Follow-the-Regularized-Leader and Mirror Descent: Equivalence Theorems and Implicit Updates

H. Brendan McMahan
mcmahan@google.com

October 19, 2018

Abstract

We study two families of online convex optimization algorithms: mirror descent and follow-the-regularized-leader. We prove that many mirror descent algorithms (such as online gradient descent) are actually instances of a more general follow-the-leader algorithm which uses proximal regularization (additional strong convexity centered at the current feasible point). Further, under certain step-size assumptions, other mirror-descent algorithms are equivalent to follow-the-leader algorithms with origin-centered regularization (such as dual averaging). Building on these observations, we provide a general analysis for the case of L2 regularization that accommodates general implicit updates as well as composite objectives. This analysis tightens (by a constant factor) and generalizes earlier analysis of the follow-the-proximally-regularized-leader algorithm.

1 Introduction

We consider the problem of online convex optimization, and in particular its application to online learning. On each round $t = 1, \ldots, T$, we must pick a point $x_t \in \mathbb{R}^n$. 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 with respect to a comparator point $\mathring{x}$ is

$$\text{Regret} \equiv \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(\mathring{x}).$$

Often, we will wish to prove that regret is low against the point $\mathring{x} = \arg\min_{x \in \mathcal{F}} f_t(x)$ for some convex feasible set $\mathcal{F}$.

The first part of the paper, Section 2, considers general formulations of mirror descent and follow-the-regularized-leader, and proves theorems relating the two. While we prove more general results, Corollaries 4 and 6 state the most intuitive and practically important cases.

In the second part of the paper, Section 3, we analyze a generalization of the FTPRL algorithm of [McMahan and Streeter, 2010]. Our generalization maintains the good properties of the original algorithm while adding support for arbitrary non-smooth regularization and implicit updates. Thanks to the results Section 2, this analysis also provides bounds for mirror descent algorithms, for example, a generalization of the composite-objective mirror descent (COMID) algorithm with quadratic regularization [Duchi et al., 2010].

Implicit and Composite Updates Much of this paper concerns follow-the-regularized-leader algorithms that perform implicit and composite updates. We briefly formalize that approach here. The standard subgradient FTRL algorithm uses the update

$$x_{t+1} = \arg\min_x \left(\sum_{s=1}^{t} \nabla f_s(x_s) \right) \cdot x + R_{1:t}(x),$$

where we write $R_{1:t}(x)$ for $\sum_{s=1}^{t} R_s(x)$. In this update, each previous (potentially non-linear) loss function f_s is approximated by the gradient at x_s (when f_s is not differentiable, we can use a subgradient at x_s in place of the gradient). The functions R_t are incremental regularization added on each round; standard learning rates are encoded by choosing the R_t such that $\eta_t R = R_{1:t}$ (see Section 2 for details).

Implicit update rules are more often defined for mirror descent algorithms, but we can define an analogous update for FTRL:

$$x_{t+1} = \arg\min_x \left(\sum_{s=1}^{t-1} \nabla f_s(x_{s+1}) \right) \cdot x + f_t(x) + R_{1:t}(x).$$

This update replaces the subgradient approximation of f_t with the possibly non-linear f_t, and is similar to the online coordinate-dual-ascent algorithm briefly mentioned in [Shalev-Shwartz and Kakade, 2008]. In general, computing the implicit update might require solving an arbitrary convex optimization problem (hence, the name implicit), however, in many useful applications it can be computed in closed form or by optimizing a one-dimensional problem. Thus, such updates can be preferable to running follow-the-regularized-leader on the full history of f_t's because the updates are generally much easier to compute. See [Kulis and Bartlett, 2010] for some additional discussion.

When f_t is not differentiable, we use the update

$$x_{t+1} = \arg\min_x g'_{1:t-1} \cdot x + f_t(x) + R_{1:t}(x), \tag{1}$$

where g'_t is a subgradient of f_t at x_{t+1} (that is, $g'_t \in \partial f_t(x_{t+1})$) such that

$$g'_{1:t-1} + g'_t + \nabla R_{1:t}(x_{t+1}) = 0.$$

The existence of such a subgradient is proved below, in Theorem 2.

In many applications, we have a fixed convex function Ψ that we also wish to include in the optimization, for example $\Psi(x) = \|x\|_1$ (L_1-regularization to induce sparsity) or the indicator function on a feasible set $\mathcal{F}$. While it is possible to approximate this function via subgradients as well, when computationally feasible it is often better to handle Ψ directly. For example, in the case where $\Psi(x) = \|x\|_1$, subgradient approximations will in general not lead to sparse solutions. In this case, closed-form updates for optimizations including Ψ are often possible, and produce much better sparsity [Xiao, 2009, Duchi and Singer, 2009]. We can include such a term directly in FTRL, giving the composite objective update

$$x_{t+1} = \arg\min_x \left(\sum_{s=1}^{t} \nabla f_s(x_s) \right) \cdot x + \Psi(x) + R_{1:t}(x). \tag{2}$$

Finally, we can combine these ideas to define Implicit-Composite FTRL. In the general case where f_t is not differentiable, we have the update

$$x_{t+1} = \arg\min_x g'_{1:t-1} \cdot x + f_t(x) + \Psi(x) + R_{1:t}(x), \tag{3}$$

where $g'_t \in \partial f_t(x_{t+1})$ such that $\exists \phi_t \in \partial \Psi(x_{t+1})$ such that

$$g'_{1:t-1} + g'_t + \phi_t + \nabla R_{1:t}(x_{t+1}) = 0.$$

The existence of such a subgradient again follows from Theorem 2.

Feasible Sets In some applications, we may be restricted to only play points from a restricted feasible set $\mathcal{F} \subseteq \mathbb{R}^n$, for example, the set of (fractional) paths between two nodes in a graph. With composite updates, Equations (2) and (3), this is accomplished for free by choosing Ψ to be the indicator function on $\mathcal{F}$,

$$\Psi_{\mathcal{F}}(x) = \begin{cases} 0 & x \in \mathcal{F} \\ \infty & x \notin \mathcal{F}. \end{cases}$$

It is straightforward to verify that

$$\arg\min_{x \in \mathbb{R}^n} g_{1:t} \cdot x + R_{1:t}(x) + \Psi_{\mathcal{F}}(x)$$

is equivalent to

$$\arg\min_{x \in \mathcal{F}} g_{1:t} \cdot x + R_{1:t}(x),$$

and so in this work we can generalize (for example) the results of [McMahan and Streeter, 2010] for specific feasible sets without specifically discussing $\mathcal{F}$, and instead considering arbitrary extended convex functions Ψ.

