Adaptive Subgradient Methods for Online Learning and Stochastic Optimization

Adaptive Subgradient Methods for Online Learning and Stochastic Optimization

John Duchi Elad Hazan Yoram Singer

January 15, 2010

Abstract

Stochastic subgradient methods are widely used, well analyzed, and constitute effective tools for optimization and online learning. Stochastic gradient methods' popularity and appeal are largely due to their simplicity, as they largely follow predetermined procedural schemes. However, most common subgradient approaches are oblivious to the characteristics of the data being observed. We present a new family of subgradient methods that dynamically incorporate knowledge of the geometry of the data observed in earlier iterations to perform more informative gradient-based learning. The adaptation, in essence, allows us to find needles in haystacks in the form of very predictive but rarely seen features. Our paradigm stems from recent advances in stochastic optimization and online learning which employ proximal functions to control the gradient steps of the algorithm. We describe and analyze an apparatus for adaptively modifying the proximal function, which significantly simplifies setting a learning rate and results in regret guarantees that are provably as good as the best proximal function that can be chosen in hindsight. In a companion paper, we validate experimentally our theoretical analysis and show that the adaptive subgradient approach outperforms state-of-the-art, but non-adaptive, subgradient algorithms.

1 Introduction

In many applications of online and stochastic learning, the input instances are of very high dimension, yet within any particular instance only a few features are non-zero. It is often the case, however, that infrequently occurring features are highly informative and discriminative. The informativeness of rare features has led practitioners to craft domain-specific feature weightings, such as TF-IDF (Salton and Buckley, 1988), which pre-emphasize infrequently occurring features. We use this old idea as a motivation for applying modern learning-theoretic techniques to the problem of online and stochastic learning, focusing specifically on (sub)gradient methods.

Standard stochastic subgradient methods largely follow a predetermined procedural scheme that is oblivious to the characteristics of the data being observed. In contrast, our algorithms dynamically incorporate knowledge of the geometry of the data observed in earlier iterations to perform more informative gradient-based learning. The adaptation facilitates finding and identifying very predictive but comparatively rare features. The following toy example of classification with the hinge loss highlights a problem with standard subgradient methods. We receive a sequence of 3-dimensional vectors $\{z_t\}$ and labels $y_t \in \{-1, +1\}$. Let u_t be independent ±1-valued random variables, each with probability $\frac{1}{2}$. Each instance z_t is associated with a label $y_t = -1$ and is equal to $(u_t, -u_t, 0)$ 99% of the time, while $z_t = (u_t, -u_t, 1)$ the remaining 1% of the cases with label $y_t = 1$. We would like find a vector x with $\|x\|_\infty \le 1$ such that $\max\{0, -y_t \langle x, z_t \rangle\} = 0$ for all t. Clearly, any solution vector takes the form $x = (a, a, 1)$ with $|a| \le 1$. Standard subgradient methods, such as the one proposed by Zinkevich (2003), iterate as follows:

$$x_{t+1} = \Pi_{\{x:\|x\|_\infty \le 1\}}(x_t - \eta_t y_t z_t) \ ,$$

where Π_X denotes Euclidean projection onto the set X and $\eta_t = 1/\sqrt{t}$. In this case, in expectation it takes the subgradient approximately 2,500 iterations until the third component of x_t is 1. In contrast, our adaptive gradient method sets the third component of x to be 1 the first time it sees a positive label, thus requiring 100 iterations in expectation.

1.1 The Adaptive Gradient Algorithm

Before introducing our adaptive gradient algorithm, which we term AdaGrad, let us establish our notation. Vectors and scalars are lower case italic letters, such as $x \in X$. We denote a sequence of vectors by subscripts, i.e. $x_t, x_{t+1}, \ldots$, and entries of each vector by an additional subscript, e.g. $x_{t,j}$. The subdifferential set of a function f evaluated at x is denoted $\partial f(x)$, and a particular vector in the subdifferential set is denoted by $f'(x) \in \partial f(x)$. We let g_t denote a particular vector of $\partial f_t(x_t)$. When a function is differentiable, we write $\nabla f(x)$. We use $\langle x, y \rangle$ to denote the inner product between x and y. Recall that the Bregman divergence associated with a strongly convex and differentiable function ψ is

$$B_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle \, .$$

We also make frequent use of the following two matrices. Let $g_{1:t} = [g_1 \ \cdots \ g_t]$ denote the matrix obtained by concatenating the subgradient sequence. We denote the ith row of this matrix, which amounts to the concatenation of the ith component of each subgradient we observe, by $g_{1:t,i}$. Last, we define the outer product matrix $G_t = \sum_{\tau=1}^{t} g_\tau g_\tau^\top$.

We consider in this paper several different online learning algorithms and their stochastic convex optimization counterparts. In online learning, the learner's goal is to achieve low regret with respect to a static predictor x^* in a (closed) convex set X on a sequence of functions $\phi_t(x)$. Each function $\phi_t(x)$ is of the form $f_t(x) + \varphi(x)$ where f_t and φ are (closed) convex functions. We would like to note that X can possibly be $\mathbb{R}^d$. In the learning settings we study, f_t is either an instantaneous loss or a stochastic estimate of the objective function in stochastic optimization. The function φ serves as a fixed regularization function and is typically used to control the complexity of x. Formally, at every round of the algorithm we make a prediction $x_t \in X$ and then receive the function f_t. We define the regret as

$$R_\phi(T) \triangleq \sum_{t=1}^{T} \left[\phi_t(x_t) - \phi_t(x^*)\right] = \sum_{t=1}^{T} \left[f_t(x_t) + \varphi(x_t) - f_t(x^*) - \varphi(x^*)\right] \, . \tag{1}$$

Our goal is to devise algorithms which are guaranteed to suffer regret that is asymptotically sub-linear, namely, $R_\phi(T) = o(T)$.

Our analysis applies to related, yet different, methods for for minimizing the regret defined in Eq. (1). The first is Nesterov's primal-dual subgradient method (Nesterov, 2009), and in particular its specialized versions: regularized dual asent (RDA) of Xiao (2009) and the follow-the-regularized-leader (FTRL) family of algorithms (see for instance Kalai and Vempala, 2003; Hazan et al., 2006; or Abernethy et al., 2008b). In the primal-daul subgradient method we make a prediction x_t on round t using the average gradient $\bar{g}_t = \frac{1}{t} \sum_{\tau=1}^{t} g_\tau$. The update encompasses a trade-off between a gradient-dependent linear term, the regularizer φ, and a strongly-convex term ψ_t for well-conditioned predictions. Here ψ is the *proximal* term. The update amounts to solving the problem

$$x_{t+1} = \operatorname*{argmin}_{x \in X} \left\{ \eta \langle \bar{g}_t, x \rangle + \eta \varphi(x) + \frac{1}{t} \psi_t(x) \right\} \, , \tag{2}$$

where η is a step-size. The second method is the composite mirror descent (Duchi et al., 2010). The composite mirror descent method employs a more immediate trade-off between the current gradient g_t, φ, and staying close to x_t using the proximal function ψ,

$$x_{t+1} = \operatorname*{argmin}_{x \in X} \left\{ \eta \langle g_t, x \rangle + \eta \varphi(x) + B_{\psi_t}(x, x_t) \right\} \, . \tag{3}$$

Our work focuses on temporal adaptation of the proximal function in a data driven way, while previous work simply sets $\psi_t \equiv \psi$, $\psi_t(\cdot) = \sqrt{t}\psi(\cdot)$, or $\psi_t(\cdot) = t\psi(\cdot)$ for some fixed ψ.

We provide formal analyses equally applicable to the above two updates and show how to automatically choose the function ψ_t so as to achieve asymptotically small regret. We describe and analyze two algorithms. Both algorithms use squared Mahalanobis norms as their proximal functions, setting $\psi_t(x) = \langle x, H_t x \rangle$ for a symmetric matrix $H_t \succeq 0$. The first uses diagonal matrices while the second constructs full dimensional matrices. Concretely, we set

$$H_t = \operatorname{diag}(G_t)^{1/2} \ \ \text{(Diagonal)} \quad ; \quad H_t = G_t^{1/2} \ \ \text{(Full)} \, . \tag{4}$$

Plugging the appropriate matrix from the above equation into ψ_t in Eq. (2) or Eq. (3) gives rise to our AdaGrad family of algorithms. Informally, we obtain algorithms which are similar to second-order gradient descent by constructing approximations to the Hessian of the functions f_t. These approximations are conservative since we rely on the square root of the gradient matrices.

1.2 Motivating Example

As mentioned in the prequel, we expect our adaptive methods to outperform standard online learning methods when the gradient vectors are sparse. We now give two concrete examples in which we receive sparse data and the adaptive algorithms achieve much lower regret than a non-adaptive version. In both examples we use the hinge loss, that is,

$$f_t(x) = \left[1 - y_t \langle z_t, x\rangle\right]_+ \ ,$$

where y_t is the label of example t and $z_t \in \mathbb{R}^d$ is the data vector. Both examples construct a sparse sequence for which there is a perfect predictor that the adaptive methods learn after d iterations, while standard online gradient descent (Zinkevich, 2003) suffers significantly higher loss. We also assume the domain X is compact so that for online gradient descent we set $\eta_t = 1/\sqrt{t}$, which gives $O(\sqrt{T})$ regret.

Diagonal Adaptation In this first example, we consider the diagonal version of our proposed update in Eq. (3) with $X = \{x : \|x\|_\infty \le 1\}$. Evidently, this choice simply results in the update $x_{t+1} = x_t - \eta \operatorname{diag}(G_t)^{-1/2} g_t$ followed by projection onto X. Let e_i denote the ith unit basis vector, and assume that for each t, $z_t = \pm e_i$ for some i. Also let $y_t = \operatorname{sign}(\langle \mathbb{1}, z_t\rangle)$ so that there exists a perfect classifier $x^* = \mathbb{1} \in X \subset \mathbb{R}^d$. We initialize x_1 to be the zero vector. On rounds $t = 1, \ldots, d$, we set $z_t = \pm e_t$, selecting the sign at random. It is clear that both diagonal adaptive descent and online gradient descent suffer a unit loss on each of the first d examples. However, the updates to parameter x_i on iteration i differ and amount to

$$x_{t+1} = x_t + e_t \ \ (\text{AdaGrad}) \qquad\qquad x_{t+1} = x_t + \frac{1}{\sqrt{t}} e_t \ \ (\text{Gradient Descent}) \ .$$

It is clear that after the first d rounds, the adaptive predictor achieves $x_{d+1} = x_{d+\tau} = \mathbb{1}$ for all $\tau \ge 1$ and ceases to suffer further losses. However, gradient descent suffers losses on rounds $t = d+1$ through $2d$ of $\sum_{t=d+1}^{2d} \left[1 - \frac{1}{\sqrt{t-d}}\right]_+ = \sum_{i=1}^{d} \left[1 - \frac{1}{\sqrt{i}}\right]_+$. Thus, the ith composite of the predictor is updated to $1/\sqrt{i} + 1/\sqrt{d+i}$ after the second d rounds (truncated so that $|x_i| \le 1$). In general, the regret of adaptive gradient descent is d, while we see that online gradient descent suffers regret

$$d + \sum_{t=0}^{T} \sum_{i=1}^{d} \left[1 - \sum_{\tau=0}^{t} \frac{1}{\sqrt{i + \tau d}}\right]_+ \ . \tag{5}$$

The next proposition lower bounds Eq. (5).

Proposition 1. *The loss suffered by online gradient descent in the problem described above is $\Omega(d\sqrt{d})$. In particular, if $T \ge \sqrt{d} + 1$,*

$$d + \sum_{t=0}^{T} \sum_{i=1}^{d} \left[1 - \sum_{\tau=0}^{t} \frac{1}{\sqrt{i + \tau d}}\right]_+ \ge d + \frac{d\sqrt{d}}{4} \ .$$

We give the proof of the proposition in Appendix A. For example, in a 10000 dimensional problem, AdaGrad suffers a cumulative loss of only $d = 10^4$, while standard stochastic gradient descent suffers loss of at least $2.6 \cdot 10^5$. We also note here that with larger stepsizes than $\eta/\sqrt{t}$, online gradient descent might suffer a lower loss in the above game. However, an adversary could simply play $z_t = e_1$ indefinitely until $\eta/\sqrt{t} \le \varepsilon$ for any $\varepsilon > 0$, in which case online gradient descent can be made to suffer regret of $\Omega(d^2)$ while AdaGrad still achieves constant regret per dimension.

