_{i=1}^{n} D_i \sqrt{\sum_{t=1}^{T} g_{t,i}^2},$$

verifying that the adaptive scheme is $\sqrt{2}$-competitive with matrices from $\mathcal{Q}_{\mathrm{diag}}$.

□

The regret guarantees of the FTPRL-Diag algorithm hold on arbitrary feasible sets, but the competitive guarantee only applies for hyperrectangles. We now extend this result, showing that a competitive guarantee can be made based on how well the feasible set is approximated by hyperrectangles.

Theorem 3. *Let $\mathcal{F}$ be an arbitrary feasible set, bounded by a hyperrectangle H^{out} of width W_i in coordinate i; further, let H^{in} be a hyperrectangle contained by $\mathcal{F}$, of width $w_i > 0$ in coordinate i. That is, $H^{in} \subseteq \mathcal{F} \subseteq H^{out}$. Let $\beta = \max_i \frac{W_i}{w_i}$. Then, the FTPRL-Diag is $\sqrt{2}\beta$-competitive with $\mathcal{Q}_{diag}$ on $\mathcal{F}$.*

Proof. By Corollary 2, the adaptive algorithm achieves regret bounded by $2\sum_{i=1}^{n} W_i\sqrt{\sum_{t=1}^{T} g_{t,i}^2}$. We now consider the best post-hoc bound achievable with diagonal matrices on $\mathcal{F}$. Considering Equation (10), it is clear that for any Q,

$$\max_{y \in \mathcal{F}_{\mathrm{sym}}} \frac{1}{2}y^\top Q y + \sum_{t=1}^{T} g_t^\top Q^{-1} g_t \geq \max_{y \in H^{\mathrm{in}}_{\mathrm{sym}}} \frac{1}{2}y^\top Q y + \sum_{t=1}^{T} g_t^\top Q^{-1} g_t,$$

since the feasible set for the maximization ($\mathcal{F}_{\mathrm{sym}}$) is larger on the left-hand side. But, on the right-hand side we have the post-hoc bound for diagonal regularization on a hyperrectangle, which we computed in the previous section to be $\sqrt{2}\sum_{i=1}^{n} w_i\sqrt{\sum_{t=1}^{T} g_{t,i}^2}$. Because $w_i \geq \frac{W_i}{\beta}$ by assumption, this is lower bounded by $\frac{\sqrt{2}}{\beta}\sum_{i=1}^{n} W_i\sqrt{\sum_{t=1}^{T} g_{t,i}^2}$, which proves the theorem. □

Having had success with L_∞, we now consider the potential benefits of diagonal adaptation for other L_p-balls.

3.3 A post-hoc bound for diagonal regularization on L_p balls

Suppose the feasible set F is an unit L_p-ball, that is $F = \{x \mid \|x\|_p \leq 1\}$. We consider the post-hoc bound optimization problem of Equation (11) with $\mathcal{Q} = \mathcal{Q}_{\mathrm{diag}}$. Our results are summarized in the following theorem.

Theorem 4. *For $p > 2$, the optimal regularization matrix for B_R in $\mathcal{Q}_{diag}$ is not coordinate-constant (i.e., not contained in $\mathcal{Q}_{const}$), except in the degenerate case where $G_i \equiv \sum_{t=1}^{T} g_{t,i}^2$ is the same for all i. However, for $p \leq 2$, the optimal regularization matrix in $\mathcal{Q}_{diag}$ always belongs to $\mathcal{Q}_{const}$.*

Proof. Since $\mathcal{F}$ is symmetric, the optimal post-hoc choice will be in the form of Equation (11). Letting $Q = \text{diag}(\lambda_1, \ldots, \lambda_n)$, we can re-write this optimization problem as

$$\max_{y:\|y\|_p \le 1} \left(2 \sum_{i=1}^{n} y_i^2 \lambda_i \right) + \sum_{i=1}^{n} \frac{G_i}{\lambda_i} . \tag{13}$$