IC-FTPRL In Section 3, we analyze a particular instance of the update rule of Equation (3), which we call *Implicit-update Composite-objective Follow-The-Proximally-Regularized-Leader* (IC-FTPRL). This algorithm adds incremental regularization R_t of the form

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

where $Q_t \in S_+^n$ is a positive-semidefinite matrix. The generalized learning rates Q_t can be chosen adaptively using techniques from [McMahan and Streeter, 2010]. We prove the following theorem on the regret of this algorithm:

Theorem 1. *Let $\Psi(x)$ be a convex function, and f_t a sequence of convex loss functions. Consider the FTPRL algorithm that plays according to Equation* (3) *using incremental regularization functions*

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

Then, there exists a $g'_t \in \partial f_t(x_{t+1})$ such that this algorithm has

$$\begin{aligned} \textit{Regret}(\mathring{x}) &\leq R_{1:T}(\mathring{x}) + \Psi(\mathring{x}) + \frac{1}{2} \sum_{t=1}^{T} \left(g_t - \frac{1}{2} g'_t\right)^\top Q_{1:t}^{-1} g'_t \\ &\leq R_{1:T}(\mathring{x}) + \Psi(\mathring{x}) + \frac{1}{2} \sum_{t=1}^{T} \|Q_{1:t}^{-\frac{1}{2}} g_t\|^2 \end{aligned}$$

versus any point $\mathring{x} \in \mathbb{R}^n$, for any $g_t \in \partial f_t(x_t)$.

If we treat Ψ as an intrinsic part of the problem, that is, as the round zero loss function, then the $\Psi(\mathring{x})$ term disappears from the regret bound.

Summary of Notation and Technical Background We use the notation $g_{1:t}$ as a shorthand for $\sum_{s=1}^{t} g_s$. Similarly we write $Q_{1:t}$ for a sum of matrices Q_t, and we use $f_{1:t}$ to denote the function $f_{1:t}(x) = \sum_{s=1}^{t} f_s(x)$. 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 write "the functions f_t" for the sequence of functions $(f_1, \ldots, f_T)$.

We write S_+^n for the set of symmetric positive semidefinite $n \times n$ matrices, with S_{++}^n the corresponding set of symmetric positive definite matrices. Recall $A \in S_{++}^n$ means $\forall x \neq 0,\ x^\top A x > 0$. Since $A \in S_+^n$ is symmetric, $x^\top A y = y^\top A x$. 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]).

Unless otherwise stated, convex functions are assumed to be extended, with domain $\mathbb{R}^n$ and range $\mathbb{R} \cup \{\infty\}$ (see, for example [Boyd and Vandenberghe, 2004, 3.1.2]). For a convex function f, we let $\partial f(x)$ denote the set of subgradients of f at x (the subdifferential of f at x). By definition, $g \in \partial f(x)$ means $f(y) \geq f(x) + g^\top (y - x)$ for all y. When f is differentiable, we write $\nabla f(x)$ for the gradient of f at x. In this case, $\partial f(x) = \{\nabla f(x)\}$. All mins and argmins are over $\mathbb{R}^n$ unless otherwise noted.

We make frequent use of the following standard results, summarized as follows:

Theorem 2. *Let $R : \mathbb{R}^n \to \mathbb{R}$ be strongly convex with continuous first partial derivatives, and let $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be an arbitrary convex function. Define $g(x) = R(x) + \Psi(x)$. Then, there exists a unique pair (x^*, ϕ^*) such that both*

$$\phi^* \in \partial \Psi(x^*) \qquad \text{and} \qquad x^* = \arg\min_x R(x) + \phi^* \cdot x.$$

Further, this x^* is the unique minimizer of g.

Note that an equivalent condition to $x^* = \arg\min_x R(x) + \phi^* \cdot x$ is

$$\nabla R(x^*) + \phi^* = 0.$$

Proof. Since R is strongly convex, g is strongly convex, and so has a unique minimizer x^* (see for example, [Boyd and Vandenberghe, 2004, 9.1.2]). Let $r = \nabla R$. Since x^* is a minimizer of g, there must exist a $\phi^* \in \partial \Psi(x^*)$ such that $r(x^*) + \phi^* = 0$, as this is a necessary (and sufficient) condition for $0 \in \partial g(x^*)$. It follows that $x^* = \arg\min_x R(x) + \phi^* \cdot x$, as $r(x^*) + \phi^*$ is the gradient of this objective at x^*. Suppose some other (x', ϕ') satisfies the conditions of the theorem. Then, $r(x') + \phi' = 0$, and so $0 \in \partial g(x')$, and so x' is a minimizer of g. Since this minimizer is unique, $x' = x^*$, and $\phi' = -r(x^*) = \phi^*$. □

2 Mirror Descent Follows the Leader

On the surface, follow-the-regularized-leader algorithms like dual averaging [Xiao, 2009] appear quite different from gradient descent (and more generally, mirror descent) algorithms like FOBOS [Duchi et al., 2010, Duchi and Singer, 2009]. However, here we show that in the case of quadratic regularization there are essentially only two differences between the algorithms:

- How they choose to center the additional strong convexity used to guarantee low regret: for example, dual averaging centers this regularization at the origin, while FOBOS centers it at the current feasible point.

- How they handle projection onto a feasible set, or more generally, how they handle an arbitrary non-smooth regularization function Ψ.

Let $f_t(x) = g_t \cdot x$ be linear functions,[1] and $f_t^R(x) = g_t \cdot x + R_t(x)$, with R_1 strongly convex and all the R_t convex. We assume that $\min_{x \in \mathbb{R}^n} R_1(x) = 0$, and assume that $x = 0$ is the unique minimizer unless otherwise noted.

2.1 Follow The Regularized Leader (FTRL)

The simplest follow-the-regularized-leader algorithm plays

$$x_{t+1} = \arg\min_x g_{1:t} \cdot x + \frac{\sigma_{1:t}}{2}\|x\|_2^2. \tag{4}$$

For $t = 1$, we typically take $x_1 = 0$. We can generalize $\frac{1}{2}\|x\|_2^2$ to an arbitrary strongly convex R by:

$$x_{t+1} = \arg\min_x g_{1:t} \cdot x + \sigma_{1:t} R(x) \tag{5}$$

Note that we could choose $\sigma_{1:t}$ independently for each t, but in practice we want $\sigma_{1:t}$ to be non-decreasing in t, and so writing it as a sum of the per-round increments $\sigma_t \geq 0$ is reasonable.