Full Matrix Adaptation We use a similar construction to the diagonal case to show a situation in which the full matrix update from Eq. (4) gives substantially lower regret than stochastic gradient descent. For full divergences we set $X = \{x : \|x\|_2 \le \sqrt{d}\}$. Let $V = [v_1 \ \ldots \ v_d] \in \mathbb{R}^{d \times d}$ be an orthonormal matrix. Instead of having z_t cycle through the unit vectors, we make z_t cycle through the v_i so that $z_t = \pm v_{(t \mod d)+1}$. We let the label $y_t = \text{sign}(\langle 1, V^\top z_t \rangle) = \text{sign}\left(\sum_{i=1}^d \langle v_i, z_t \rangle\right)$. We provide an elaborated explanation in Appendix A. Intuitively, we see that with $\psi_t(x) = \langle x, H_t x \rangle$ and H_t set to be the full matrix from Eq. (4), AdaGrad again needs to observe each orthonormal vector v_i only once while the stochastic gradient descent's loss is again $\Omega(d\sqrt{d})$.

1.3 Outline of Results

We now outline our results, deferring formal statements of the theorems to later sections. Recall the definitions of $g_{1:t}$ as the matrix of concatenated subgradients and G_t as the outer product matrix in the prequel. The AdaGrad algorithm with full matrix divergences entertains bounds of the form

$$R_\phi(T) = O\left(\|x^*\|_2 \operatorname{tr}(G_T^{1/2})\right) \quad \text{and} \quad R_\phi(T) = O\left(\max_{t \le T} \|x_t - x^*\|_2 \operatorname{tr}(G_T^{1/2})\right).$$

We further show that

$$\operatorname{tr}\left(G_T^{1/2}\right) = d^{1/2} \sqrt{\inf_S \left\{ \sum_{t=1}^T \langle g_t, S^{-1} g_t \rangle \ : \ S \succeq 0, \operatorname{tr}(S) \le d \right\}}.$$

These results are formally given in Theorem 8 and its corollaries. When our proximal function $\psi_t(x) = \left\langle x, \operatorname{diag}(G_t)^{1/2} x \right\rangle$ we have bounds, attainable in time at most linear in the dimension d of our problems, of the form

$$R_\phi(T) = O\left(\|x^*\|_\infty \sum_{i=1}^d \|g_{1:T,i}\|_2\right) \quad \text{and} \quad R_\phi(T) = O\left(\max_{t \le T} \|x_t - x^*\|_\infty \sum_{i=1}^d \|g_{1:T,i}\|_2\right).$$

Similar to the above, we will show that

$$\sum_{i=1}^d \|g_{1:T,i}\|_2 = d^{1/2} \sqrt{\inf_s \left\{ \sum_{t=1}^T \langle g_t, \operatorname{diag}(s)^{-1} g_t \rangle \ : \ s \succeq 0, \langle 1, s \rangle \le d \right\}}.$$

We formally state the above two regret bounds in Theorem 6 and its corollaries.

We give here one simple example and a corollary to Theorem 6 to illustrate one regime in which we expect substantial improvements. For simplicity let $\varphi \equiv 0$. Consider Zinkevich's 2003 online gradient descent algorithm. Given a compact convex set $X \subseteq \mathbb{R}^d$ and sequence of convex functions f_t, Zinkevich's algorithm makes the sequence of predictions $x_1, \ldots, x_T$ with $x_{t+1} = \Pi_X(x_t - (\eta/\sqrt{t}) g_t)$. If the diameter of X is bounded, so $\sup_{x,y \in X} \|x - y\|_2 \le D_2$, then Zinkevich's algorithm—with the optimal choice in *hindsight* for the stepsize η (see Eq. (7))—achieves regret

$$\sum_{t=1}^T f_t(x_t) - \inf_{x \in X} \sum_{t=1}^T f_t(x) \le \sqrt{2} D_2 \sqrt{\sum_{t=1}^T \|g_t\|_2^2}. \tag{6}$$

When X is bounded via $\sup_{x,y \in X} \|x - y\|_\infty \le D_\infty$, the following corollary is an easy consequence of our Theorem 6, and we give a brief proof in Appendix B.

Corollary 2. *Let the sequence $\{x_t\} \subset \mathbb{R}^d$ be generated by the update in Eq. (3) and* $\max_t \|x^* - x_t\|_\infty \le$ *D_∞. Then with stepsize $\eta = D_\infty/\sqrt{2}$, for any x^*, we have*

$$R_\phi(T) \le \sqrt{2d} D_\infty \sqrt{\inf_{s \succeq 0, \langle 1, s \rangle \le d} \sum_{t=1}^T \|g_t\|_{\operatorname{diag}(s)^{-1}}^2} = \sqrt{2} D_\infty \sum_{i=1}^d \|g_{1:T,i}\|_2.$$

The important feature of the bound above is the infimum under the square root, which allows us to do better than simply using the identity matrix, and the fact that the stepsize is easy to set a priori. For example, if the set $X = \{x : \|x\|_\infty \le 1\}$, then $D_2 = 2\sqrt{d}$ while $D_\infty = 2$, which suggests that if we are learning a dense predictor over a box, the adaptive method should perform well. Indeed, in this case we are guaranteed that the bound in Corollary 2 is better than Eq. (6) as the identity matrix belongs to the set over which we take the infimum.[1]

1.4 Related Work

Many successful algorithms have been developed over the past few years to minimize regret in the online learning setting. A modern view of these algorithms cast the problem as the task of following the (regularized) leader (see Rakhlin, 2009 and the references therein) or FTRL in short. Informally, FTRL methods choose the best decision in hindsight at every iteration. Plain usage of the FTRL approach fails to achieve low regret, however, adding a proximal[2] term to the past predictions leads to numerous low regret algorithms (Kalai and Vempala, 2003; Hazan and Kale, 2008; Rakhlin, 2009). The proximal term strongly affects the performance of the learning algorithm. Therefore, adapting the proximal function to the characteristics of the problem at hand is desirable.

Our approach is thus motivated by two goals. The first is to generalize the agnostic online learning paradigm to the meta-task of specializing an algorithm to fit a particular dataset. Specifically, we change the proximal function to achieve performance guarantees which are competitive with the best proximal term found in hindsight. The second, as alluded to earlier, is to automatically tune the learning rates for online learning and stochastic gradient descent on a per-feature basis. The latter goal can potentially be very useful when our gradient vectors g_t are sparse, for example, in a classification setting where examples have only a small number of features on at a particular time. As we demonstrate in the examples above, it is rather deficient to to employ exactly the same learning rate for a feature seen hundreds of times and for a feature seen only once or twice.

Our techniques stem from a variety of research directions, and as a byproduct we extend a few well-known algorithms. In particular, we consider variants of the follow-the-regularized leader (FTRL) algorithms mentioned above, which are close in spirit to Zinkevich's lazy projection algorithm (Zinkevich, 2003). We use the recently analyzed regularized dual ascent (RDA) algorithm of Xiao (2009), which builds upon Nesterov's 2009 primal-dual subgradient method. We also consider the forward-backward splitting (FOBOS) algorithmic framework (Duchi and Singer, 2009) and its composite mirror-descent generalizations (Duchi et al., 2010), which in turn include as special cases the method of projected gradient descent (Zinkevich, 2003) and mirror descent (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003). Prior to the analysis presented in this paper, the strongly convex function ψ in the update equations (2) and (3) either remained intact or was simply multiplied by a time-dependent scalar throughout the run of the algorithm. Zinkevich's projected gradient, for example, uses $\psi_t(x) = \|x\|_2^2$, while RDA (Xiao, 2009) employs $\psi_t(x) = \sqrt{t}\psi(x)$ where ψ is a strongly convex function. The bounds for both types of algorithms are similar, and both rely on the norm $\|\cdot\|$ (and its associated dual $\|\cdot\|_*$) with respect to which ψ is strongly convex. Mirror-descent type first order algorithms, such as projected gradient methods, attain regret bounds of the form (Zinkevich, 2003; Bartlett et al., 2007; Duchi et al., 2010)

$$R_\phi(T) \le \frac{1}{\eta} B_\psi(x^*, x_1) + \frac{\eta}{2} \sum_{t=1}^{T} \|f_t'(x_t)\|_*^2 \ . \tag{7}$$

Choosing $\eta \propto 1/\sqrt{T}$ gives $R_\phi(T) = O(\sqrt{T})$. When $B_\psi(x, x^*)$ is bounded for all $x \in X$, we choose step sizes $\eta_t \propto 1/\sqrt{t}$ which is equivalent to setting $\psi_t(x) = \sqrt{t}\psi(x)$. Therefore, no assumption on the time horizon is necessary. For RDA and follow-the-leader algorithms, the bounds are similar (Xiao, 2009,

[1] In general Zinkevich's bound is tight and cannot be improved in the worst case (Abernethy et al., 2008a). Therefore we rely on specific reasonable data generating mechanisms for which our bounds are better.

[2] The proximal term is also referred to as regularization in the online learning literature. We use the phrase proximal term in order to avoid confusion with the statistical regularization function φ.

Theorem 3):

$$R_\phi(T) \le \sqrt{T}\psi(x^*) + \frac{1}{2\sqrt{T}} \sum_{t=1}^{T} \|f_t'(x_t)\|_*^2 \ . \tag{8}$$

The problem of adapting to data and obtaining tighter data-dependent bounds for algorithms such as those above is a natural one and has been studied extensively in the past.

The framework that is most related to ours is probably confidence weighted learning (Crammer et al., 2008) and the adaptive regularization of weights algorithm (AROW) of Crammer et al. (2009). These papers give a mistake-bound analysis for second-order algorithms for the Perceptron, which are similar in spirit to the second-order Perceptron itself (Cesa-Bianchi et al., 2005). Crammer et al. (2009) maintain an inverse covariance matrix $\Sigma_t^{-1} = I - \sum_{\tau=1}^{t} \beta_t z_\tau z_\tau^\top$, where z_t is the tth example and β_t is a multiplier depending on whether a mistake is made on round t. They update a mean prediction vector

$$\mu_{t+1} = \mu_t + \alpha_t \Sigma_t y_t z_t \ ,$$

where y_t is the label of example t and α_t is a function of the previous predictor's margin on example t. In contrast, the AdaGrad algorithm uses the *root* of the inverse covariance matrix, a choice which nicely matches the formal analysis. Crammer et al.'s algorithm and our algorithms have similar run times—generally linear in the dimension d—when using diagonal matrices. However, when using full matrices the runtime of their algorithm is $O(d^2)$, which is faster than ours.

There are also other lines of work on adaptivity less directly related to ours but nonetheless relevant. Tighter regret bounds using the variation of the cost functions were proposed in Cesa-Bianchi et al. (2007) and derived in Hazan and Kale (2008). Another adaptation technique for η_t was explored by Bartlett et al. (2007), who adapt the step size to accommodate both strongly and weakly convex functions.

Our approach differs from previous approaches in the following sense. Instead of focusing on a particular loss function or mistake bound, we view the problem of adapting the proximal function as an online (meta) learning problem and obtain a bound comparable to the bound obtained using the best proximal function chosen in hindsight.

2 Adaptive Proximal Functions

Examining the bounds in Eq. (7) and Eq. (8), we see that most of the regret depends on dual norms of $f_t'(x_t)$, and the dual norms in turn depend on the choice of ψ. This naturally leads to the question of whether we can modify the proximal term ψ along the run of the algorithm in order to lower the contribution of the aforementioned norms. We achieve this goal by keeping second order information about the sequence f_t and allow ψ to vary on each round of the algorithms.

We begin by providing two corollaries based on previous work that give the regret of each of the base algorithms when the proximal function ψ_t is allowed to change. These corollaries are used in the sequel in our regret analysis. We assume that ψ_t is monotonically non-decreasing, that is, $\psi_{t+1}(x) \ge \psi_t(x)$. We also assume that ψ_t is 1-strongly convex with respect to a time-dependent seminorm $\|\cdot\|_{\psi_t}$. Formally, ψ is 1-strongly convex with respect to $\|\cdot\|_\psi$ if

$$\psi(y) \ge \psi(x) + \langle \nabla\psi(x), y - x \rangle + \frac{1}{2} \|x - y\|_\psi^2 \ .$$

Recall that strong convexity is guaranteed if and only if $B_{\psi_t}(x, y) \ge \frac{1}{2} \|x - y\|_{\psi_t}^2$. We also denote the dual norm of $\|\cdot\|_{\psi_t}$ by $\|\cdot\|_{\psi_t^*}$. For completeness, we provide the proofs of following two corollaries in Appendix B, though they build straightforwardly on Duchi et al. (2010) and Xiao (2009). For the primal-dual subgradient update, the following regret bound holds.