To determine the optimal λ vector, we first derive a closed form for the solution to the maximization problem on the left hand side, assuming $p \ge 2$ (we handle the case $p < 2$ separately below). First note that the inequality $\|y\|_p \le 1$ is equivalent to $\sum_{i=1}^n |y_i|^p \le 1$. Making the change of variable $z_i = y_i^2$, this is equivalent to $\sum_{i=1}^n z_i^{\frac{p}{2}} \le 1$, which is equivalent to $\|z\|_{\frac{p}{2}} \le 1$ (the assumption $p \ge 2$ ensures that $\|\cdot\|_{\frac{p}{2}}$ is a norm). Hence, the left-hand side optimization reduces to

$$\max_{z:\|z\|_{\frac{p}{2}} \le 1} 2 \sum_{i=1}^{n} z_i \lambda_i = 2\|\lambda\|_q ,$$

where $q = \frac{p}{p-2}$, so that $\|\cdot\|_{\frac{p}{2}}$ and $\|\cdot\|_q$ are dual norms (allowing $q = \infty$ for $p = 2$). Thus, for $p \ge 2$, the above bound simplifies to

$$B(\lambda) = 2\|\lambda\|_q + \sum_{i=1}^{n} \frac{G_i}{\lambda_i}. \tag{14}$$

First suppose $p > 2$, so that q is finite. Then, taking the gradient of $B(\lambda)$,

$$\nabla B(\lambda)_i = \frac{2}{q} \left(\sum_{i=1}^{n} \lambda_i^q \right)^{\frac{1}{q}-1} \cdot q\lambda_i^{q-1} - \frac{G_i}{\lambda_i^2} = 2 \left(\frac{\lambda_i}{\|\lambda\|_q} \right)^{q-1} - \frac{G_i}{\lambda_i^2},$$

using $\frac{1}{q} - 1 = -\frac{1}{q}(q-1)$. If we make all the λ_i's equal (say, to λ_1), then for the left-hand side we get

$$\left(\frac{\lambda_i}{\|\lambda\|_q} \right)^{q-1} = \left(\frac{\lambda_1}{(n\lambda_1^q)^{\frac{1}{q}}} \right)^{q-1} = \left(\frac{1}{n^{\frac{1}{q}}} \right)^{q-1} = n^{\frac{1}{q}-1} .$$

Thus the i^{th} component of the gradient is $2n^{\frac{1}{q}-1} - \frac{G_i}{\lambda_1^2}$, and so if not all the G_i's are equal, some component of the gradient is non-zero. Because $B(\lambda)$ is differentiable and the $\lambda_i \ge 0$ constraints cannot be tight (recall $g_1 > 0$), this implies a constant λ_i cannot be optimal, hence the optimal regularization matrix is not in $\mathcal{Q}_{\text{const}}$.

For $p \in [1, 2]$, we show that the solution to Equation (13) is

$$B_\infty(\lambda) \equiv 2\|\lambda\|_\infty + \sum_{i=1}^{n} \frac{G_i}{\lambda_i}. \tag{15}$$

For $p = 2$ this follows immediately from Equation (14), because when $p = 2$ we have $q = \infty$. For $p \in [1, 2)$, the solution to Equation (13) is at least $B_\infty(\lambda)$, because we can always set $y_i = 1$ for whatever λ_i is largest and set $y_j = 0$ for $j \ne i$. If $p < 2$ then the feasible set $\mathcal{F}$ is a subset of the unit L_2 ball, so the solution to Equation (13) is upper bounded by the solution when $p = 2$, namely $B_\infty(\lambda)$. It follows that the solution is exactly $B_\infty(\lambda)$. Because the left-hand term of $B_\infty(\lambda)$ only penalizes for the largest λ_i, and on the right-hand we would like all λ_i as large as possible, a solution of the form $\lambda_1 = \lambda_2 = \cdots = \lambda_n$ must be optimal. □