The most general update is

$$x_{t+1} = \arg\min_x g_{1:t} \cdot x + R_{1:t}(x). \tag{6}$$

where we add an additional convex function R_t on each round. Letting $R_t(x) = \sigma_t R(x)$ recovers the previous formulation.

When

$$\arg\min_{x \in \mathbb{R}^n} R_t(x) = 0,$$

we call the functions R_t (and associated algorithms) *origin-centered*. We can also define *proximal* versions of FTRL[2] that center additional regularization at the current point rather than at the origin. In this section (only), we write $\tilde{R}_t(x) = R_t(x - x_t)$ and reserve the R_t notation for origin-centered functions. Note that $\tilde{R}_t$ is only needed to select x_{t+1}, and x_t is known to the algorithm at this point, ensuring the algorithm only needs access to the first t loss functions when computing x_{t+1} (as required). The general update is

$$x_{t+1} = \arg\min_x g_{1:t} \cdot x + \tilde{R}_{1:t}(x), \tag{7}$$

In the simplest case, this becomes

$$x_{t+1} = \arg\min_x g_{1:t} \cdot x + \sum_{s=1}^{t} \frac{\sigma_s}{2}\|x - x_s\|_2^2. \tag{8}$$

[1] In this section we do not consider implicit updates for the loss functions f_t; if f_t is not linear, we can take $g_t \in \partial f_t(x_t)$.

[2] We adapt the name "proximal" from [Do et al., 2009], but note that while similar proximal regularization functions were considered, that paper deals only with gradient descent algorithms, not FTRL.

2.2 Mirror Descent

The simplest version of mirror descent is gradient descent using a constant step size η, which plays

$$x_{t+1} = x_t - \eta g_t = -\eta g_{1:t}. \tag{9}$$

In order to get low regret, T must be known in advance so η can be chosen accordingly.[3] But, since there is a closed-form solutions for the point x_{t+1} in terms of $g_{1:t}$ and η, we generalize this to a "revisionist" algorithm that on each round plays the point that gradient descent with constant step size would have played if it had used step size η_t on rounds 1 through $t-1$. That is,

$$x_{t+1} = -\eta_t g_{1:t}.$$

When $R_t(x) = \frac{\sigma_t}{2}\|x\|_2^2$ and $\eta_t = \frac{1}{\sigma_{1:t}}$, this is equivalent to the FTRL of Equation (4).

In general, we will be more interested in gradient descent algorithms which use an adaptive step size that depends (at least) on the round t. Using a variable step size η_t on each round, gradient descent plays:

$$x_{t+1} = x_t - \eta_t g_t. \tag{10}$$

An intuition for this update comes from the fact it can be re-written as

$$x_{t+1} = \arg\min_x g_t \cdot x + \frac{1}{2\eta_t}\|x - x_t\|_2^2.$$

This version captures the notion (in online learning terms) that we don't want to change our hypothesis x_t too much (for fear of predicting badly on examples we have already seen), but we do want to move in a direction that decreases the loss of our hypothesis on the most recently seen example (which is approximated by the linear function g_t).

Mirror descent algorithms use this intuition, replacing the L_2-squared penalty with an arbitrary Bregman divergence. For a differentiable, strictly convex R, the corresponding Bregman divergence is

$$\mathcal{B}_R(x, y) = R(x) - \big(R(y) + \nabla R(y) \cdot (x - y)\big)$$

for any $x, y \in \mathbb{R}^n$. We then have a generalized update of

$$x_{t+1} = \arg\min_x g_t \cdot x + \frac{1}{\eta_t}\mathcal{B}_R(x, x_t), \tag{11}$$

or more explicitly

$$x_{t+1} = r^{-1}(r(x_t) - \eta_t g_t) \tag{12}$$

where $r = \nabla R$. Letting $R(x) = \frac{1}{2}\|x\|_2^2$ so that $\mathcal{B}_R(x, x_t) = \frac{1}{2}\|x - x_t\|_2^2$ recovers the algorithm of Equation (10). One way to see this is to note that $r(x) = r^{-1}(x) = x$ in this case.

We can generalize this even further by adding a new strongly convex function R_t to the Bregman divergence on each round. Namely, let

$$\mathcal{B}_{1:t}(x, y) = \sum_{s=1}^{t} \mathcal{B}_{R_s}(x, y),$$

[3] Or a doubling trick can be used.

so the update becomes

$$x_{t+1} = \arg\min_x g_t \cdot x + \mathcal{B}_{1:t}(x, x_t) \tag{13}$$

or equivalently

$$x_{t+1} = (r_{1:t})^{-1}(r_{1:t}(x_t) - g_t) \tag{14}$$

where $r_{1:t} = \sum_{s=1}^{t} \nabla R_t = \nabla R_{1:t}$ and $(r_{1:t})^{-1}$ is the inverse of $r_{1:t}$. Note that the step size η_t is now encoded implicitly in the choice of R_t.

The COMID algorithm [Duchi et al., 2010] handles Ψ functions[4] as part of the objective on each round: $f_t(x) = g_t \cdot x + \Psi(x)$, where typically g_t is a subgradient of the loss function of the underlying learning problem. Using our notation, the COMID update is

$$x_{t+1} = \arg\min_x \eta g_t \cdot x + \mathcal{B}(x, x_t) + \eta\Psi(x),$$

which can be generalized to

$$x_{t+1} = \arg\min_x g_t \cdot x + \Psi(x) + \mathcal{B}_{1:t}(x, x_t), \tag{15}$$

where the learning rate η has been rolled into the definition of $R_1, \ldots, R_t$.

As with FTRL, an analysis of COMID that supports arbitrary extend convex Ψ also applies to standard mirror descent on a bounded feasible set $\mathcal{F}$, by noting the equivalence of the updates

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^n} \eta g_t \cdot x + \mathcal{B}(x, x_t) + \eta\Psi_{\mathcal{F}}(x), \qquad \text{and} \qquad x_{t+1} = \arg\min_{x \in \mathcal{F}} \eta g_t \cdot x + \mathcal{B}(x, x_t).$$

2.3 An Equivalence Theorem for Proximal Regularization

In Theorem 3, we show that mirror descent algorithms can be viewed as implicit-update follow-the-leader algorithms (Equation (1)):