Corollary 3. *Let the sequence $\{x_t\}$ be defined by the update in Eq. (2). Then for any x^*, we have*

$$R_\phi(T) \le \frac{1}{\eta}\psi_T(x^*) + \frac{\eta}{2} \sum_{t=1}^{T} \|f_t'(x_t)\|_{\psi_{t-1}^*}^2 \ . \tag{9}$$

Algorithm 1 AdaGrad with Diagonal Matrices

```
Input: $\eta > 0$, $\delta \in (0,1)$
Variables: $s \in \mathbb{R}^d$, $H \in \mathbb{R}^{d \times d}$, $g_{1:t,i} \in \mathbb{R}^t$ for $i \in \{1, \dots, d\}$
Initialize $x_1 = 0$, $g_{1:0} = []$
for $t = 1$ to $T$ do
    Suffer loss $f_t(x_t)$
    Receive subgradient $g_t \in \partial f_t(x_t)$ of $f_t$ at $x_t$
    Update $g_{1:t} = [g_{1:t-1} \; g_t]$, $s_{t,i} = \|g_{1:t,i}\|_2$
    Set $H_t = \delta I + \text{diag}(s_t)$, $\psi_t(x) = \frac{1}{2} \langle x, H_t x \rangle$
    Primal-Dual Subgradient Update (Eq. (2)):
        $x_{t+1} = \underset{x \in X}{\text{argmin}} \left\{ \eta \left\langle \frac{1}{t} \sum_{\tau=1}^{t} g_\tau, x \right\rangle + \eta \varphi(x) + \frac{1}{t} \psi_t(x) \right\}.$
    Composite Mirror Descent Update (Eq. (3)):
        $x_{t+1} = \underset{x \in X}{\text{argmin}} \left\{ \eta \langle g_t, x \rangle + \eta \varphi(x) + B_{\psi_t}(x, x_t) \right\}.$
end for
```

For composite mirror descent algorithms we have a similar corollary.

Corollary 4. *Let the sequence $\{x_t\}$ be defined by the update in Eq. (3). Then for any x^*, we have*

$$R_\psi(T) \le \frac{1}{\eta} B_{\psi_1}(x^*, x_1) + \varphi(x_1) + \frac{1}{\eta} \sum_{t=1}^{T-1} \left[B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1}) \right] + \frac{\eta}{2} \sum_{t=1}^{T} \|f_t'(x_t)\|_{\psi_t^*}^2 \; . \qquad (10)$$

The above corollaries allow us to prove regret bounds of a family of algorithms that iteratively modify the proximal functions ψ_t in attempt to lower the regret bounds.

3 Diagonal Matrix Proximal Functions

We begin by restricting ourselves to using diagonal matrices to define matrix proximal functions and (semi)norms. This restriction serves a two-fold purpose. First, the analysis for the general case is somewhat complicated and thus the analysis of the diagonal restriction serves as a proxy for better understanding. Second, in problems with high dimension where we expect this type of modification to help, maintaining more complicated proximal functions is likely to be prohibitively expensive. Whereas earlier analysis requires a learning rate to slow changes between predictors x_t and x_{t+1}, we will instead automatically grow the proximal function we use to achieve asymptotically low regret. To remind the reader, $g_{1:t,i}$ is the ith row of the matrix obtained by concatenating the subgradients from iteration 1 through t in the online algorithm.

To provide some intuition for the algorithm we show in Alg. 1, let us examine the problem

$$\min_s \; \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{g_{t,i}^2}{s_i} \quad \text{s.t. } s \succeq 0, \;\; \langle 1, s \rangle \le c \; .$$

This problem is solved by setting $s_i = \|g_{1:T,i}\|_2$ and scaling s so that $\langle s, 1 \rangle = c$. To see this, we can write the Lagrangian of the minimization problem by introducing multipliers $\lambda \succeq 0$ and $\theta \ge 0$ to get

$$\mathcal{L}(s, \lambda, \theta) = \sum_{i=1}^{d} \frac{\|g_{1:T,i}\|_2^2}{s_i} - \langle \lambda, s \rangle + \theta(\langle 1, s \rangle - c).$$

Taking partial derivatives to find the infimum of $\mathcal{L}$, we see that $-\left\|g_{1:T,i}\right\|_2^2/s_i^2 - \lambda_i + \theta = 0$, and complimentarity conditions on $\lambda_i s_i$ (Boyd and Vandenberghe, 2004) imply that $\lambda_i = 0$. Thus we have $s_i = \theta^{-\frac{1}{2}} \left\|g_{1:T,i}\right\|_2$, and normalizing appropriately using θ gives that $s_i = c\left\|g_{1:T,i}\right\|_2 / \sum_{j=1}^d \left\|g_{1:T,j}\right\|_2$. As one final note, we can plug s_i in to the above to see that

$$\inf_s \left\{ \sum_{t=1}^T \sum_{i=1}^d \frac{g_{t,i}^2}{s_i} \ : \ s \succeq 0, \langle 1, s\rangle \le c \right\} = \frac{1}{c}\left(\sum_{i=1}^d \left\|g_{1:T,i}\right\|_2\right)^2 . \tag{11}$$

Letting $\operatorname{diag}(v)$ denote the diagonal matrix with diagonal v, it is natural to suspect that if we use a proximal function similar to $\psi(x) = \langle x, \operatorname{diag}(s)x\rangle$ with associated squared dual norm $\|x\|_{\psi^*}^2 = \left\langle x, \operatorname{diag}(s)^{-1}x\right\rangle$, we should do well lowering the gradient terms in the regret in Eq. (9) and Eq. (10).

To prove a regret bound for our Alg. 1, we note that both types of updates suffer losses which include a term depending solely on the gradients obtained along their run. Thus, the following lemma is applicable to both updates.

Lemma 5. *Let $g_t = f_t'(x_t)$ and $g_{1:t}$ and s_t be defined as in Alg. 1, then*

$$\sum_{t=1}^T \left\langle g_t, \operatorname{diag}(s_t)^{-1} g_t\right\rangle \le 2\sum_{i=1}^d \left\|g_{1:T,i}\right\|_2 .$$

Proof. We prove the lemma by considering an arbitrary $\mathbb{R}$-valued sequence $\{a_i\}$ and its vector representation $a_{1:i} = [a_1 \ \cdots \ a_i]$. We are going to show that (where we set $0/0 = 0$)

$$\sum_{t=1}^T \frac{a_t^2}{\left\|a_{1:t}\right\|_2} \le 2\left\|a_{1:T}\right\|_2 . \tag{12}$$

We use induction on T to prove Eq. (12). For $T = 1$, the inequality trivially holds. Assume Eq. (12) holds true for $T-1$, then

$$\sum_{t=1}^T \frac{a_t^2}{\left\|a_{1:t}\right\|_2} = \sum_{t=1}^{T-1} \frac{a_t^2}{\left\|a_{1:t}\right\|_2} + \frac{a_T^2}{\left\|a_{1:T}\right\|_2} \le 2\left\|a_{1:T-1}\right\|_2 + \frac{a_T^2}{\left\|a_{1:T}\right\|_2} ,$$

where the inequality follows from the inductive hypothesis. We now define $b_T = \sum_{t=1}^T a_t^2$ and use first-order inequality for concavity to obtain that so long as $b_T - a_T^2 \ge 0$, $\sqrt{b_T - a_T^2} \le \sqrt{b} - a_T^2 \frac{1}{2\sqrt{b_T}}$ (as a forward pointer, we note that we use an identical technique in the full-matrix case; see Lemma 19 in the appendix). Thus we have

$$2\left\|a_{1:T-1}\right\|_2 + \frac{a_T^2}{\left\|a_{1:T}\right\|_2} = 2\sqrt{b_T - a_T^2} + \frac{a_T^2}{\sqrt{b_T}} \le 2\sqrt{b_T} = 2\left\|a_{1:T}\right\|_2 .$$

Having proved Eq. (12), we note that by construction $s_{t,i} = \left\|g_{1:t,i}\right\|_2$, thus,

$$\sum_{t=1}^T \left\langle g_t, \operatorname{diag}(s_t)^{-1} g_t\right\rangle = \sum_{t=1}^T \sum_{i=1}^d \frac{g_{t,i}^2}{\left\|g_{1:t,i}\right\|_2} \le 2\sum_{i=1}^d \left\|g_{1:T,i}\right\|_2 . \qquad \square$$

To get a regret bound, we need to consider the terms consisting of the dual-norm of the subgradient in Eq. (9) and Eq. (10), $\|f_t'(x_t)\|_{\psi_t^*}^2$. When $\psi_t(x) = \langle x, (\delta I + \operatorname{diag}(s_t))x\rangle$, it is easy to see that the associated dual-norm is

$$\|g\|_{\psi_t^*}^2 = \left\langle g, (\delta I + \operatorname{diag}(s_t))^{-1} g\right\rangle .$$

From the definition of s_t in Alg. 1, we clearly have $\|f_t'(x_t)\|_{\psi_t^*}^2 \le \left\langle g_t, \operatorname{diag}(s_t)^{-1} g_t\right\rangle$. Note that we replace the inverse with a pseudo-inverse if needed, which is well defined since g_t is always in the column-space of $\operatorname{diag}(s_t)$. Thus, Lemma 5 gives

$$\sum_{t=1}^T \|f_t'(x_t)\|_{\psi_t^*}^2 \le 2\sum_{i=1}^d \left\|g_{1:T,i}\right\|_2 .$$

To obtain a bound for a primal-dual subgradient method, we set $\delta \ge \max_t \|g_t\|_\infty$, in which case $\|g_t\|_{\psi_{t-1}^*}^2 \le \left\langle g_t, \operatorname{diag}(s_t)^{-1} g_t \right\rangle$, and we follow the same lines of reasoning.

It remains to bound the various Bregman divergence terms for Corollary 4 and the term $\psi_T(x^*)$ for Corollary 3. We focus first on composite mirror-descent updates. Examing Eq. (10) and Alg. 1, we notice that

$$\begin{aligned} B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1}) &= \frac{1}{2} \langle x^* - x_{t+1}, \operatorname{diag}(s_{t+1} - s_t)(x^* - x_{t+1}) \rangle \\ &\le \frac{1}{2} \max_i (x_i^* - x_{t+1,i})^2 \|s_{t+1} - s_t\|_1 . \end{aligned}$$

Since $\|s_{t+1} - s_t\|_1 = \langle s_{t+1} - s_t, 1 \rangle$ and $\langle s_T, 1 \rangle = \sum_{i=1}^d \|g_{1:T,i}\|_2$, we have

$$\begin{aligned} \sum_{t=1}^{T-1} B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1}) &\le \frac{1}{2} \sum_{t=1}^{T-1} \|x^* - x_{t+1}\|_\infty^2 \langle s_{t+1} - s_t, 1 \rangle \\ &\le \frac{1}{2} \max_{t \le T} \|x^* - x_t\|_\infty^2 \sum_{i=1}^d \|g_{1:T,i}\|_2 - \frac{1}{2} \|x^* - x_1\|_\infty^2 \langle s_1, 1 \rangle \ . \end{aligned} \tag{13}$$

We also have

$$\psi_T(x^*) = \delta \|x^*\|_2^2 + \langle x^*, \operatorname{diag}(s_T) x^* \rangle \le \delta \|x^*\|_2^2 + \|x^*\|_\infty^2 \sum_{i=1}^d \|g_{1:T,i}\|_2 .$$

Combining the above arguments with Corollaries 3 and 4, and combining Eq. (13) with the fact that $B_{\psi_1}(x^*, x_1) \le \frac{1}{2} \|x^* - x_1\|_\infty^2 \langle 1, s_1 \rangle$, we have proved the following theorem.

Theorem 6. *Let the sequence $\{x_t\}$ be defined by Algorithm 1. If we generate x_t using the primal-dual subgradient update of Eq. (2) and $\delta \ge \max_t \|g_t\|_\infty$, then for any $x^* \in X$ we have*

$$R_\phi(T) \le \frac{\delta}{\eta} \|x^*\|_2^2 + \frac{1}{\eta} \|x^*\|_\infty^2 \sum_{i=1}^d \|g_{1:T,i}\|_2 + \eta \sum_{i=1}^d \|g_{1:T,i}\|_2 . \tag{14}$$

If we use Algorithm 1 with the composite mirror-descent update of Eq. (3), then for any $x^* \in X$

$$R_\phi(T) \le \varphi(x_1) + \frac{1}{2\eta} \max_{t \le T} \|x^* - x_t\|_\infty^2 \sum_{i=1}^d \|g_{1:T,i}\|_2 + \eta \sum_{i=1}^d \|g_{1:T,i}\|_2 . \tag{15}$$

The above theorem is a bit unwieldy. We thus perform a few algebraic simplifications to get the next corollary. Let us assume that X is compact and set $D_\infty = \sup_{x \in X} \|x - x^*\|_\infty$. Furthermore, let

$$\gamma_T = \sum_{i=1}^d \|g_{1:T,i}\|_2 = \sqrt{\inf_s \left\{ \sum_{t=1}^T \langle g_t, \operatorname{diag}(s)^{-1} g_t \rangle : \langle 1, s \rangle \le \sum_{i=1}^d \|g_{1:T,i}\|_2 \, , \ s \succeq 0 \right\}} \, .$$

Also w.l.o.g. let $0 \in X$. The following corollary is immediate.