3.4 Full matrix regularization on hyperspheres and hyperellipsoids

In this section, we develop an algorithm for feasible sets $\mathcal{F} \subseteq \{x \mid \|Ax\|_p \leq 1\}$, where $p \in [1,2]$ and $A \in S^n_{++}$. When $\mathcal{F} = \{x \mid \|Ax\|_2 \leq 1\}$, this algorithm, FTPRL-Scale, is $\sqrt{2}$-competitive with arbitrary $Q \in S^n_+$. For $\mathcal{F} = \{x \mid \|Ax\|_p \leq 1\}$ with $p \in [1,2)$ it is $\sqrt{2}$-competitive with $\mathcal{Q}_{\text{diag}}$.

First, we show that rather than designing adaptive schemes specifically for linear transformations of norm balls, it is sufficient (from the point of view of analyzing FTPRL) to consider unit norm balls if suitable pre-processing is applied. In the same fashion that pre-conditioning may speed batch subgradient descent algorithms, we show this approach can produce significantly improved regret bounds when A is poorly conditioned (i.e., the ratio of the largest to smallest eigenvalue is large).

Theorem 5. *Fix an arbitrary norm $\|\cdot\|$, and define an online linear optimization problem $\mathcal{I} = (\mathcal{F}, (g_1, \ldots, g_T))$ where $\mathcal{F} = \{x \mid \|Ax\| \leq 1\}$ with $A \in S^n_{++}$. We define the related instance $\hat{\mathcal{I}} = (\hat{\mathcal{F}}, (\hat{g}_1, \ldots, \hat{g}_T))$, where $\hat{\mathcal{F}} = \{\hat{x} \mid \|\hat{x}\| \leq 1\}$ and $\hat{g}_t = A^{-1} g_t$. Then:*

- *If we run any algorithm dependent only on subgradients on $\hat{\mathcal{I}}$, and it plays $\hat{x}_1, \ldots, \hat{x}_T$, then by playing the corresponding points $x_t = A^{-1}\hat{x}_t$ on $\mathcal{I}$ we achieve identical loss and regret.*
- *The post-hoc optimal bound over arbitrary $Q \in S^n_{++}$ is identical for these two instances.*

Proof. First, we note that for any function h where $\min_{x:\|Ax\|\leq 1} h(x)$ exists,

$$\min_{x:\|Ax\|\leq 1} h(x) = \min_{\hat{x}:\|\hat{x}\|\leq 1} h(A^{-1}\hat{x}), \tag{16}$$

using the change of variable $\hat{x} = Ax$. For the first claim, note that $\hat{g}_t^\top = g_t^\top A^{-1}$, and so for all t, $\hat{g}_t^\top \hat{x}_t = g_t^\top A^{-1} A x_t = g_t^\top x_t$, implying the losses suffered on $\hat{\mathcal{I}}$ and $\mathcal{I}$ are identical. Applying Equation (16), we have

$$\min_{x:\|Ax\|\leq 1} g_{1:t}^\top x = \min_{\hat{x}:\|\hat{x}\|\leq 1} g_{1:t}^\top A^{-1}\hat{x} = \min_{\hat{x}:\|\hat{x}\|\leq 1} \hat{g}_{1:t}^\top \hat{x},$$

implying the post-hoc optimal feasible points for the two instances also incur identical loss. Combining these two facts proves the first claim. For the second claim, it is sufficient to show for any $Q \in S^n_{++}$ applied to the post-hoc bound for problem $\mathcal{I}$, there exists a $\hat{Q} \in S^n_{++}$ that achieves the same bound for $\hat{\mathcal{I}}$ (and vice versa). Consider such a Q for $\mathcal{I}$. Then, again applying Equation (16), we have

$$\max_{y:\|Ay\|_p\leq 1} \left(2y^\top Q y\right) + \sum_{t=1}^{T} g_t^\top Q^{-1} g_t = \max_{\hat{y}:\|\hat{y}\|\leq 1} \left(2\hat{y}^\top A^{-1} Q A^{-1} \hat{y}\right) + \sum_{t=1}^{T} \hat{g}_t^\top A Q^{-1} A \hat{g}_t.$$