Theorem 3. *Let R_t be a sequence of differentiable origin-centered convex functions ($\nabla R_t(0) = 0$), with R_1 strongly convex, and let Ψ be an arbitrary convex function. Let $x_1 = \hat{x}_1 = 0$. For a sequence of loss functions $f_t(x) = g_t \cdot x + \Psi(x)$, let the sequence of points played by the generalized composite-objective mirror descent algorithm be*

$$\hat{x}_{t+1} = \arg\min_x \; g_t \cdot x + \Psi(x) + \tilde{\mathcal{B}}_{1:t}(x, \hat{x}_t), \tag{16}$$

where $\tilde{R}_t(x) = R_t(x - \hat{x}_t)$, and $\tilde{\mathcal{B}}_t = \mathcal{B}_{\tilde{R}_t}$, so $\tilde{\mathcal{B}}_{1:t}$ is the Bregman divergence with respect to $\tilde{R}_1 + \cdots + \tilde{R}_t$. Consider the alternative sequence of points x_t played by implicit FTRL, Equation (1), applied to these same f_t, defined by

$$x_{t+1} = \arg\min_x \; (g_{1:t} + \phi_{1:t-1}) \cdot x + \tilde{R}_{1:t}(x) + \Psi(x) \tag{17}$$

where $\phi_t \in \partial\Psi(x_{t+1})$ such that $g_{1:t} + \phi_{1:t-1} + \nabla\tilde{R}_{1:t}(x_{t+1}) + \phi_t = 0$. Then, these algorithms are equivalent, in that $x_t = \hat{x}_t$ for all $t > 0$.

[4] Our Ψ is denoted r in [Duchi et al., 2010]

We present a few comments and a Corollary before proceeding to the proof. The Bregman divergences used by mirror descent in the theorem are with respect to the proximal functions $R_{1:t}$, whereas typically (as in Equation (13)) these functions would not depend on the previous points played. We will show when $R_t(x) = \frac{1}{2}\|Q_t^{\frac{1}{2}}x\|_2^2$, this issue disappears. Considering arbitrary Ψ functions also complicates the theorem statement somewhat. The following Corollary sidesteps these complexities, and is perhaps the most interesting and practically important consequence of the Theorem:

Corollary 4. *Let $f_t(x) = g_t \cdot x$. Then, the following algorithms play identical points:*

- *Gradient descent with generalized learning rates $Q_t \in S_+^n$, defined by:*

$$x_{t+1} = x_t - Q_{1:t}^{-1} g_t.$$

- *Implicit FTPRL with regularization functions $\tilde{R}(x) = \frac{1}{2}\|Q_t^{\frac{1}{2}}(x - x_t)\|_2^2$, which plays*

$$x_{t+1} = \arg\min_x g_{1:t} \cdot x + \tilde{R}_{1:t}(x).$$

Proof. Let $R_t(x) = \frac{1}{2}x^\top Q_t x$. It is easy to show that $R_{1:t}$ and $\tilde{R}_{1:t}$ differ by only a linear function, and so (by a standard result) $\mathcal{B}_{1:t}$ and $\tilde{\mathcal{B}}_{1:t}$ are equal, and simple algebra reveals

$$\mathcal{B}_{1:t}(x, y) = \tilde{\mathcal{B}}_{1:t}(x, y) = \frac{1}{2}\|Q_{1:t}^{\frac{1}{2}}(x - y)\|_2^2.$$

Then, it follows from Equation (12) that the first algorithm is a mirror descent algorithm using this Bregman divergence. Taking $\Psi(x) = 0$ and hence $\phi_t = 0$, the result follows from Theorem 3. □

The behavior of composite FTPRL can be different from mirror descent when a non-trivial Ψ is used. For example, suppose Ψ is the indicator function on a feasible set $\mathcal{F}$. Then, Theorem 3 shows that mirror descent on $f_t(x) = g_t \cdot x + \Psi(x)$ (equivalent to COMID in this case) approximates previously seen Ψs by their subgradients, whereas composite FTRL optimizes over Ψ explicitly. In this case, the mirror-descent update corresponds to the standard greedy projection [Zinkevich, 2003], whereas the composite FTRL algorithm corresponds to a lazy projection (for example, [McMahan and Streeter, 2010]).[5]

We can also derive a previously known result (for example, [Gordon, 1999, Sec. 3.6]) about constant step-size gradient descent, namely that gradient descent with constant step size η, which plays

$$x_{t+1} = x_t - \eta g_t = -\eta g_{1:t}$$

is equivalent to FTRL, which plays

$$x_{t+1} = \arg\min_x g_{1:t} \cdot x + \frac{1}{2\eta}\|x\|_2^2.$$

A proof using Theorem 3 follows by choosing $R_1(x) = \frac{1}{2\eta}\|x\|_2^2 = \frac{1}{2\eta}\|x - x_1\|_2^2$ and $R_t(x) = 0$ for $t > 1$, though a trivial direct proof is also possible.

[5] Zinkevich [2004, Sec. 5.2.3] describes a different lazy projection algorithm, which requires an appropriately chosen constant step-size to get low regret. Composite FTPRL does not suffer from this problem, because it always centers the additional regularization R_t at points in $\mathcal{F}$, whereas the algorithm of Zinkevich can be shown to be centering the additional regularization *outside* of $\mathcal{F}$, at the optimum of the unconstrained FTRL optimization.

Proof of Theorem 3. The proof is by induction. For the base case, we have $x_1 = \hat{x}_1 = 0$. For the induction step, assume $x_t = \hat{x}_t$.

Theorem 2 guarantees the existence of a suitable ϕ_t for use in the update of Equation (17), and so in particular there exists a $\phi_{t-1} \in \partial\Psi(x_t)$ such that

$$g_{1:t-1} + \phi_{1:t-2} + \nabla \tilde{R}_{1:t-1}(x_t) + \phi_{t-1} = 0,$$

and so applying the induction hypothesis,

$$-\nabla \tilde{R}_{1:t-1}(\hat{x}_t) = g_{1:t-1} + \phi_{1:t-1}. \tag{18}$$

Then, for some $\phi'_t \in \partial\Psi(\hat{x}_{t+1})$,

$$\begin{aligned}
\hat{x}_{t+1} &= \arg\min_x g_t \cdot x + \tilde{\mathcal{B}}_{1:t}(x, \hat{x}_t) + \Psi(x) \\
&= \arg\min_x g_t \cdot x + \tilde{\mathcal{B}}_{1:t}(x, \hat{x}_t) + \phi'_t \cdot x && \text{Theorem 2} \\
&= \arg\min_x g_t \cdot x + \tilde{R}_{1:t}(x) - \tilde{R}_{1:t}(x_t) - \nabla \tilde{R}_{1:t}(\hat{x}_t)(x - x_t) + \phi'_t \cdot x,
\end{aligned}$$

by the definition of $\tilde{\mathcal{B}}_{1:t}$. Dropping terms independent of x,

$$\begin{aligned}
&= \arg\min_x g_t \cdot x + \tilde{R}_{1:t}(x) - \nabla \tilde{R}_{1:t}(\hat{x}_t)x + \phi'_t \cdot x \\
&= \arg\min_x g_t \cdot x + \tilde{R}_{1:t}(x) - \nabla \tilde{R}_{1:t-1}(\hat{x}_t)x + \phi'_t \cdot x && \text{since } \nabla \tilde{R}_t(\hat{x}_t) = 0 \\
&= \arg\min_x g_t \cdot x + \tilde{R}_{1:t}(x) + (g_{1:t-1} + \phi_{1:t-1})x + \phi'_t \cdot x && \text{by Equation (18)} \\
&= x_{t+1}
\end{aligned}$$