Corollary 7. *Assume that $\varphi(x_1) = 0$ and D_∞ and γ_T are defined as above. If we generate the sequence $\{x_t\}$ be given by Algorithm 1 using the primal-dual subradient update Eq. (2) with $\eta = \|x^*\|_\infty$, then for any $x^* \in X$ we have*

$$R_\phi(T) \le 2 \|x^*\|_\infty \gamma_T + \delta \frac{\|x^*\|_2^2}{\|x^*\|_\infty} \le 2 \|x^*\|_\infty \gamma_T + \delta \|x^*\|_1 \ .$$

Using the composite mirror descent update of Eq. (2) to generate $\{x_t\}$ and setting $\eta = D_\infty / \sqrt{2}$, we have

$$R_\phi(T) \le \sqrt{2} D_\infty \gamma_T \ .$$

Algorithm 2 AdaGrad with Full Matrices

```
Variables $x \in \mathbb{R}^d$, $S_t \in \mathbb{R}^{d\times d}$, $H_t \in \mathbb{R}^{d\times d}$, $G_t \in \mathbb{R}^{d\times d}$
Initialize $x = 0$, $S_0 = 0$, $H_0 = 0$, $G_0 = 0$
for $t = 1$ to $T$ do
    Suffer loss $f_t(x_t)$
    Recieve subgradient $g_t \in \partial f_t(x_t)$ of $f_t$ at $x_t$.
    Update $G_t = G_{t-1} + g_t g_t^\top$, $S_t = G_t^{\frac{1}{2}}$.
    Let $H_t = \delta I + S_t$, $\psi_t(x) = \frac{1}{2}\langle x, H_t x\rangle$
    Primal-Dual Subgradient Update (Eq. (2))
        $x_{t+1} = \operatorname*{argmin}_{x\in X}\left\{\eta\left\langle\frac{1}{t}\sum_{\tau=1}^{t} g_t, x\right\rangle + \eta\varphi(x) + \frac{1}{t}\psi_t(x)\right\}$
    Composite Mirror Descent Update (Eq. (3))
        $x_{t+1} = \operatorname*{argmin}_{x\in X}\left\{\eta\langle g_t, x\rangle + \eta\varphi(x) + B_{\psi_t}(x, x_t)\right\}$
end for
```

Intuitively, as discussed in the introduction, Alg. 1 should have good properties on sparse data. For example, suppose that our gradient terms are based on 0/1-valued features for a logistic regression task. Then the gradient terms in the bound $\sum_{i=1}^{d}\|g_{1:t,i}\|_2$ should all be much smaller than $\sqrt{T}$. If we assume that some features appear much more frequently than others, then the infimal representation of γ_T and the infimal equality in Corollary 2 show that we can have much lower learning rates on commonly appearing features and higher rates on uncommon features, and this will significantly lower the bound on the regret. Further, if we are constructing a relatively dense predictor x as is often the case in sparse prediction problems, then $\|x^*\|_\infty$ is the best p-norm we can have in the regret.

4 Full Matrix Proximal Functions

In this section we derive and analyze new updates when we estimate a full matrix for the divergence ψ_t instead of diagonal ones. In this generalized case, we use the the square-root of the matrix of outer products of the gradients that we have observed to update our parameters. As in the diagonal case, we build on intuition garnered from the following constrained optimization problem. We seek a matrix S which is the solution to the following minimization problem

$$\min_S \sum_{t=1}^{T}\left\langle g_t, S^{-1} g_t\right\rangle \quad \text{s.t. } S \succeq 0, \;\; \mathrm{tr}(S) \le c\ .$$

The solution is obtained by defining $G_t = \sum_{\tau=1}^{t} g_\tau g_\tau^\top$, and then setting S to be a normalized version of the root of G_T, that is, $S = c\, G_T^{1/2}/\,\mathrm{tr}(G_T^{1/2})$. For a proof, see Lemma 21 in Appendix C, which also shows that when G_T is not full rank we can instead use its pseudo-inverse. If we iteratively use divergences of the form $\psi_t(x) = \left\langle x, G_t^{1/2} x\right\rangle$, we might expect as earlier to attain low regret and collect gradient information. We can achieve our low regret goal by employing a similar doubling lemma to Lemma 5 and bounding the gradient norm terms. The resulting algorithm is given in Alg. 2, and the next theorem provides a quantative analysis of the brief motivation above.

Theorem 8. *Let G_t be the outer product matrix defined above and the sequence $\{x_t\}$ be defined by Algorithm 2. If we generate x_t using the primal-dual subgradient update of Eq. (2) and $\delta \ge \max_t \|g_t\|_2$, then for any $x^* \in X$ we have*

$$R_\phi(T) \le \frac{\delta}{\eta}\|x^*\|_2^2 + \frac{1}{\eta}\|x^*\|_2^2\,\mathrm{tr}(G_T^{1/2}) + \eta\,\mathrm{tr}(G_T^{1/2}). \tag{16}$$

If we use Algorithm 2 with the composite mirror-descent update of Eq. (3), then for any x^* and $\delta \geq 0$

$$R_\phi(T) \leq \frac{\delta}{\eta}\left\|x^*\right\|_2^2 + \varphi(x^1) + \frac{1}{2\eta}\max_{t\leq T}\left\|x^* - x_t\right\|_2^2 \operatorname{tr}(G_T^{1/2}) + \eta \operatorname{tr}(G_T^{1/2}). \tag{17}$$

Proof. To begin, we consider the difference between the divergence terms at time $t+1$ and time t from Eq. (10) in Corollary 4. Let $\lambda_{\max}(M)$ denote the largest eigenvalue of a matrix M. We have

$$\begin{aligned} B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1}) \; &= \; \frac{1}{2}\left\langle x^* - x_{t+1}, ({G_{t+1}}^{1/2} - {G_t}^{1/2})(x^* - x_{t+1})\right\rangle \\ \leq \; \frac{1}{2}\left\|x^* - x_{t+1}\right\|_2^2 \lambda_{\max}(G_{t+1}^{1/2} - G_t^{1/2}) \; &\leq \; \frac{1}{2}\left\|x^* - x_{t+1}\right\|_2^2 \operatorname{tr}(G_{t+1}^{1/2} - G_t^{1/2}) \; . \end{aligned}$$

For the last inequality we used the fact that the trace of a matrix is equal to the sum of its eigenvalues along with the property ${G_{t+1}}^{1/2} - {G_t}^{1/2} \succeq 0$ (see Lemma 17) and therefore $\operatorname{tr}(G_{t+1}^{1/2} - G_t^{1/2}) \geq \lambda_{\max}(G_{t+1}^{1/2} - G_t^{1/2})$. Thus, we get

$$\begin{aligned} \sum_{t=1}^{T-1} B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1}) \; &\leq \; \frac{1}{2}\sum_{t=1}^{T-1}\left\|x^* - x_{t+1}\right\|_2^2 \left(\operatorname{tr}(G_{t+1}^{1/2}) - \operatorname{tr}(G_t^{1/2})\right) \\ &\leq \; \frac{1}{2}\max_{t\leq T}\left\|x^* - x_t\right\|_2^2 \operatorname{tr}({G_T}^{1/2}) - \frac{1}{2}\left\|x^* - x_1\right\|_2^2 \operatorname{tr}(G_1^{1/2}) \; . \end{aligned} \tag{18}$$

For the last inequality we used the fact that G_1 is a rank 1 PSD matrix with non-negative trace. What remains is to bound the gradient terms common to all our bounds. The following lemma is directly applicable.

Lemma 9. *Let $S_t = {G_t}^{1/2}$ be as defined in Alg. 2. Then, using the pseudo-inverse when necessary,*

$$\sum_{t=1}^{T}\left\langle g_t, S_t^{-1} g_t\right\rangle \leq 2\sum_{t=1}^{T}\left\langle g_t, S_T^{-1} g_t\right\rangle = 2\operatorname{tr}({G_T}^{1/2}) \; .$$

Proof. We prove the lemma by induction. The base case is immediate, since we have

$$\left\langle g_1, G_1^{-1/2} g_1\right\rangle = \frac{\langle g_1, g_1\rangle}{\left\|g_1\right\|_2} = \left\|g_1\right\|_2 \leq 2\left\|g_1\right\|_2 .$$

Now, assume the lemma is true for $T-1$, so from the inductive assumption we get

$$\sum_{t=1}^{T}\left\langle g_t, S_t^{-1} g_t\right\rangle \quad \leq \quad 2\sum_{t=1}^{T-1}\left\langle g_t, S_{T-1}^{-1} g_t\right\rangle + \left\langle g_T, S_T^{-1} g_T\right\rangle \; .$$

Since S_{T-1} does not depend on t we can rewrite $\sum_{t=1}^{T-1}\left\langle g_t, S_{T-1}^{-1} g_t\right\rangle$ as

$$\operatorname{tr}\left(S_{T-1}^{-1}, \sum_{t=1}^{T-1} g_t g_t^\top\right) = \operatorname{tr}(G_{T-1}^{-1/2} G_{T-1}) \; ,$$

where the right-most equality follows from the definitions of S_t and G_t. Therefore, we get that

$$\begin{aligned} \sum_{t=1}^{T}\left\langle g_t, S_t^{-1} g_t\right\rangle \quad &\leq \quad 2\operatorname{tr}(G_{T-1}^{-1/2} G_{T-1}) + \left\langle g_T, G_T^{-1/2} g_T\right\rangle \\ &= \quad 2\operatorname{tr}(G_{T-1}^{1/2}) + \left\langle g_T, G_T^{-1/2} g_T\right\rangle \; . \end{aligned}$$

Finally, Lemma 19 in the appendix, which also justifies the use of pseudo-inverses, lets us exploit the concavity of the function $\operatorname{tr}(A^{1/2})$ to bound the above sum by $2\operatorname{tr}(G_T^{1/2})$, which yields the proof. △

We can now finalize our proof of the theorem. As in the diagonal case, we have that the squared dual norm (seminorm when $\delta = 0$) associated with ψ_t is

$$\|x\|_{\psi_t^*}^2 = \left\langle x, (\delta I + S_t)^{-1} x \right\rangle \ .$$

Thus it is clear that $\|g_t\|_{\psi_t^*}^2 \le \left\langle g_t, S_t^{-1} g_t \right\rangle$. For the dual-ascent algorithms, we use Lemma 20 from the appendix to see that $\|g_t\|_{\psi_{t-1}^*}^2 \le \left\langle g_t, S_t^{-1} g_t \right\rangle$ so long as $\delta \ge \|g_t\|_2$. The doubling inequality from Lemma 9 implies that $\sum_{t=1}^T \|f_t'(x_t)\|_{\psi_t^*}^2 \le 2\,\mathrm{tr}(G_T^{1/2})$ for the mirror-descent algorithms and that $\sum_{t=1}^T \|f_t'(x_t)\|_{\psi_{t-1}^*}^2 \le 2\,\mathrm{tr}(G_T^{1/2})$ for primal-dual subgradient algorithms.

Note that $B_{\psi_1}(x^*, x_1) \le \frac{1}{2} \|x^* - x_1\|_2^2 \,\mathrm{tr}(G_1^{1/2})$ when $\delta = 0$. Combining the first of the last bounds in the previous paragraph with this and the bound on $\sum_{t=1}^{T-1} B_{\psi_{t+1}}(x^*, x^{t+1}) - B_{\psi_t}(x^*, x^{t+1})$ from Eq. (18), we see that Corollary 4 gives the bound for the mirror-descent family of algorithms. Combining the second of the bounds in the previous paragraph and Eq. (18) with Corollary 3 gives the desired bound on $R_\phi(T)$ for the primal-dual subgradient algorithms, which completes the proof of the theorem. □

As before, we can give a corollary that clarifies the bound implied by Theorem 8. The second bound in the corollary hinges on the assumption that the set X is compact and uses Lemma 21 in Appendix C for the equality.