The left-hand side is the value of the post-hoc bound on $\mathcal{I}$ from Equation (11). Noting that $(A^{-1}QA^{-1})^{-1} = AQ^{-1}A$, the right-hand side is the value of the post hoc bound for $\hat{\mathcal{I}}$ using $\hat{Q} = A^{-1}QA^{-1}$. The fact A^{-1} and Q are in S^n_{++} guarantees $\hat{Q} \in S^n_{++}$ as well, and the theorem follows. □

We can now define the adaptive algorithm FTPRL-Scale: given a $\mathcal{F} \subseteq \{x \mid \|Ax\|_p \leq 1\}$, it uses the transformation suggested by Theorem 5, applying the coordinate-constant algorithm of Corollary 1 to the transformed instance, and playing the corresponding point mapped back into $\mathcal{F}$.[3] Pseudocode is given as Algorithm 2.

[3] By a slightly more cumbersome argument, it is possible to show that instead of applying this transformation, FTPRL can be run directly on $\mathcal{F}$ using appropriately transformed Q_t matrices.

Theorem 6. *The diagonal-constant algorithm analyzed in Corollary 1 is $\sqrt{2}$-competitive with S^n_+ when $\mathcal{F} = \{x \mid \|x\|_p \leq 1\}$ for $p = 2$, and $\sqrt{2}$-competitive against $\mathcal{Q}_{diag}$ when $p \in [1, 2)$. Furthermore, when $\mathcal{F} = \{x \mid \|Ax\|_p \leq 1\}$ with $A \in S^n_{++}$, the FTPRL-Scale algorithm (Algorithm 2) achieves these same competitive guarantees. In particular, when $\mathcal{F} = \{x \mid \|x\|_2 \leq 1\}$, we have*

$$\textit{Regret} \leq \sqrt{2} \inf_{Q \in S^n_+} \left(\max_{y \in \mathcal{F}} \left(2y^\top Q y\right) + \sum_{t=1}^{T} g_t^\top Q^{-1} g_t \right).$$

Proof. The results for $\mathcal{Q}_{\text{diag}}$ with $p \in [1, 2)$ follow from Theorems 4 and 5 and Corollary 1. We now consider the case $p = 2$. Consider a $Q \in S^n_{++}$ for Equation (11) (recall only a $Q \in S^n_{++}$ could be optimal since $g_1 > 0$). We can write $Q = PDP^\top$ where $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix of positive eigenvalues and $PP^\top = I$. It is then easy to verify $Q^{-1} = PD^{-1}P^\top$.

When $p = 2$ and $\mathcal{F} = \{x \mid \|x\|_p \leq 1\}$, Equation (15) is tight, and so the post-hoc bound for Q is

$$2 \max_i (\lambda_i) + \sum_{t=1}^{T} g_t^\top (PD^{-1}P^\top) g_t.$$

Let $z_t = P^\top g_t$, so each right-hand term is $\sum_{i=1}^n \frac{z_{t,i}^2}{\lambda_i}$. It is clear this quantity is minimized when each λ_i is chosen as large as possible, while on the left-hand side we are only penalized for the largest eigenvalue of Q (the largest λ_i). Thus, a solution where $D = \alpha I$ for $\alpha > 0$ is optimal. Plugging into the bound, we have

$$B(\alpha) = 2\alpha + \sum_{t=1}^{T} g_t^\top \left(P \left(\frac{1}{\alpha} I \right) P^\top \right) g_t = 2\alpha + \frac{1}{\alpha} \sum_{t=1}^{T} g_t^\top g_t = 2\alpha + \frac{G_T}{\alpha}$$

where we have used the fact that $PP^\top = I$. Setting $\alpha = \sqrt{G_T/2}$ produces a minimal post-hoc bound of $2\sqrt{2G_T}$. The diameter D is 2, so the coordinate-constant algorithm has regret bound $4\sqrt{G_T}$ (Corollary 1), proving the first claim of the theorem for $p = 2$. The second claim follows from Theorem 5. $\square$