The last line follows from Theorem 2, as $(\hat{x}_{t+1}, \phi'_t)$ satisfy the conditions of the theorem with respect to the objective in the optimization defining x_{t+1}. □

2.4 An Equivalence Theorem for Origin-Centered Regularization

So far, we have shown conditions under which gradient descent on $f_t(x) = g_t \cdot x$ with an adaptive step size is equivalent to follow-the-proximally-regularized-leader. In this section, we show that mirror descent on the *regularized* functions $f_t^R(x) = g_t \cdot x + R_t(x)$, with a certain natural step-size, is equivalent to a follow-the-regularized-leader algorithm with origin-centered regularization.

When $R_t(x) = \frac{\sigma_t}{2}\|x\|_2^2$, the gradient descent algorithm we consider is

$$\begin{aligned}
x_{t+1} &= x_t - \eta_t \nabla f_t^R(x_t) \\
&= x_t - \eta_t(g_t + \nabla R_t(x_t)) \\
&= x_t - \eta_t(g_t + \sigma_t x_t)).
\end{aligned}$$

Such an algorithm is proposed, for example, in [Bartlett et al., 2007, Theorem 2.1], with a step size of $\eta_t = \frac{1}{\sigma_{1:t}}$. We show (in Corollary 6) that this algorithm is identical to follow-the-leader on the functions $f_t^R(x) = g_t \cdot x + R_t(x)$, an algorithm that is minimax optimal in terms of regret against quadratic functions like f^R [Abernethy et al., 2008]. As with the

previous theorem, the difference between the two is how they handle the feasible set or in general, an arbitrary Ψ.

We will prove this result for the generalized versions of these algorithms. Instead of vanilla gradient descent, we analyze the mirror descent algorithm of Equation (15), but now g_t is replaced by $\nabla f_t^R(x_t)$:

$$x_{t+1} = \arg\min_x \nabla f_t^R(x_t) \cdot x + \mathcal{B}_{1:t}(x, x_t)$$

Note that since f_t^R is not linear, this update no longer exactly solves the optimization $\arg\min_x f_t^R(x) + \mathcal{B}_{1:t}(x, x_t)$.

Theorem 5. *Let $f_t(x) = g_t \cdot x$, and let $f_t^R(x) = g_t \cdot x + R_t(x)$, where R_t is a differentiable convex function. Let Ψ be an arbitrary convex function. Consider the composite-objective mirror-descent algorithm which plays*

$$\hat{x}_{t+1} = \arg\min_x \nabla f_t^R(\hat{x}_t) \cdot x + \Psi(x) + \mathcal{B}_{1:t}(x, \hat{x}_t), \tag{19}$$

and the FTRL algorithm which plays

$$x_{t+1} = \arg\min_x f_{1:t}^R(x) + \phi_{1:t-1} \cdot x + \Psi(x), \tag{20}$$

for $\phi_t \in \partial\Psi(x_{t+1})$ such that $g_{1:t} + \nabla R_{1:t}(x_{t+1}) + \phi_{1:t-1} + \phi_t = 0$. If both algorithms play $\hat{x}_1 = x_1 = 0$, then they are equivalent, in that $x_t = \hat{x}_t$ for all $t > 0$.

The FTRL algorithm of Equation (20) is doing full follow-the-leader on $R_{1:t} + g_{1:t}$, and an implicit update on each round's Ψ. Again, before proceeding to the proof of the main theorem it is useful to specialize to the simpler case when $\Psi(x) = 0$ and the regularization is quadratic:

Corollary 6. *Let $f_t(x) = g_t \cdot x$ and*

$$f_t^R(x) = g_t \cdot x + \frac{\sigma_t}{2}\|x\|_2^2.$$

Then following update algorithms play identical points:

- *FTRL, which plays*
$$x_{t+1} = \arg\min_x f_{1:t}^R(x).$$
- *Gradient descent on the functions f^R using the step size $\eta_t = \frac{1}{\sigma_{1:t}}$, which plays*
$$x_{t+1} = x_t - \eta_t \nabla f_t^R(x_t)$$
- *Revisionist constant-step size gradient descent with $\eta_t = \frac{1}{\sigma_{1:t}}$, which plays*
$$x_{t+1} = -\eta_t g_{1:t}.$$

The last equivalence in the corollary follows from deriving the closed form for the point played by FTRL. We now proceed to the proof of the general theorem:

Proof of Theorem 5. The proof is by induction, using the induction hypothesis $\hat{x}_t = x_t$. The base case for $t = 1$ follows by inspection. Suppose the induction hypothesis holds for t; we will show it also holds for $t + 1$.

Again let $r_t = \nabla R_t$ and consider Equation (20). Since R_1 is assumed to be strongly convex, applying Theorem 2 gives us that x_t is the unique solution to $\nabla f^R_{1:t-1}(x_t) + \phi_{1:t-1} = 0$ and so $g_{1:t-1} + r_{1:t-1}(x_t) + \phi_{1:t-1} = 0$. Then, by the induction hypothesis,

$$-r_{1:t-1}(\hat{x}_t) = g_{1:t-1} + \phi_{1:t-1}. \tag{21}$$

Now consider Equation (19). Since R_1 is strongly convex, $\mathcal{B}_{1:t}(x, \hat{x}_t)$ is strongly convex in its first argument, and so by Theorem 2 we have that $\hat{x}_{t+1}$ and some $\phi'_t \in \partial\Psi(\hat{x}_{t+1})$ are the unique solution to

$$\nabla f^R_t(\hat{x}_t) + \phi'_t + r_{1:t}(\hat{x}_{t+1}) - r_{1:t}(\hat{x}_t) = 0,$$

since $\nabla_p \mathcal{B}_R(p, q) = r(p) - r(q)$. Beginning from this equation,

$$\begin{aligned}
0 &= \nabla f^R_t(\hat{x}_t) + \phi'_t + r_{1:t}(\hat{x}_{t+1}) - r_{1:t}(\hat{x}_t) \\
&= g_t + r_t(\hat{x}_t) + \phi'_t + r_{1:t}(\hat{x}_{t+1}) - r_{1:t}(\hat{x}_t) \\
&= g_t + r_{1:t}(\hat{x}_{t+1}) + \phi'_t - r_{1:t-1}(\hat{x}_t) \\
&= g_t + r_{1:t}(\hat{x}_{t+1}) + \phi'_t + g_{1:t-1} + \phi_{1:t-1} && \text{by Equation (21)} \\
&= g_{1:t} + r_{1:t}(\hat{x}_{t+1}) + \phi_{1:t-1} + \phi'_t.
\end{aligned}$$