Corollary 10. *Assume further that $\varphi(x_1) = 0$. Then, the regret of the sequence $\{x_t\}$ generated by Algorithm 2 when using the primal-dual subgradient update with $\eta = \|x^*\|_2$ is*

$$R_\phi(T) \le 2\|x^*\|_2 \,\mathrm{tr}(G_T^{1/2}) + \delta \|x^*\|_2 \ .$$

Let X be compact set so that $\sup_{x \in X} \|x - x^*\|_2 \le D$. Taking $\eta = D/\sqrt{2}$ and using the composite mirror descent update with $\delta = 0$, we have

$$R_\phi(T) \le \sqrt{2} D \,\mathrm{tr}(G_T^{1/2}) = \sqrt{2d} D \sqrt{\inf_S \left\{ \sum_{t=1}^T g_t^\top S^{-1} g_t \ : \ S \succeq 0, \mathrm{tr}(S) \le d \right\}} \ .$$

5 Lowering the Regret for Strongly Convex Functions

It is now well established that strong convexity of the functions f_t can give significant improvements in the regret of online (or stochastic) convex optimization algorithms (Hazan et al., 2006; Shalev-Shwartz and Singer, 2007). We can likewise derive lower regret bounds in the presence of strong convexity

Let us now assume that our functions $f_t + \varphi$ are strongly convex with respect to a norm $\|\cdot\|$. For simplicity, we assume that each has the same strong convexity parameter λ, that is,

$$f_t(y) + \varphi(y) \ge f_t(x) + \varphi(x) + \langle f_t'(x), y - x \rangle + \langle \varphi'(x), y - x \rangle + \frac{\lambda}{2} \|x - y\|^2 .$$

We focus on composite mirror descent algorithms here, as the analysis of strongly convex variants of primal-dual subgradient algorithms does not seem to lend itself to dynamic learning rate adaptation.[3] For composite mirror descent, we obtain a simple corollary by combining the original convergence results from Corollary 4 with the strongly-convex function results from Duchi et al. (2010). We assume without loss of generality that $\varphi(x_1) = 0$ and $x_1 = 0$.

Corollary 11. *Let the sequence $\{x_t\}$ be generated by the update given in Eq. (3) and assume that φ is λ-strongly convex with respect to a norm $\|\cdot\|$. Then for any $x^* \in X$,*

$$R_\phi(T) \le c + \frac{1}{\eta} \sum_{t=1}^{T-1} \left[B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1}) - \frac{\lambda\eta}{2} \|x^* - x_{t+1}\|^2 \right] + \frac{\eta}{2} \sum_{t=1}^T \|f_t'(x_t)\|_{\psi_t^*}^2$$

[3] This is a consequence of the analysis of the strongly-convex primal-dual method, which keeps the function ψ constant rather than letting it grow at a rate of $\sqrt{t}$ as in standard RDA. Allowing ψ to grow breaks the stronger regret. It may be possible to give an analysis for RDA in which the *regularization* function φ is time-dependent, but this is outside the scope of this work.

where $c = (1/\eta) B_{\psi_1}(x^*, x_1) - (\lambda\eta/2) \left\| x^* - x_1 \right\|^2$.

Based on the above corollary, we can derive a logarithmic regret algorithm for strongly convex losses $f_t + \varphi$. Such a bound when using full matrix information was derived by Hazan et al. (2006). We thus focus on the diagonal matrix case. In this case, we let ψ_t grow somewhat faster than when $f_t + \varphi$ are merely convex, which follows the ideas from prior logarithmic regret algorithms with faster step rates. The algorithm is identical to Alg. 1 except that the proximal function ψ_t grows more quickly.

We now assume for simplicity that φ is λ-strongly convex with respect to the 2-norm. We let $s_{t,i} = \left\| g_{1:t,i} \right\|_2^2$. Then, setting $\psi_t(x) = \frac{1}{2} \langle x, (\delta I + \operatorname{diag}(s_t)), x \rangle$ and $\eta \geq \frac{1}{\lambda} \max_t \left\| g_t \right\|_\infty^2$, we have

$$\begin{aligned} & B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1}) - \frac{\lambda\eta}{2} \left\| x^* - x_{t+1} \right\|_p^2 \\ & \quad = \quad \frac{1}{2} \left\langle x^* - x_{t+1}, \operatorname{diag}\left(g_{t,i}^2 \right) (x^* - x_{t+1}) \right\rangle - \frac{\lambda\eta}{2} \left\| x^* - x_{t+1} \right\|_p^2 \\ & \quad \leq \quad \frac{1}{2} \left\| g_t \right\|_\infty^2 \left\| x^* - x_{t+1} \right\|_2^2 - \frac{\lambda}{2\lambda} \max_t \left\| g_t \right\|_\infty^2 \left\| x^* - x_{t+1} \right\|_2^2 \;\leq\; 0 \;, \end{aligned}$$

where $\operatorname{diag}(g_{t,i}^2))$ denotes the diagonal matrix whose ith diagonal element is $g_{t,i}^2$. The only term remaining in the regret bound from Corollary 11 is a constant term and the gradient terms. We bound these terms using the following lemma.

Lemma 12. *Let $\{x_t\}$ be the sequence of vectors generated by Algorithm 1 with mirror-descent updates using the proximal function $\psi_t(x) = \langle x, (\delta I + \operatorname{diag}(s_t)) x \rangle$ with $s_{t,i} = \left\| g_{1:t,i} \right\|_2^2$. Then*

$$\sum_{t=1}^{T} \left\| f_t'(x_t) \right\|_{\psi_t^*}^2 \leq \sum_{i=1}^{d} \log \left(\frac{\left\| g_{1:T,i} \right\|_2^2}{\delta} + 1 \right) .$$

Proof. We begin by noting that for $a, b > 0$, the concavity oflog$(\cdot)$ implies that $\log(b) \leq \log(a) + \frac{1}{a}(b - a)$ and therefore $\frac{1}{a}(a - b) \leq \log \frac{a}{b}$. Now, consider a sequence $a_i \geq 0$ and define $v_i = a_0 + \sum_{j=1}^{i} a_j$ with $a_0 > 0$. We have

$$\sum_{i=1}^{n} \frac{a_i}{v_i} = \sum_{i=1}^{n} \frac{1}{v_i}(v_i - v_{i-1}) \leq \sum_{i=1}^{n} \log \frac{v_i}{v_{i-1}} = \log \frac{v_n}{v_0} = \log \frac{a_0 + \sum_{i=1}^{n} a_i}{a_0} \;.$$

Recalling the definition of ψ_t so $\left\| x \right\|_{\psi_t^*}^2 = \left\langle x, (\delta I + \operatorname{diag}(s_t))^{-1} x \right\rangle$ and the fact that this term is separable we get

$$\sum_{t=1}^{T} \left\| g_t \right\|_{\psi_t^*}^2 = \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{(g_{t,i})^2}{\delta + \left\| g_{1:t,i} \right\|_2^2} \leq \sum_{i=1}^{d} \log \frac{\left\| g_{1:T,i} \right\|_2^2 + \delta}{\delta} \;. \qquad \square$$

Based on the above lemma, we immediately obtain the following theorem.

Theorem 13. *Assume that φ is λ-strongly convex with respect to a p-norm with $p \geq 2$ over the set X. Assume further that $\left\| g \right\|_\infty \leq G_\infty$ for all $g \in \partial f_t(x)$ for $x \in X$. Let $\{x_t\}$ be the sequence of vectors generated by Algorithm 1 with the diagonal divergence ψ_t used in Lemma 12. Setting $\eta = \frac{G_\infty^2}{\lambda}$, we have*

$$R_\phi(T) \leq \frac{2 G_\infty^2 \delta}{\lambda} \left\| x_1 - x^* \right\|_2^2 + \frac{G_\infty^2}{\lambda} \sum_{i=1}^{d} \log \left(\frac{\left\| g_{1:T,i} \right\|_2^2}{\delta} + 1 \right) = O\left(\frac{d G_\infty^2}{\lambda} \log(T G_\infty) \right) \;.$$

6 Derived Algorithms

In this section, we derive updates using concrete regularization functions with the diagonal matrix version of the algorithms we have presented. We focus on the diagonal case for two reasons. First, the updates take closed-form in this case and carry some intuition. Second, we expect the diagonal case to be feasible to implement in very high dimensions whereas the full version is likely to be confined to a few thousand dimensions. We also discuss how to efficiently compute the updates when the gradient vectors are sparse.

ℓ_1-regularization We begin by considering how to solve the minimization problems necessary for Alg. 1 with a diagonal diagonal matrix divergences and $\varphi(x) = \lambda \left\| x \right\|_1$. We consider the two updates we proposed and denote the ith diagonal element of the matrix $H_t = \delta I + \mathrm{diag}(s_t)$ from Alg. 1 by $h_{t,i} = \delta + \left\| g_{1:t,i} \right\|_2$. For the primal-dual subgradient update, we need to solve Eq. (2), which amounts to the following:

$$\min_x \ \eta \left\langle \bar{g}_t, x \right\rangle + \frac{1}{2t}\delta \left\| x \right\|_2^2 + \frac{1}{2t} \left\langle x, \mathrm{diag}(s_t) x \right\rangle + \eta\lambda \left\| x \right\|_1 \ .$$

Let $\hat{x}$ denote the optimal solution of the above optimization problem. Standard subgradient calculus implies that when $|\bar{g}_{t,i}| \le \lambda$ the solution is $\hat{x}_i = 0$. Similarly, when $\bar{g}_{t,i} < -\lambda$, then $\hat{x}_i > 0$, the objective is differentiable, and the solution is obtained by setting the gradient to zero:

$$\eta \bar{g}_{t,i} + \frac{h_{t,i}}{t} \hat{x}_i + \eta\lambda = 0 \ , \quad \text{so that} \quad \hat{x}_i = \frac{\eta t}{h_{t,i}} \left(-\bar{g}_{t,i} - \lambda \right) \ .$$

Likewise, when $\bar{g}_{t,i} > \lambda$ then $\hat{x}_i < 0$, and the solution is $\hat{x}_i = \frac{\eta t}{h_{t,i}}(-\bar{g}_{t,i} + \lambda)$. Combining the three cases, we obtain the following simple update for $x_{t+1,i}$:

$$x_{t+1,i} = \mathrm{sign}\left(-\bar{g}_{t,i}\right) \frac{\eta t}{h_{t,i}} \left[|\bar{g}_{t,i}| - \lambda \right]_+ \ . \tag{19}$$

We can now compare Eq. (19) to the dual averaging update (Xiao, 2009) update, which amounts to

$$x_{t+1,i} = \mathrm{sign}\left(-\bar{g}_{t,i}\right) \eta \sqrt{t} \left[|\bar{g}_{t,i}| - \lambda \right]_+ \ . \tag{20}$$

The difference between Eq. (19) and Eq. (20) distills to the step size employed for each coordinate. For RDA, this step size is common to all coordinates and depends on the root of the number of iterations. Our generalization yields a dedicated step size for each coordinate which is inversely proportional to the time-based norm of the corresponding coordinate in the sequence of gradients. Moreover, due to the normalization by this term the step size scales *linearly* with t. The generalized update is likely to outperform RDA when features vary in their importance and dynamic range.

We now move our focus to the composite mirror-descent update Eq. (3). Using the same techniques of subgradient calculus and summarizing the different solutions by taking into account the sign of $x_{t+1,i}$ (see also Section 5 in Duchi and Singer, 2009), we get that

$$x_{t+1,i} = \mathrm{sign}\left(x_{t,i} - \frac{\eta}{h_{t,i}} g_{t,i} \right) \left[\left| x_{t,i} - \frac{\eta}{h_{t,i}} g_{t,i} \right| - \frac{\lambda\eta}{h_{t,i}} \right]_+ \ . \tag{21}$$

We compare the actual performance of the newly derived algorithms to previously studied versions in the next section.

For both updates it is clear that we can perform "lazy" computation when the gradient vectors are sparse, a frequently occurring setting when learning from text data. Suppose that from time step t_0 through t, the ith component of the gradient is 0. Then we can evaluate the above updates on demand since $h_{t,i}$ remains intact. At time t when $x_{t,i}$ is needed, we rewrite the composite mirror-descent update,

$$x_{t,i} = \mathrm{sign}(x_{t_0,i}) \left[|x_{t_0,i}| - \frac{\lambda\eta}{h_{t_0,i}} (t - t_0) \right]_+ \ .$$

Even simpler just in time evaluation can be performed for the the primal-dual subgradient update. Here we need to keep an unnormalized version of the average $\bar{g}_t$. Concretely, we keep track of $v_t = t\bar{g}_t = \sum_{\tau=1}^{t} g_\tau = v_{t-1} + g_t$ and rewrite the update in terms of v_t. We then use Eq. (19):

$$x_{t,i} = \mathrm{sign}(-v_{t,i}) \frac{\eta t}{h_{t,i}} \left[\frac{|v_{t,i}|}{t} - \lambda \right]_+ \ ,$$

where h_t can clearly be updated lazily in a similar fasion to v_t.