Suppose we have a problem instance where $\mathcal{F} = \{x \mid \|Ax\|_2 \leq 1\}$ where $A = \text{diag}(1/a_1, \ldots, 1/a_n)$ with $a_i > 0$. To demonstrate the advantage offered by this transformation, we can compare the regret bound obtained by directly applying the algorithm of Corollary 1 to that of the FTPRL-Scale algorithm. Assume WLOG that $\max_i a_i = 1$, implying the diameter of $\mathcal{F}$ is 2. Let $g_1, \ldots, g_T$ be the loss functions for this instance. Recalling $G_i = \sum_{t=1}^T g_{t,i}^2$, applying Corollary 1 directly to this problem gives

$$\text{Regret} \leq 4\sqrt{\sum_{i=1}^{n} G_i}. \tag{17}$$

This is the same as the bound obtained by online subgradient descent and related algorithms as well.

We now consider FTPRL-Scale, which uses the transformation of Theorem 5. Noting $D = 2$ for the hypersphere and applying Corollary 1 to the transformed problem gives an adaptive scheme with

$$\text{Regret} \leq 4\sqrt{\sum_{i=1}^{n} \sum_{t=1}^{T} \hat{g}_{t,i}^2} = 4\sqrt{\sum_{i=1}^{n} a_i^2 \sum_{t=1}^{T} g_{t,i}^2} = 4\sqrt{\sum_{i=1}^{n} a_i^2 G_i}.$$

This bound is never worse than the bound of Equation (17), and can be arbitrarily better when many of the a_i are much smaller than 1.

4 Related work

In the batch convex optimization setting, it is well known that convergence rates can often be dramatically improved through the use of preconditioning, accomplished by an appropriate change of coordinates taking into account both the shape of the objective function and the feasible region [Boyd and Vandenberghe, 2004]. To our knowledge, this is the first work that extends these concepts (necessarily in a quite different form) to the problem of online convex optimization, where they can provide a powerful tool for improving regret (the online analogue of convergence rates).

Perhaps the closest algorithms in spirit to our diagonal adaptation algorithm are confidence-weighted linear classification [Drezde et al., 2008] and AROW [Crammer et al., 2009], in that they make different-sized adjustments for different coordinates. Unlike our algorithm, these algorithms apply only to classification problems and not to general online convex optimization, and the guarantees are in the form of mistake bounds rather than regret bounds.

FTPRL is similar to the lazily-projected gradient descent algorithm of [Zinkevich, 2004, Sec. 5.2.3], but with a critical difference: the latter effectively centers regularization outside of the current feasible region (at u_t rather than x_t). As a consequence, lazily-projected gradient descent only attains low regret via a re-starting mechanism or a constant learning rate (chosen with knowledge of T). It is our technique of always centering additional regularization inside the feasible set that allows us to make guarantees for adaptively-chosen regularization.

Most recent state-of-the-art algorithms for online learning are in fact general algorithms for online convex optimization applied to learning problems. Many of these algorithms can be thought of as (significant) extensions of online subgradient descent, including [Duchi and Singer, 2009, Do et al., 2009, Shalev-Shwartz et al., 2007]. Apart from the very general work of [Kalai and Vempala, 2005], few general follow-the-regularized-leader algorithms have been analyzed, with the notable exception of the recent work of Xiao [2009].

The notion of proving competitive ratios for regret bounds that are functions of regularization parameters is not unique to this paper. Bartlett et al. [2008] and Do et al. [2009] proved guarantees of this form, but for a different algorithm and class of regularization parameters.