Applying Theorem 2 to Equation (20), (x_{t+1}, ϕ_t) are the unique pair such that

$$g_{1:t} + r_{1:t}(x_{t+1}) + \phi_{1:t-1} + \phi_t = 0$$

and $\phi_t \in \partial\Psi(x_{t+1})$, and so we conclude $\hat{x}_{t+1} = x_{t+1}$ and $\phi'_t = \phi_t$. □

3 Analysis of IC-FTPRL

In this section, we prove the regret bound of Theorem 1 for the IC-FTPRL algorithm with quadratic regularization. Recall the update is

$$x_{t+1} = \arg\min_x g'_{1:t-1} \cdot x + f_t(x) + \Psi(x) + R_{1:t}(x) \tag{3}$$

where $g'_t \in \partial f_t(x_{t+1})$ such that (x_{t+1}, g'_t) satisfy Theorem 2.

For analysis it will be useful to consider the equivalent (by Theorem 2) update

$$x_{t+1} = \arg\min_x g'_{1:t} \cdot x + \Psi(x) + R_{1:t}(x). \tag{22}$$

We can view this alternative update as running composite FTPRL on the linear approximations of f_t taken at x_{t+1},

$$\bar{f}_t(x) = f_t(x_{t+1}) + g'_t \cdot (x - x_{t+1}).$$

To see the equivalence, note the constant terms in $\bar{f}$ change neither the argmin nor regret. This is still an implicit update, as implementing the update requires an oracle to compute the the appropriate gradient g'_t (say, by finding x_{t+1} via Equation (3)).

This re-interpretation is essential, as it lets us analyze IC-FTPRL as a follow-the-leader algorithm on convex functions; note that the objective function of Equation (3) is not the sum of one convex function per round, as when moving from x_{t-1} to x_t we effectively add $g'_{t-1} \cdot x - f_{t-1}(x) + f_t(x)$ to the objective, which is not in general convex. By immediately applying an appropriate linearization of the loss functions, we avoid this non-convexity, which simplifies the analysis.

The linear functions $\bar{f}$ lower bound f_t, and so can be used to lower bound the loss of any $\mathring{x}$; however, in contrast to the more typical subgradient approximations taken at x_t, these linear functions are not tight at x_t, and so our analysis must also account for the additional loss $f_t(x_t) - \bar{f}_t(x_t)$. Before formalizing these arguments in the proof of Theorem 1, we prove the following lemma. We will use this lemma to get a tight bound on the regret of the algorithm against the linearized functions $\bar{f}$, but it is in fact much more general.

Lemma 7. *Let f_t be a sequence of arbitrary loss functions, and let R_t be arbitrary non-negative regularization functions. Define $f_t^R(x) = f_t(x) + R_t(x)$. Then, if we play $x_{t+1} = \arg\min_x f_{1:t}^R(x)$, our regret against the functions f_t versus an arbitrary point $\mathring{x}$ is bounded by*

$$\textit{Regret} \leq R_{1:T}(\mathring{x}) + \sum_{t=1}^{T} f_{1:t}^R(x_t) - f_{1:t}^R(x_{t+1}).$$

A weaker (though sometimes easier to use) version of this lemma, stating

$$\text{Regret} \leq R_{1:T}(\mathring{x}) + \sum_{t=1}^{T} f_t(x_t) - f_t(x_{t+1}),$$

has been used previously [Kalai and Vempala, 2005, Hazan, 2008, McMahan and Streeter, 2010]. In the case of linear functions with quadratic regularization, as in the analysis of of [McMahan and Streeter, 2010], the weaker version loses a factor of $\frac{1}{2}$ (corresponding to a $\sqrt{2}$ in the final bound). The key is that in this case, being the leader is *strictly better* than playing the post-hoc optimal point. Quantifying this difference leads to the improved bounds.

Proof of Lemma 7. First, we consider regret against the functions f^R for not playing $\mathring{x}$:

$$\begin{aligned}
\text{Regret}(f^R) &= \sum_{t=1}^{T} (f_t^R(x_t) - f_t^R(\mathring{x})) && \text{by definition} \\
&= \sum_{t=1}^{T} f_t^R(x_t) - f_{1:T}^R(\mathring{x}) \\
&= \sum_{t=1}^{T} (f_{1:t}^R(x_t) - f_{1:t-1}^R(x_t)) - f_{1:T}^R(\mathring{x}) && \text{where } f_{1:0}(x) = 0 \\
&\leq \sum_{t=1}^{T} (f_{1:t}^R(x_t) - f_{1:t-1}^R(x_t)) - f_{1:T}^R(x_{T+1}) && \text{since } x_{T+1} \text{ minimizes } f_{1:T}^R \\
&= \sum_{t=1}^{T} (f_{1:t}^R(x_t) - f_{1:t}^R(x_{t+1})),
\end{aligned}$$

where the last line follows by simply re-indexing the $-f^R_{1:t}$ terms. Equivalently, applying the definitions of regret and f^R,

$$\sum_{t=1}^{T}(f_t(x_t) + R_t(x_t)) - f_{1:T}(\mathring{x}) - R_{1:T}(\mathring{x}) \leq \sum_{t=1}^{T}(f^R_{1:t}(x_t) - f^R_{1:t}(x_{t+1})).$$

Dropping the non-negative terms $R_t(x_t)$ and re-arranging proves the theorem. □

Observe that the $R_t(x_t)$ terms are exactly zero in the case of proximal regularization, and so in this case the *only* inequality in Lemma 7 is the fact that the point $\mathring{x}$ used for comparison (typically, the minimizer of $f_{1:T}$) will in general not be x_{T+1}, the minimizer of $f^R_{1:T}$.