ℓ_2^2-regularization We now consider the case in which our regularizer $\varphi(x) = \frac{\lambda}{2} \|x\|_2^2$. This regularization is used for support vector machines and a voluminous number of learning algorithms have been developed for this setting. We focus on mirror-descent updates, as we have not derived an adaptive strongly convex primal-dual algorithm. When using the update of Eq. (3) we face the following optimization problem:

$$\min_{x \in X} \ \eta \langle g_t, x \rangle + \frac{\eta\lambda}{2} \|x\|_2^2 + \frac{\delta}{2} \|x\|_2^2 + \frac{1}{2} \left\langle (x - x_t), \operatorname{diag}(s^t)(x - x_t) \right\rangle \ .$$

Assuming for simplicity that $X = \mathbb{R}^d$, we compute the gradient of the above and equate it with the zero vector. We get

$$\eta g_t + \eta\lambda x + \delta x + \operatorname{diag}(s_t)(x - x_t) = 0 \ ,$$

whose solution is

$$x_{t+1} = ((\eta\lambda + \delta)I + \operatorname{diag}(s_t))^{-1} (\operatorname{diag}(s_t) x_t - \eta g_t) \ . \tag{22}$$

As in the case of ℓ_1 regularization, when the gradient vectors $g_1, \ldots, g_t, \ldots$ are sparse, we can evaluate the updates lazily by computing $x_{t,i}$ on demand with additional rather minor book keeping. Indeed, suppose that $g_{\tau,i} = 0$ for $\tau = t_0, \ldots, t$. Then to get $x_{t,i}$, we note that $\|g_{1:t_0,i}\|_2^2 = \|g_{1:t,i}\|_2^2$, so applying Eq. (22) $t - t_0$ times,

$$x_{t,i} = \frac{\|g_{1:t_0,i}\|_2^{2(t-t_0)}}{\left(\eta\lambda + \delta + \|g_{1:t_0,i}\|_2^2\right)^{t-t_0}} x_{t_0,i} \ .$$

In many prediction problems, the regularizer does not include the full prediction vector x because there is an unregularized bias term b. In this case, our analysis from Sec. 5 is still applicable, but we use the slower stepping for the bias term. That is, we use Eq. (22) to update the regularized predictors x, and if b is the bias term with associated gradients $g_{1:t,b}$, then $b_{t+1} = b_t - \frac{\eta}{\delta + \|g_{1:t,b}\|_2} g_{t,b}$.

7 Conclusions

We have presented a paradigm that adapts subgradient methods to the geometry of the problem at hand. The adaptation allows us to derive strong regret guarantees, which for some natural data distributions achieve better performance guarantees than previous algorithms. Our online convergence results can be naturally converted into rate of convergence and generalization bounds (Cesa-Bianchi et al., 2004). We believe that there are a few questions still unanswered in this line of work. The first is whether we can efficiently use full matrices in the proximal functions, as in Section 4, or whether a different algorithm is necessary. A second open issue is whether non-Euclidean proximal functions can be used. For example, is it possible to adapt the KL divergence to characteristics of the problem at hand? In our forthcoming companion paper, we give an expansive experimental evaluation of AdaGrad and other algorithms, showing that the algorithms derived in this paper are easy to implement, efficient, and significantly improve performance in real world domains.

A Proofs of Motivating Examples

Proof of Proposition 1. We begin with an integral approximation, noting that

$$\frac{2}{d}\left(\sqrt{dt+i} - \sqrt{i}\right) = \int_0^t \frac{1}{\sqrt{\tau d + i}} d\tau \le \sum_{\tau=0}^{t} \frac{1}{\sqrt{\tau d + i}} \le \frac{1}{\sqrt{i}} + \int_0^t \frac{1}{\sqrt{\tau d + i}} d\tau = \frac{1}{\sqrt{i}} + \frac{2}{d}\left(\sqrt{dt+i} - \sqrt{i}\right).$$

This implies that Eq. (5) is lower bounded by

$$d + \sum_{t=0}^{T} \sum_{i=1}^{d} \left[1 - \frac{2}{d}\left(\sqrt{dt+i} - \sqrt{i}\right) - \frac{1}{\sqrt{i}}\right]_+ .$$

By concavity of $\sqrt{\cdot}$,

$$\sqrt{dt+i} - \sqrt{i} \le \frac{1}{2\sqrt{i}} dt \quad \text{so} \quad 1 - \frac{2}{d}\left(\sqrt{dt+i} - \sqrt{i}\right) - \frac{1}{\sqrt{i}} \ge 1 - \frac{t}{\sqrt{i}} - \frac{1}{\sqrt{i}} = 1 - \frac{t+1}{\sqrt{i}}.$$

Assuming that $T \ge \sqrt{d} + 1$, we have

$$d + \sum_{t=0}^{T} \sum_{i=1}^{d} \left[1 - \sum_{\tau=0}^{t} \frac{1}{\sqrt{i + \tau d}}\right]_{+} \ge d + \sum_{i=1}^{d} \sum_{t=0}^{T} \left[1 - \frac{t+1}{\sqrt{i}}\right]_{+} \ge d + \sum_{i=1}^{d} \sum_{t=1}^{\sqrt{i}} \left(1 - \frac{t}{\sqrt{i}}\right) \ge d + \sum_{i=1}^{d} \frac{\sqrt{i}}{2}.$$

But $\sum_{i=1}^{d} \sqrt{i} \ge d\sqrt{d}/2$, so we have that Eq. (5) is lower bounded by $d + d\sqrt{d}/4$. □

Comparison of online gradient descent with AdaGrad in the full matrix case Zinkevich's projected gradient descent simply sets $x_{t+1} = x_t + \frac{1}{\sqrt{t}} v_t$ for $t = 1, \ldots, d$, as x_t remains in X. After iteration d, we are in a situation similar to the prequel and have suffered d losses in the first d rounds. In the remaining rounds, it is clear that $x_t = \alpha_{t,1} v_1 + \ldots + \alpha_{t,d} v_d = V\alpha_t$ for some multipliers $\alpha_t \succeq 0$. Since $\langle v_i, v_j \rangle = 0$ for $i \ne j$, $\sqrt{\sum_{i=1}^{d} \|v_i\|_2^2} = \sqrt{d}$, and projection simply shrinks the α multipliers for x_t, gradient descent suffers losses at least those of Eq. (5). For adaptive descent, for $t \le d$ we will construct the outer product matrix

$$G_t = \sum_{\tau=1}^{t} v_\tau v_\tau^\top \quad \text{with} \quad G_t^\dagger = \sum_{\tau=1}^{t} v_\tau v_\tau^\top, \quad G_t^{\frac{1}{2}} = G_t, \quad \text{and} \quad G_t^{-\frac{1}{2}} = G_t$$

since the v_i are orthonormal. This is easiest to see by an inductive argument. Clearly, $G_1 = G_1^\dagger = G_1^{\frac{1}{2}}$, since $g_1 = v_1$. To construct x_2, adaptive descent sets

$$x_2 = x_1 + G_1^\dagger v_1 = x_1 + v_1 v_1^\top v_1 = x_1 + v_1,$$

so $\langle x_2, z_2 \rangle = \pm \langle x_2, v_2 \rangle = 0$. Thus $G_2 = G_1 + v_2 v_2^\top = v_1 v_1^\top + v_2 v_2^\top$, the same argument applies, and the obvious inductive argument follows. Of course, we then have

$$x_{d+1} = \sum_{i=1}^{d} v_i \quad \text{and} \quad \|x_{d+1}\|_2 = \sqrt{\sum_{i=1}^{d} \|v_i\|_2^2} = \sqrt{d}$$

since the updates are all unconstrained because $X = \{x : \|x\|_2 \le \sqrt{d}\}$ is large enough. Then for adaptive descent, $\langle x_{d+1}, z_t \rangle = \langle 1, V^\top z_t \rangle = \left\langle \sum_{i=1}^{d} v_i, z_t \right\rangle$, which is the prediction achieving margin 1 and suffering no losses. Again, we see that the adaptive descent suffers only d losses, while projected gradient descent suffers $\Omega(d\sqrt{d})$.

B Proofs of Corollaries

Proof of Corollary 2. The proof corollary simply uses Theorem 6, Corollary 7, and the fact that

$$\inf_{s} \left\{ \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{g_{t,i}^2}{s_i} \ : \ s \succeq 0, \langle 1, s \rangle \le d \right\} = \frac{1}{d} \left(\sum_{i=1}^{d} \|g_{1:T,i}\|_2 \right)^2 .$$

as in Eq. (11) in the beginning of Sec. 3. Plugging the γ_T term in from Corollary 7 and multiplying D_∞ by $\sqrt{d}$ completes the proof of the corollary. □

In the remainder of this appendix we provide proofs for some of the more technical corollaries presented in the paper. We begin by stating an immediate corollary to Lemma 1 from Duchi et al. (2010). We provide the proof for completeness.

Corollary 14. *Let $\{x_t\}$ be the sequence defined by the update in Eq. (3) and assume that $B_{\psi_t}(\cdot,\cdot)$ is σ-strongly convex with respect to a norm $\|\cdot\|_{\psi_t}$. Let $\|\cdot\|_{\psi_t^*}$ be the associated dual norm. Then for any x^*,*

$$\eta\left(f_t(x_t) - f_t(x^*)\right) + \eta\left(\varphi(x_{t+1}) - \varphi(x^*)\right) \le B_{\psi_t}(x^*, x_t) - B_{\psi_t}(x^*, x_{t+1}) + \frac{\eta^2}{2\sigma}\left\|f'_t(x_t)\right\|^2_{\psi_t^*}$$

Proof. The optimality of x_{t+1} for Eq. (3) implies for all $x \in X$ and $\varphi'(x_{t+1}) \in \partial\varphi(x_{t+1})$ (Bertsekas, 1999)

$$\langle x - x_{t+1}, \eta f'(x_t) + \nabla\psi_t(x_{t+1}) - \nabla\psi_t(x_t) + \eta\varphi'(x_{t+1})\rangle \ge 0. \tag{23}$$

In particular, this obtains for $x = x^*$. From the subgradient inequality for convex functions, we have $f_t(x^*) \ge f_t(x_t) + \langle f'_t(x_t), x^* - x_t\rangle$, or $f_t(x_t) - f_t(x^*) \le \langle f'_t(x_t), x_t - x^*\rangle$, and likewise for $\varphi(x_{t+1})$. We thus have

$$\begin{aligned}
&\eta\left[f_t(x_t) + \varphi(x_{t+1}) - f_t(x^*) - \varphi(x^*)\right] \\
&\quad\le \eta\left\langle x_t - x^*, f'_t(x_t)\right\rangle + \eta\left\langle x_{t+1} - x^*, \varphi'(x_{t+1})\right\rangle \\
&\quad= \eta\left\langle x_{t+1} - x^*, f'_t(x_t)\right\rangle + \eta\left\langle x_{t+1} - x^*, \varphi'(x_{t+1})\right\rangle + \eta\left\langle x_t - x_{t+1}, f'_t(x_t)\right\rangle \\
&\quad= \left\langle x^* - x_{t+1}, \nabla\psi_t(x_t) - \nabla\psi_t(x_{t+1}) - \eta f'_t(x_t) - \eta\varphi'(x_{t+1})\right\rangle + \left\langle x^* - x_{t+1}, \nabla\psi_t(x_{t+1}) - \nabla\psi_t(x_t)\right\rangle \\
&\qquad+ \eta\left\langle x_t - x_{t+1}, f'_t(x_t)\right\rangle.
\end{aligned}$$

Now, by Eq. (23), the first term in the last equation is non-positive. Thus we have that

$$\begin{aligned}
&\eta\left[f_t(x_t) + \varphi(x_{t+1}) - f_t(x^*) - \varphi(x^*)\right] \\
&\quad\le \left\langle x^* - x_{t+1}, \nabla\psi_t(x_{t+1}) - \nabla\psi_t(x_t)\right\rangle + \eta\left\langle x_t - x_{t+1}, f'_t(x_t)\right\rangle \\
&\quad= B_{\psi_t}(x^*, x_t) - B_{\psi_t}(x_{t+1}, x_t) - B_{\psi_t}(x^*, x_{t+1}) + \eta\left\langle x_t - x_{t+1}, f'_t(x_t)\right\rangle \\
&\quad= B_{\psi_t}(x^*, x_t) - B_{\psi_t}(x_{t+1}, x_t) - B_{\psi_t}(x^*, x_{t+1}) + \eta\left\langle \sqrt{\frac{\sigma}{\eta}}(x_t - x_{t+1}), \sqrt{\frac{\eta}{\sigma}} f'_t(x_t)\right\rangle \\
&\quad\le B_{\psi_t}(x^*, x_t) - B_{\psi_t}(x_{t+1}, x_t) - B_{\psi_t}(x^*, x_{t+1}) + \frac{\sigma}{2}\left\|x_t - x_{t+1}\right\|^2_{\psi_t} + \frac{\eta^2}{2\sigma}\left\|f'_t(x_t)\right\|^2_{\psi_t^*} \\
&\quad\le B_{\psi_t}(x^*, x_t) - B_{\psi_t}(x^*, x_{t+1}) + \frac{\eta^2}{2\sigma}\left\|f'_t(x_t)\right\|^2_{\psi_t^*}.
\end{aligned}$$

In the above, the first equality follows from simple algebra with Bregman divergences, the second to last inequality follows from Fenchel's inequality applied to the conjugate functions $\frac{1}{2}\|\cdot\|^2_{\psi_t}$ and $\frac{1}{2}\|\cdot\|^2_{\psi_t^*}$ (Boyd and Vandenberghe, 2004, Example 3.27), and the last inequality follows from the assumed strong convexity of B_{ψ_t} with respect to the norm $\|\cdot\|_{\psi_t}$. □

Proof of Corollary 4. Simply sum the equation in the conclusion of the above corollary. □

We next move to the proof of Corollary 3. The proof of the corollary essentially builds upon Xiao (2009) and Nesterov (2009), with a slight modification to deal with the indexing of ψ_t. We include the proof only for completeness.