In concurrent work [Streeter and McMahan, 2010], the authors proved bounds similar to those of Corollary 2 for online gradient descent with per-coordinate learning rates. These results were significantly less general that the ones presented here, and in particular were restricted to the case where $\mathcal{F}$ was exactly a hyperrectangle. The FTPRL algorithm and bounds proved in this paper hold for arbitrary feasible sets, with the bound depending on the shape of the feasible set as well as the width along each dimension. Some results similar to those in this work were developed concurrently by Duchi et al. [2010], though for a different algorithm and using different analysis techniques.

5 Conclusions

In this work, we analyzed a new algorithm for online convex optimization, which takes ideas both from online subgradient descent as well as follow-the-regularized-leader. In our analysis of this algorithm, we show that the learning rates that occur in standard bounds can be replaced by positive semidefinite matrices. The extra degrees of freedom offered by these generalized learning rates provide the key to proving better regret bounds. We characterized the types of feasible sets where this technique can lead to significant gains, and showed that while it does not help on the hypersphere, it can have dramatic impact when the feasible set is a hyperrectangle.

The diagonal adaptation algorithm we introduced can be viewed as an incremental optimization of the formula for the final bound on regret. In the case where the feasible set really is a hyperrectangle, this allows us to guarantee our final regret bound is within a small constant factor of the best bound that could have been obtained had the full problem been known in advance. The diagonal adaptation algorithm is efficient, and exploits exactly the kind of structure that is typical in large-scale real-world learning problems such as click-through rate prediction and text classification.

Our work leaves open a number of interesting directions for future work, in particular the development of competitive algorithms for arbitrary feasible sets (without resorting to bounding norm-balls), and the development of algorithms that optimize over richer families of regularization functions.

References

Jacob Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In *COLT*, 2008.

Peter Auer and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. In *COLT*, 2000.

Peter L. Bartlett, Elad Hazan, and Alexander Rakhlin. Adaptive online gradient descent. In *NIPS*, 2008.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *NIPS*, volume 20, 2008.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. New York, NY, USA, 2004. ISBN 0521833787.

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

Koby Crammer, Alex Kulesza, and Mark Drezde. Adaptive regularization of weight vectors. In *NIPS*, 2009.

Chuong B. Do, Quoc V. Le, and Chuan-Sheng Foo. Proximal regularization for online and batch learning. In *ICML*, 2009.

Mark Drezde, Koby Crammer, and Fernando Pereira. Confidence-weighted linear classification. In *ICML*, 2008.

John Duchi and Yoram Singer. Efficient learning using forward-backward splitting. In *NIPS*, 2009.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. Manuscript, 2010.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and Systems Sciences*, 71(3), 2005. ISSN 0022-0000.

Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *ICML*, 2007.

Matthew Streeter and H. Brendan McMahan. Less regret via online conditioning. http://arxiv.org/abs/1002.4862, 2010.

Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *NIPS*, 2009.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *ICML*, 2004.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, 2003.

Martin Zinkevich. *Theoretical guarantees for algorithms in multi-agent settings*. PhD thesis, Pittsburgh, PA, USA, 2004.

A A Proof of the FTRL Bound

In this section we provide a proof of Lemma 1. The high-level structure of our proof follows Kalai and Vempala's analysis of the *follow the perturbed leader* algorithm, in that we prove bounds on three quantities:

1. the regret of a hypothetical *be the leader* algorithm (BTL), which on round t plays
$$x_t^* = \arg\min_{x \in \mathcal{F}} f_{1:t}(x),$$
2. the difference between the regret of BTL and that of the *be the regularized leader* algorithm (BTRL), which plays
$$\hat{x}_t = \arg\min_{x \in \mathcal{F}} \left(r_{1:t}(x) + f_{1:t}(x)\right) = x_{t+1}, \tag{18}$$
and
3. the difference between the regret of BTRL and that of FTRL.