Proof of Theorem 1. Let $\bar{f}_t(x) = f_t(x_{t+1}) + g'_t \cdot (x - x_{t+1})$, a linear approximation of f_t taken at the next point, x_{t+1}. We can bound the regret of our algorithm (expressed as an FTRL algorithm on the functions $\bar{f}_t$, Equation (22)) against the functions $\bar{f}_t$ by applying Lemma 7 to the functions $\bar{f}_t$ with regularization functions R'_t where $R'_1(x) = R_1(x) + \Psi(x)$ and $R'_t(x) = R_t(x)$ for $t > 1$. However, because we are taking the linear approximation at x_{t+1} instead of x_t, it may be the case that our actual loss $f_t(x_t)$ on round t is greater than the loss under $\bar{f}_t$, that is we may have $f_t(x_t) > \bar{f}_t(x_t)$. Thus, we must account for this additional regret. From the definition of regret we have

$$\text{Regret}(f) = \text{Regret}(\bar{f}) + \sum_{t=1}^{T}(f_t(x_t) - \bar{f}_t(x_t)) + (\bar{f}_{1:t}(\mathring{x}) - f_{1:t}(\mathring{x})),$$

and since $\bar{f}_t$ lower bounds f_t,

$$\leq \text{Regret}(\bar{f}) + \sum_{t=1}^{T}(f_t(x_t) - \bar{f}_t(x_t))$$

and letting $\bar{f}^R_t(x) = \bar{f}_t(x) + R'_t(x)$,

$$\leq \underbrace{R'_{1:T}(\mathring{x}) + \sum_{t=1}^{T}(\bar{f}^R_{1:t}(x_t) - \bar{f}^R_{1:t}(x_{t+1}))}_{\text{Lemma 7 on } \bar{f}_t \text{ and } R'_t} \quad + \quad \underbrace{\sum_{t=1}^{T}(f_t(x_t) - \bar{f}_t(x_t))}_{\text{Underestimate of real loss at } x_t}.$$

Let Δ_t be the contribution of the non-regularization terms for a particular t,

$$\Delta_t = \bar{f}^R_{1:t}(x_t) - \bar{f}^R_{1:t}(x_{t+1}) + f_t(x_t) - \bar{f}_t(x_t),$$

and expanding the definition of $\bar{f}^R_{1:t}$ and re-arranging terms,

$$\begin{aligned}
&= \bar{f}_{1:t}(x_t) + R_{1:t}(x_t) + \Psi(x_t) - \bar{f}_{1:t}(x_{t+1}) - R_{1:t}(x_{t+1}) - \Psi(x_{t+1}) + f_t(x_t) - \bar{f}_t(x_t),\\
&= \bar{f}_{1:t-1}(x_t) + R_{1:t}(x_t) + \Psi(x_t) - \bar{f}_{1:t}(x_{t+1}) - R_{1:t}(x_{t+1}) - \Psi(x_{t+1}) + f_t(x_t),\\
&= (\bar{f}_{1:t-1}(x_t) + R_{1:t}(x_t) + \Psi(x_t) + f_t(x_t)) - (\bar{f}_{1:t}(x_{t+1}) + R_{1:t}(x_{t+1}) + \Psi(x_{t+1})).
\end{aligned}$$

For the terms containing x_{t+1}, using the fact that $\bar{f}_t(x_{t+1}) = f(x_{t+1})$, we have

$$\bar{f}_{1:t}(x_{t+1}) + R_{1:t}(x_{t+1}) + \Psi(x_{t+1}) = \\ \bar{f}_{1:t-1}(x_{t+1}) + R_{1:t}(x_{t+1}) + \Psi(x_{t+1}) + f_t(x_{t+1}). \tag{23}$$

Let $\hat{h}_1(x) = \bar{f}_{1:t-1}(x) + R_{1:t}(x) + \Psi(x)$ and let $\hat{h}_2(x) = \hat{h}_1(x) + f_t(x)$, so

$$\Delta_t = \hat{h}_2(x_t) - \hat{h}_2(x_{t+1}).$$

Beginning from Equation (22), we have

$$\begin{aligned} x_t &= \arg\min_x g'_{1:t-1} \cdot x + \Psi(x) + R_{1:t-1}(x) \\ &= \arg\min_x \bar{f}_{1:t-1}(x) + \Psi(x) + R_{1:t-1}(x) \\ &= \arg\min_x \bar{f}_{1:t-1}(x) + \Psi(x) + R_{1:t}(x) \qquad \text{since } R_t(x) \text{ is minimized by } x_t \\ &= \arg\min_x \hat{h}_1(x). \end{aligned}$$

Note $\hat{h}_2(x)$ is equivalent to the objective of Equation (3) for $t+1$ (it differs only by a constant term), and so $x_{t+1} = \arg\min_x \hat{h}_2(x)$. The remainder of the proof is accomplished with the aid of two additional lemmas, stated and proved below. Applying Lemma 8, we can write

$$\hat{h}_1(x) = \frac{1}{2}\|Q_{1:t}^{\frac{1}{2}}(x - x_t)\| + \hat{\Psi}(x) + k.$$

Then, applying Lemma 9 to $\hat{h}_1$ and $\hat{h}_2$ gives the desired bound,

$$\Delta_t = \hat{h}_2(x_t) - \hat{h}_2(x_{t+1}) \leq (g_t - \frac{1}{2}g'_t)^\top Q_{1:t}^{-1} g'_t \leq \frac{1}{2} g_t^\top Q_{1:t}^{-1} g_t.$$

Summing this bound over t and adding back the $R_{1:t}(\mathring{x})$ term completes the proof. □

We now prove the two lemmas used in bounding the $\hat{h}_2(x_t) - \hat{h}_2(x_{t+1})$ terms.

Lemma 8. *Let Ψ be a convex function defined on $\mathbb{R}^n$, and let $Q \in S_{++}^n$. Define*

$$h(x) = \frac{1}{2}x^\top Q x + b \cdot x + \Psi(x),$$

and let $x_2 = \arg\min_x h(x)$. Then, we can rewrite h as

$$h(x) = \frac{1}{2}\|Q(x - x_2)\|_2^2 + \hat{\Psi}(x) + k,$$

where $k \in \mathbb{R}$ and $\hat{\Psi}$ is convex with $0 \in \partial\hat{\Psi}(x_2)$.