Proof of Corollary 3. In the original analysis of Xiao (2009), the functions ψ_t are constrained to be $\psi_t(x) = \sqrt{t}\psi(x)$ for a pre-specified strongly-convex function ψ. We assume as in Eq. (2) now that ψ_t changes, however the step-size η is a constant and not time-dependent. We start by defining conjugate-like functions

$$U_t(g) = \sup_{x \in X : \psi_t(x) \le \psi_t(x^*)} \left[\langle g, x\rangle - t\varphi(x)\right] \tag{24}$$

$$V_t(g) = \sup_{x \in X} \left[\langle g, x\rangle - t\varphi(x) - \frac{1}{\eta}\psi_t(x)\right]. \tag{25}$$

In the original analysis, ψ was fixed and was simply upper bounded by a scalar D^2, while we bound $\psi_t(x)$ by $\psi_t(x^*)$. Since φ is closed and ψ_t is strongly convex, it is clear that the supremum in V_t is attained at a unique point and the supremum in U_t is also attained. In order to proceed, we use the following lemma, which is also used by Nesterov (2009) and Xiao (2009).

Lemma 15. *For any g and any $t \geq 0$, $U_t(g) \leq V_t(g) + \frac{1}{\eta}\psi_t(x^*)$.*

Proof. We have

$$\begin{aligned} U_t(g) &= \sup_{x \in X} \left[\langle g, x\rangle - t\varphi(x) \;:\; \psi_t(x) \leq \psi_t(x^*)\right] \;=\; \sup_x \inf_{\beta \geq 0} \left[\langle g, x\rangle - t\varphi(x) + \beta\left(\psi_t(x^*) - \psi_t(x)\right)\right] \\ &\leq \inf_{\beta \geq 0} \sup_x \left[\langle g, x\rangle - t\varphi(x) + \beta\left(\psi_t(x^*) - \psi_t(x)\right)\right] \;\leq\; \sup_x \left[\langle g, x\rangle - t\varphi(x) + \frac{1}{\eta}\psi_t(x^*) - \frac{1}{\eta}\psi_t(x)\right] \\ &= V_t(g) + \frac{1}{\eta}\psi_t(x^*) \ . \end{aligned}$$

The first inequality is a consequence of the min-max inequality that $\sup_a \inf_b g(a,b) \leq \inf_b \sup_a g(a,b)$, and the second is a consequence of convex duality (Boyd and Vandenberghe, 2004). △

We now define the generalized projection operator π_t

$$\pi_t(-g) = \operatorname*{argmin}_{x \in X} \left[\langle g, x\rangle + t\varphi(x) + \frac{1}{\eta}\psi_t(x)\right] \ . \tag{26}$$

We next use the following standard properties of $V_t(g)$ (see, for Theorem 1 from Nesterov, 2005). The function V_t is convex and differentiable with gradient $\nabla V_t(g) = \pi_t(g)$. Moreover, its gradient is Lipschitz continuous with constant η, specifically,

$$\forall\; g_1, g_2 : \quad \left\|\nabla V_t(g_1) - \nabla V_t(g_2)\right\|_{\psi_t} \leq \eta \left\|g_1 - g_2\right\|_{\psi_t^*} \ .$$

The main consequence of the above reasoning with which we are concerned is the following consequence of the fundamental theorem of calculus (Nesterov, 2004, Theorem 2.1.5):

$$V_t(g + h) \leq V_t(g) + \langle h, \nabla V_t(g)\rangle + \frac{\eta}{2}\left\|h\right\|_{\psi_t^*}^2 \ . \tag{27}$$

For the remainder of this proof, we set $\bar{g}_t = \frac{1}{t}\sum_{\tau=1}^t g_\tau$. To conclude the proof we need the following lemma characterizing V_t as a function of $\bar{g}_t$.

Lemma 16. *For any $t \geq 1$, we have $V_t(-t\bar{g}_t) + \varphi(x_{t+1}) \leq V_{t-1}(-t\bar{g}_t)$.*

Proof. The proof is almost identical to the proof of Lemma 6 from Xiao (2009). Recalling our assumption that $\psi_{t-1}(x) \leq \psi_t(x)$ and that x_{t+1} attains the supremum in $V_t(-t\bar{g}_t)$, we have

$$\begin{aligned} V_t(-t\bar{g}_t) + \varphi(x_{t+1}) &\leq V_t(-t\bar{g}_t) + \varphi(x_{t+1}) + \frac{1}{\eta}\left(\psi_t(x_{t+1}) - \psi_{t-1}(x_{t+1})\right) \\ &= \left[\langle -t\bar{g}_t, x_{t+1}\rangle - t\varphi(x_{t+1}) - \frac{1}{\eta}\psi_t(x_{t+1})\right] + \varphi(x_{t+1}) + \frac{1}{\eta}\left(\psi_t(x_{t+1}) - \psi_{t-1}(x_{t+1})\right) \\ &= \langle -t\bar{g}_t, x_{t+1}\rangle - (t-1)\varphi(x_{t+1}) - \frac{1}{\eta}\psi_{t-1}(x_{t+1}) \\ &\leq \sup_{x \in X} \left[\langle -t\bar{g}_t, x\rangle - (t-1)\varphi(x) - \frac{1}{\eta}\psi_{t-1}(x)\right]. \end{aligned}$$

△

We now proceed in the same fashion as Xiao (2009) and Nesterov (2009), defining duality gap variables

$$\delta_t \;\triangleq \sup_{x \in X} \left\{\sum_{\tau=1}^t \left[\langle g_\tau, x_\tau - x\rangle + \varphi(x_\tau)\right] - t\varphi(x) \;:\; \psi_t(x) \leq \psi_t(x^*)\right\} \ .$$

The above definition along with the convexity of f_τ implies

$$\delta_t \geq \sum_{\tau=1}^t \left[\langle g_\tau, x_\tau - x^*\rangle + \varphi(x_\tau)\right] - t\varphi(x^*) \geq \sum_{\tau=1}^t \left[f_\tau(x_\tau) - f_\tau(x^*) + \varphi(x_\tau) - \varphi(x^*)\right] \ . \tag{28}$$

Using Lemma 15, we upper bound δ_t by

$$\delta_t \;=\; \sum_{\tau=1}^{t} \left[\langle g_\tau, x_\tau\rangle + \varphi(x_\tau)\right] + U_t(-t\bar{g}_t) \;\leq\; \sum_{\tau=1}^{t} \left[\langle g_\tau, x_\tau\rangle + \varphi(x_\tau)\right] + V_t(-t\bar{g}_t) + \frac{1}{\eta}\psi_t(x^*)\;. \tag{29}$$

Finally, we upper bound V_t, and rearrange terms to obtain our desired inequality. First, Lemma 16 and Eq. (27) imply

$$\begin{aligned} V_t(-t\bar{g}^t) + \varphi(x_{t+1}) \;&\leq\; V_{t-1}(-t\bar{g}^t) \;=\; V_{t-1}(-(t-1)\bar{g}^{t-1} - g_t) \\ &\leq\; V_{t-1}(-(t-1)\bar{g}^{t-1}) - \left\langle g_t, \nabla V_{t-1}(-(t-1)\bar{g}^{t-1})\right\rangle + \frac{\eta}{2}\left\|g_t\right\|_{\psi_{t-1}^*}^2 \\ &\leq\; V_{t-1}(-(t-1)\bar{g}^{t-1}) - \langle g_t, x_t\rangle + \frac{\eta}{2}\left\|g_t\right\|_{\psi_{t-1}^*}^2\;. \end{aligned}$$

Using Eq. (29), we sum the above equation from $\tau = 1$ through t to get that

$$\begin{aligned} \delta_t - V_t(-t\bar{g}^t) - \frac{1}{\eta}\psi_t(x^*) \;&\leq\; \sum_{\tau=1}^{t}\left[\left\langle g^\tau, x^\tau - x^0\right\rangle + \varphi(x^{\tau+1})\right] \\ &\leq\; V_0(-0\cdot\bar{g}^0) - V_t(-t\bar{g}^t) + \frac{\eta}{2}\sum_{\tau=1}^{t}\left\|g^\tau\right\|_{\psi_{\tau-1}^*}^2\;. \end{aligned}$$

Since that $0\cdot\bar{g}_0 = 0$ and $V_0(0) = 0$, we can add $V_t(-t\bar{g}^t)$ to both sides of the above inequality to get

$$\delta_t \leq \frac{1}{\eta}\psi_t(x^*) + \frac{\eta}{2}\sum_{\tau=1}^{t}\left\|g^\tau\right\|_{\psi_{\tau-1}^*}^2\,.$$

Combining the above equation with the lower bound on δ_t from Eq. (28) finishes the proof. □

C Technical Lemmas

Lemma 17 (Example 3, Davis, 1963)**.** *Let $A \succeq B \succeq 0$ be symmetric $d\times d$ PSD matrices. Then $A^{1/2} \succeq B^{1/2}$.*

The gradient of the function $\operatorname{tr}(X^p)$ is easy to compute for integer values of p. However, when p is real we need the following lemma. The lemma tacitly uses the fact that there is a unique positive semidefinite X^p when $X \succeq 0$ (Horn and Johnson, 1985, Theorem 7.2.6).

Lemma 18. *Let $p \in \mathbb{R}$ and $X \succ 0$. Then $\nabla_X \operatorname{tr}(X^p) = pX^{p-1}$.*

Proof. We do a first order expansion of $(X+A)^p$ when $X \succ 0$ and A is symmetric. Let $X = U\Lambda U^\top$ be the symmetric eigen-decomposition of X and $VDV^\top$ be the decomposition of $\Lambda^{-1/2}U^\top AU\Lambda^{-1/2}$. Then

$$\begin{aligned} (X+A)^p &= (U\Lambda U^\top + A)^p = U(\Lambda + U^\top AU)^p U^\top = U\Lambda^{p/2}(I + \Lambda^{-1/2}U^\top AU\Lambda^{-1/2})^p\Lambda^{p/2}U^\top \\ &= U\Lambda^{p/2}V^\top(I+D)^pV\Lambda^{p/2}U^\top = U\Lambda^{p/2}V^\top(I + pD + o(D))V\Lambda^{p/2}U^\top \\ &= U\Lambda^pU^\top + pU\Lambda^{p/2}V^\top DV\Lambda^{p/2}U^\top + o(U\Lambda^{-/2}V^\top DV\Lambda^{p/2}U^\top) \\ &= X^p + U\Lambda^{(p-1)/2}U^\top AU\Lambda^{(p-1)/2}U^\top + o(A) = X^p + pX^{(p-1)/2}AX^{(p-1)/2} + o(A). \end{aligned}$$

In the above, $o(A)$ is a matrix that goes to zero faster than $A \to 0$, and the second line follows via a first-order Taylor expansion of $(1+d_i)^p$. From the above, we immediately see that

$$\operatorname{tr}((X+A)^p) = \operatorname{tr}X^p + p\operatorname{tr}(X^{p-1}A) + o(\operatorname{tr}A),$$

which completes the proof. □

Lemma 19. *Let $B \succeq 0$ and assume that $g \in \mathrm{Range}(B)$. Let $B^{-1/2}$ denote the root of the inverse of B when $B \succ 0$ and the root of the pseudo-inverse of B otherwise. Then, for any t such that $B - tgg^\top \succeq 0$ the following inequality holds*

$$2\,\mathrm{tr}((B - tgg^\top)^{1/2}) \le 2\,\mathrm{tr}(B^{1/2}) - t\,\mathrm{tr}(B^{-1/2}gg^\top)\ .$$

Proof. The core of the proof is based on the concavity of the function $\mathrm{tr}(A^{1/2})$. However, careful analysis is required as A might not be strictly positive definite. We also use the previous lemma which implies that the gradient of $\mathrm{tr}(A^{1/2})$ is $\frac{1}{2}A^{-1/2}$ when $A \succ 0$.