As shown in [Kalai and Vempala, 2005], the BTL algorithm has regret ≤ 0 even without any restrictions on the loss functions or the feasible set. The proof is a straightforward induction, which we reproduce here for completeness.

Lemma 6 ([Kalai and Vempala, 2005])**.** *Let $f_1, f_2, \ldots, f_T$ be an arbitrary sequence of functions, and let $\mathcal{F}$ be an arbitrary set. Define $x_t^* \equiv \arg\min_{x \in \mathcal{F}} \sum_{\tau=1}^{t} f_\tau(x)$. Then*
$$\sum_{t=1}^{T} f_t(x_t^*) \leq \sum_{t=1}^{T} f_t(x_T^*) \, .$$

Proof. We prove this by induction on T. For $T = 1$ it is trivially true. Suppose that it holds for $T - 1$. Then
$$\begin{aligned}
\sum_{t=1}^{T} f_t(x_t^*) &= f_T(x_T^*) + \sum_{t=1}^{T-1} f_t(x_t^*) \\
&\leq f_T(x_T^*) + \sum_{t=1}^{T-1} f_t(x_{T-1}^*) && \text{(Induction hypothesis)} \\
&\leq f_T(x_T^*) + \sum_{t=1}^{T-1} f_t(x_T^*) && \text{(Definition of } x_{T-1}^*\text{)} \\
&= \sum_{t=1}^{T} f_t(x_T^*) \, .
\end{aligned}$$
□

We next prove a bound on the regret of BTRL.

Lemma 7. *Let $r_1, r_2, \ldots, r_T$ be a sequence of non-negative functions. Then BTRL, which on round t plays $\hat{x}_t$ as defined by equation* (18)*, has regret at most $r_{1:T}(\mathring{x})$ where $\mathring{x}$ is the post-hoc optimal solution.*

Proof. Define $f'_t(x) = f_t(x) + r_t(x)$. Observe that $\hat{x}_t = \arg\min_{x \in \mathcal{F}} f'_{1:t}(x)$. Thus, by Lemma 6, we have

$$\sum_{t=1}^{T} f'_t(\hat{x}_t) \leq \min_{x \in \mathcal{F}} f'_{1:T}(x) \leq f'_{1:T}(\mathring{x})$$

or equivalently,

$$\sum_{t=1}^{T} f_t(\hat{x}_t) + r_t(\hat{x}_t) \leq r_{1:T}(\mathring{x}) + f_{1:T}(\mathring{x}).$$

Dropping the non-negative $r_t(\hat{x}_t)$ terms on the left hand side proves the lemma. □

By definition, the total loss of FTRL (which plays x_t) exceeds that of BTRL (which plays $\hat{x}_t = x_{t+1}$) by $\sum_{t=1}^{T} f_t(x_t) - f_t(x_{t+1})$. Putting these facts together proves Lemma 1.

B Proof of Lemma 5

Proof. The lemma is clearly true for $n = 1$. Fix some n, and assume the lemma holds for $n - 1$. Thus,

$$\begin{aligned} \sum_{i=1}^{n} \frac{x_i}{\sqrt{\sum_{j=1}^{i} x_j}} &\leq 2\sqrt{\sum_{i=1}^{n-1} x_i} + \frac{x_n}{\sqrt{\sum_{i=1}^{n} x_i}} \\ &= 2\sqrt{Z - x} + \frac{x}{\sqrt{Z}} \end{aligned}$$

where we define $Z = \sum_{i=1}^{n} x_i$ and $x = x_n$. The derivative of the right hand side with respect to x is $\frac{-1}{\sqrt{Z-x}} + \frac{1}{\sqrt{Z}}$, which is negative for $x > 0$. Thus, subject to the constraint $x \geq 0$, the right hand side is maximized at $x = 0$, and is therefore at most $2\sqrt{Z}$. □