Proof. Since $Q \in S_{++}^n$, the function $\frac{1}{2}x^\top Q x$ is strongly convex, and so using Theorem 2, h has a unique minimizer x_2 and there exists a (unique) ϕ such that

$$Qx_2 + b + \phi = 0 \tag{24}$$

with $\phi \in \partial\Psi(x_2)$. Define $\hat{\Psi}(x) = \Psi(x) - \phi \cdot x$, and note $0 \in \partial\hat{\Psi}(x_2)$. Then,

$$\begin{aligned}
h(x) =& \frac{1}{2}x^\top Qx + b \cdot x + \Psi(x) \\
=& \frac{1}{2}x^\top Qx + (b+\phi) \cdot x + \hat{\Psi}(x) && \text{Defn. } \hat{\Psi}(x) \\
=& \frac{1}{2}x^\top Qx - x^\top Qx_2 + \hat{\Psi}(x) && \text{Eq. (24)} \\
=& \frac{1}{2}x^\top Qx - x^\top Qx_2 + \frac{1}{2}x_2^\top Qx_2 + \hat{\Psi}(x) - \frac{1}{2}x_2^\top Qx_2 \\
=& \frac{1}{2}\|Q^{\frac{1}{2}}(x - x_2)\|_2^2 + \hat{\Psi}(x) - \frac{1}{2}\|Q^{\frac{1}{2}}x_2\|_2^2,
\end{aligned}$$

where $\hat{\Psi}$ and $k = -\frac{1}{2}\|Q^{\frac{1}{2}}x_2\|_2^2$ satisfy the requirements of the theorem. □

Now, we show how to bound the cost of not being the leader for functions in this form:

Lemma 9. *Let*

$$h_1(x) = \frac{1}{2}\|Q^{\frac{1}{2}}(x - x_2)\|_2^2$$

where $x_2 = \arg\min_x h_1(x)$ *and* $Q \in S_{++}^n$. *Let* $h_2(x) = h_1(x) + f(x)$ *for an arbitrary convex function* f, *and let* $x_3 = \arg\min_x h_2(x)$. *Then, there exists a* $g' \in \partial f(x_3)$ *such that* $Q(x_3 - x_2) + g' = 0$, *and for any* $g \in \partial f(x_2)$,

$$h_2(x_2) - h_2(x_3) \leq (g - \frac{1}{2}g')^\top Q^{-1}g' \leq \frac{1}{2}g^\top Q^{-1}g. \tag{25}$$

Further, adding an additional convex function $\hat{\Psi}$ *centered at* x_2 *to each* h *cannot make the gap any larger: Let* $\hat{\Psi}$ *be a convex function such that* $0 \in \partial\hat{\Psi}(x_2)$, *and define* $\hat{h}_1(x) = h_1(x) + \hat{\Psi}(x)$ *and* $\hat{h}_2(x) = h_2(x) + \hat{\Psi}(x)$. *Let* $\hat{x}_2 = \arg\min_x \hat{h}_1(x)$ *and* $\hat{x}_3 = \arg\min_x \hat{h}_2(x)$. *Then,*

$$\hat{h}_2(\hat{x}_2) - \hat{h}_2(\hat{x}_3) \leq h_2(x_2) - h_2(x_3). \tag{26}$$

Proof. We first prove Equation (25). Theorem 2 guarantees the needed g' exists. Then, for the first inequality we have

$$\begin{aligned}
&h_2(x_2) - h_2(x_3) \\
&\quad = f(x_2) - \frac{1}{2}\|Q^{\frac{1}{2}}(x_3 - x_2)\|_2^2 - f(x_3) \\
&\quad \leq -\frac{1}{2}\|Q^{\frac{1}{2}}(x_3 - x_2)\|_2^2 + g(x_2 - x_3) && \text{because } f(x_3) \geq f(x_2) + g(x_3 - x_2) \\
&\quad = -\frac{1}{2}\|Q^{\frac{1}{2}}(Q^{-1}g')\|_2^2 + g^\top Q^{-1}g' && \text{because } Q(x_3 - x_2) + g' = 0 \\
&\quad = -\frac{1}{2}g'^\top Q^{-1}g' + g^\top Q^{-1}g' \\
&\quad = (g - \frac{1}{2}g')^\top Q^{-1}g'.
\end{aligned}$$

For the second inequality of Equation (25), observe an equivalent statement is

$$g^\top Q^{-1}g' \leq \frac{1}{2}g^\top Q^{-1}g + \frac{1}{2}g'^\top Q^{-1}g'.$$

Recall by Hölder's inequality, for any vectors $x, y \in \mathbb{R}^n$,

$$x \cdot y \leq \|x\|\|y\| \leq \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2.$$

Letting $x \leftarrow Q^{-\frac{1}{2}} g$ and $y \leftarrow Q^{-\frac{1}{2}} g'$ then gives the desired inequality.

We now consider Equation (26). First, since h_1 is strongly convex and $0 \in \partial\hat{\Psi}(x_2)$, we have $\hat{x}_2 = x_2$. Also note that $\hat{h}_2(\hat{x}_3) = h_2(\hat{x}_3) + \hat{\Psi}(\hat{x}_3) \geq h_2(x_3) + \hat{\Psi}(x_2)$, since x_3 minimizes h_2 and x_2 minimizes $\hat{\Psi}$. Then, for Equation (26) we have

$$\begin{aligned}
\hat{h}_2(\hat{x}_2) - \hat{h}_2(\hat{x}_3) &= h_2(x_2) + \hat{\Psi}(x_2) - \hat{h}_2(\hat{x}_3) \\
&\leq h_2(x_2) + \hat{\Psi}(x_2) - (h_2(x_3) + \hat{\Psi}(x_2)) \\
&= h_2(x_2) - h_2(x_3).
\end{aligned}$$

□

Acknowledgements

The author wishes to thank Matt Streeter for numerous helpful discussions, as well as detailed comments on earlier drafts of this work.

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 Bartlett, Elad Hazan, and Alexander Rakhlin. Adaptive online gradient descent. Technical Report UCB/EECS-2007-82, EECS Department, University of California, Berkeley, Jun 2007. URL `http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-82.html`.

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

Chuong B. Do, Quoc V. Le, and Chuan-Sheng Foo. Proximal regularization for online and batch learning. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 257–264, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: http://doi.acm.org/10.1145/1553374.1553407.

John Duchi and Yoram Singer. Efficient learning using forward-backward splitting. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 495–503. 2009.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *COLT*, 2010.

Geoffrey J. Gordon. *Approximate solutions to markov decision processes*. PhD thesis, Carnegie Mellon University, 1999. Chair - Tom Mitchell.

Elad Hazan. Extracting certainty from uncertainty: Regret bounded by variation in costs. In *In COLT*, 2008.

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

Brian Kulis and Peter Bartlett. Implicit online learning. In *ICML*, 2010.

H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *COLT*, 2010.

Shai Shalev-Shwartz and Sham M. Kakade. Mind the duality gap: Logarithmic regret algorithms for online optimization. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Lon Bottou, editors, *NIPS*, pages 1457–1464. MIT Press, 2008.

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.

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.