First, A^p is matrix-concave for $A \succ 0$ and $0 \le p \le 1$ (see, for example, Corollary 4.1 in Ando, 1979 or Theorem 16.1 in Bondar, 1994). That is, for $A, B \succ 0$ and $\alpha \in [0,1]$ we have

$$(\alpha A + (1-\alpha)B)^p \succeq \alpha A^p + (1-\alpha)B^p\ . \tag{30}$$

Now suppose simply $A, B \succeq 0$ (but neither is necessarily strict). Then for any $\delta > 0$, we have $A + \delta I \succ 0$ and $B + \delta I \succ 0$ and therefore

$$(\alpha(A+\delta I) + (1-\alpha)(B+\delta I))^p \succeq \alpha(A+\delta I)^p + (1-\alpha)(B+\delta I)^p \succeq \alpha A^p + (1-\alpha)B^p\ ,$$

where we used Lemma 17 for the second matrix inequality. Moreover, $\alpha A + (1-\alpha)B + \delta I \to \alpha A + (1-\alpha)B$ as $\delta \to 0$. Since A^p is continuous (when we use the unique PSD root), this line of reasoning proves that Eq. (30) holds for $A, B \succeq 0$. Thus, we proved that

$$\mathrm{tr}((\alpha A + (1-\alpha)B)^p) \ge \alpha\,\mathrm{tr}(A^p) + (1-\alpha)\,\mathrm{tr}(B^p) \ \text{ for } 0 \le p \le 1\ .$$

Recall now that Lemma 18 implies that the gradient of $\mathrm{tr}(A^{1/2})$ is $\frac{1}{2}A^{-1/2}$ when $A \succ 0$. Therefore, from the concavity of $A^{1/2}$ and the form of its gradient, we can use the standard first-order inequality for concave functions so that for any $A, B \succ 0$,

$$\mathrm{tr}(A^{1/2}) \le \mathrm{tr}(B^{1/2}) + \frac{1}{2}\,\mathrm{tr}(B^{-1/2}(A-B))\ . \tag{31}$$

Let $A = B - tgg^\top \succeq 0$ and suppose only that $B \succeq 0$. We must take some care since $B^{-1/2}$ may not necessarily exist, and the above inequality does not hold true in the pseudo-inverse sense when $B \not\succ 0$. However, for any $\delta > 0$ we know that $2\nabla_B\,\mathrm{tr}((B+\delta I)^{1/2}) = (B+\delta I)^{-1/2}$, and $A - B = -tgg^\top$. From Eq. (31) and Lemma 17, we have

$$\begin{aligned} 2\,\mathrm{tr}(B - tgg^\top)^{1/2} &= 2\,\mathrm{tr}(A^{1/2}) \le 2\,\mathrm{tr}((A+\delta I)^{1/2}) \\ &\le 2\,\mathrm{tr}(B+\delta I)^{1/2} - t\,\mathrm{tr}((B+\delta I)^{-1/2}gg^\top)\ . \end{aligned} \tag{32}$$

Now let $B = V\,\mathrm{diag}(\lambda)V^\top$ be the eigen-decomposition of B, and note that since $g \in \mathrm{Range}(B)$,

$$\begin{aligned} g^\top(B+\delta I)^{-1/2}g &= g^\top V\,\mathrm{diag}\Big(1/\sqrt{\lambda_i+\delta}\Big)\,V^\top g \\ &= \sum_{i:\lambda_i>0} \frac{1}{\sqrt{\lambda_i+\delta}}(g^\top v_i)^2 \ \underset{\delta\downarrow 0}{\longrightarrow}\ \sum_{i:\lambda_i>0} \lambda_i^{-1/2}(g^\top v_i)^2 \ = \ g^\top B^{-1/2}g\ . \end{aligned}$$

Thus, by taking $\delta \downarrow 0$ in Eq. (32), and since both $\mathrm{tr}(B+\delta I)^{1/2}$ and $\mathrm{tr}((B+\delta I)^{-1/2}gg^\top)$ are evidently continuous in δ, we complete the proof. □

Lemma 20. *Let $\delta \ge \|g\|_2$ and $A \succeq 0$, then $\langle g, (\delta I + A^{1/2})^{-1}g\rangle \le \Big\langle g, \big((A+gg^\top)^\dagger\big)^{1/2} g\Big\rangle$.*

Proof. We begin by noting that $\delta^2 I \succeq gg^\top$, so from Lemma 17 we get $(A + gg^\top)^{1/2} \preceq (A + \delta^2 I)^{1/2}$. Since A and I are simultaneously diagonalizable, we can generalize the inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, which holds for $a, b \ge 0$, to positive semi-definite matrices, thus,

$$(A + \delta^2 I)^{1/2} \preceq A^{1/2} + \delta I\ .$$

Therefore, if $A + gg^\top$ is of full rank, we have $(A+gg^\top)^{-1/2} \succeq (A^{1/2} + \delta I)^{-1}$ (Horn and Johnson, 1985, Corollary 7.7.4(a)). Since $g \in \mathrm{Range}((A+gg^\top)^{1/2})$, we can apply an analogous limiting argument to the one used in the proof of Lemma 19 and discard all zero eigenvalues of $A + gg^\top$, which completes the lemma. □

We next prove another technical lemma that is useful in characterizing the solution of the optimization problem below. Note that the second part of the lemma implies that we can treat the inverse of the solution matrix S^{-1} as $S^{\dagger}$. We consider solving

$$\min_{S} \ \operatorname{tr}(S^{-1}A) \quad \text{subject to} \quad S \succeq 0, \ \operatorname{tr}(S) \leq c \ \text{ where } \ A \succeq 0 \ . \tag{33}$$

Lemma 21. *If A is of full rank, then the minimizer of Eq. (33) is $S = cA^{\frac{1}{2}} / \operatorname{tr}(A^{\frac{1}{2}})$. If A is not of full rank, then setting $S = cA^{\frac{1}{2}} / \operatorname{tr}(A^{\frac{1}{2}})$ gives*

$$\operatorname{tr}(S^{\dagger}A) = \inf_{S} \left\{ \operatorname{tr}(S^{-1}A) : S \succeq 0, \ \operatorname{tr}(S) \leq c \right\} \ .$$

In either case, $\operatorname{tr}(S^{\dagger}A) = \operatorname{tr}(A^{\frac{1}{2}})^2/c$.

Proof. Both proofs rely on constructing the Lagrangian for Eq. (33). We introduce $\theta \in \mathbb{R}_+$ for the trace constraint and $Z \succeq 0$ for the positive semidefinite constraint on S. In this case, the Lagrangian is

$$\mathcal{L}(S, \theta, Z) = \operatorname{tr}(S^{-1}A) + \theta(\operatorname{tr}(S) - c) - \operatorname{tr}(SZ).$$

The derivatives of $\mathcal{L}$ with respect to all of the elements of S are

$$-S^{-1}AS^{-1} + \theta I - Z. \tag{34}$$

If S is full rank, then to satisfy the generalized complimentarity conditions (Boyd and Vandenberghe, 2004) for the problem, we must have $Z = 0$. Therefore, we get that $S^{-1}AS^{-1} = \theta I$. We now can multiply by S on the right and the left to get that $A = \theta S^2$, which implies that $S \propto A^{\frac{1}{2}}$. If A is of full rank, the optimal solution for $S \succ 0$ forces θ to be positive so that $\operatorname{tr}(S) = c$. This yields the solution $S = cA^{\frac{1}{2}} / \operatorname{tr}(A^{\frac{1}{2}})$. In order to verify optimality of this solution, we set $Z = 0$ and $\theta = c^{-2} \operatorname{tr}(A^{1/2})^2$ which gives $\nabla_S \mathcal{L}(S, \theta, Z) = 0$, as is indeed required.

Suppose now that A is not full rank and that $A = Q \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} Q^{\top}$ is the eigen-decomposition of A. Let n be the dimension of the null-space of A (so the rank of A is $d - n$). Define the variables

$$Z(\theta) = \begin{bmatrix} 0 & 0 \\ 0 & \theta I \end{bmatrix}, \quad S(\theta, \delta) = \frac{1}{\sqrt{\theta}} Q \begin{bmatrix} \Lambda^{\frac{1}{2}} & 0 \\ 0 & \delta I \end{bmatrix} Q^{\top}, \quad S(\delta) = \frac{c}{\operatorname{tr}(A^{\frac{1}{2}}) + \delta n} Q \begin{bmatrix} \Lambda^{\frac{1}{2}} & 0 \\ 0 & \delta I \end{bmatrix} Q^{\top}.$$

It is easy to see that $\operatorname{tr} S(\delta) = c$, and clearly $\lim_{\delta \to 0} \operatorname{tr}(S(\delta)^{-1}A) = \operatorname{tr}(S(0)^{\dagger}A) = \operatorname{tr}(A^{\frac{1}{2}}) \operatorname{tr}(\Lambda^{\frac{1}{2}})/c = \operatorname{tr}(A^{\frac{1}{2}})^2/c$. Further, let $g(\theta) = \inf_S \mathcal{L}(S, \theta, Z(\theta))$ be the dual of Eq. (33). From the above analysis and Eq. (34), it is evident that

$$-S(\theta, \delta)^{-1}AS(\theta, \delta)^{-1} + \theta I - Z(\theta) = -\theta Q \begin{bmatrix} \Lambda^{-\frac{1}{2}}\Lambda\Lambda^{-\frac{1}{2}} & 0 \\ 0 & \delta^{-2}I \cdot 0 \end{bmatrix} Q^{\top} + \theta I - \begin{bmatrix} 0 & 0 \\ 0 & \theta I \end{bmatrix} = 0.$$

So $S(\theta, \delta)$ achieves the infimum in the dual for *any* $\delta > 0$, $\operatorname{tr}(S(0)Z(\theta)) = 0$, and

$$g(\theta) = \sqrt{\theta} \operatorname{tr}(\Lambda^{\frac{1}{2}}) + \sqrt{\theta} \operatorname{tr}(\Lambda^{\frac{1}{2}}) + \sqrt{\theta}\delta n - \theta c.$$

Setting $\theta = \operatorname{tr}(\Lambda^{\frac{1}{2}})^2/c^2$ gives $g(\theta) = \operatorname{tr}(\Lambda^{\frac{1}{2}})^2/c - \delta n \operatorname{tr}(\Lambda^{\frac{1}{2}})/c$. Taking $\delta \to 0$ gives $g(\theta) = \operatorname{tr}(A^{\frac{1}{2}})^2/c$, which means that $\lim_{\delta \to 0} \operatorname{tr}(S(\delta)^{-1}A) = \operatorname{tr}(A^{\frac{1}{2}})^2/c = g(\theta)$. Thus the duality gap for the original problem is 0 so $S(0)$ is the limiting solution.

The last statement of the lemma is simply plugging $S^{\dagger} = (A^{\dagger})^{\frac{1}{2}} \operatorname{tr}(A^{\frac{1}{2}})/c$ in to the objective being minimized. □

References

J. Abernethy, P. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the Twenty First Annual Conference on Computational Learning Theory*, 2008a.

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the Twenty First Annual Conference on Computational Learning Theory*, 2008b.

T. Ando. Concavity of certain maps on positive definite matrices and applications to Hadamard products. *Linear Algebra and its Applications*, 26:203–241, 1979.

P. L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems 20*, 2007.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

J. V. Bondar. Comments on and complements to *Inequalities: Theory of Majorization and Its Applications*. *Linear Algebra and its Applications*, 199:115–129, 1994.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

N. Cesa-Bianchi, A. Conconi, , and C. Gentile. A second-order perceptron algorithm. *SIAM Journal on Computing*, 34(3):640–668, 2005.

N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007.

K. Crammer, M. Dredze, and F. Pereira. Exact convex confidence-weighted learning. In *Advances in Neural Information Processing Systems 22*, 2008.

K. Crammer, M. Dredze, and A. Kulesza. Adaptive regularization of weight vectors. In *Advances in Neural Information Processing Systems 23*, 2009.

C. Davis. Notions generalizing convexity for functions defined on spaces of matrices. In *Proceedings of the Symposia in Pure Mathematics*, volume 7, pages 187–201. American Mathematical Society, 1963.

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2873–2908, 2009.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. *Submitted*, 2010. URL `cs.berkeley.edu/~jduchi/projects/DuchiShSiTe10.html`.

E. Hazan and S. Kale. Extracting certainty from uncertainty: regret bounded by variation in costs. In *Proceedings of the Twenty First Annual Conference on Computational Learning Theory*, 2008.

E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2006.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2003.

A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Efficiency in Optimization*. John Wiley and Sons, 1983.

Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

Y. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.

Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1): 221–259, 2009.

A. Rakhlin. Lecture notes on online learning. For the Statistical Machine Learning Course at University of California, Berkeley, 2009.

G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 1988.

S. Shalev-Shwartz and Y. Singer. Logarithmic regret algorithms for strongly convex repeated games. Technical report, The Hebrew University, 2007. URL `http://www.cs.huji.ac.il/~shais`.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. In *Advances in Neural Information Processing Systems 23*, 2009.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.