Adaptive Subgradient Methods for Online Learning and Stochastic Optimization

John Duchi
Elad Hazan
Yoram Singer

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2010-24
http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-24.html

March 3, 2010

Acknowledgement

There are many people to whom we owe our thanks for this research. Fernando Pereira helped push us in the direction of working on adaptive methods and has been a constant source of discussion and helpful feedback. Samy Bengio provided us with a version of the ImageNet dataset and was instrumental in helping to get our experiments running, and Adam Sadovsky gave many coding suggestions. Lastly, Sam Roweis was a sounding board for our earlier ideas on the subject, and we will miss him dearly.

Adaptive Subgradient Methods for Online Learning and Stochastic Optimization

John Duchi jduchi@cs.berkeley.edu
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720 USA

Elad Hazan hazan@cs.princeton.edu
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120

Yoram Singer singer@google.com
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043 USA

Abstract

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. Metaphorically, the adaptation 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. We give several efficient algorithms for empirical risk minimization problems with common and important regularization functions and domain constraints. We experimentally study our theoretical analysis and show that adaptive subgradient methods significantly outperform state-of-the-art, yet non-adaptive, subgradient algorithms.

Keywords: Subgradient methods, adaptivity, online learning, stochastic convex optimization

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 concretely 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. Informally, our procedures give frequently occurring features very low learning rates and infrequent features high learning rates, where the intuition is that each time an infrequent feature is seen, the learner should "take notice." Thus, the adaptation facilitates finding and identifying very predictive but comparatively rare features.

1.1 The Adaptive Gradient Algorithm

Before introducing our adaptive gradient algorithm, which we term AdaGrad, we establish notation. Vectors and scalars are lower case italic letters, such as $x \in \mathcal{X}$. 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 $f'(x) \in \partial f(x)$ or $g_t \in \partial f_t(x_t)$. 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. 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 $g_{1:t,i}$. We also define the outer product matrix $G_t = \sum_{\tau=1}^{t} g_\tau g_\tau^\top$.

Online learning and stochastic optimization are closely related and basically interchangeable (Cesa-Bianchi et al., 2004). In order to keep our presentation simple, we confine our discussion and algorithmic descriptions to the online setting with the regret bound model. In online learning, the learner repeatedly predicts a point $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$, which often represents a weight vector assigning importance values to various features. The learner's goal is to achieve low regret with respect to a static predictor x^* in the (closed) convex set $\mathcal{X} \subseteq \mathbb{R}^d$ (possibly $\mathcal{X} = \mathbb{R}^d$) on a sequence of functions $f_t(x)$, measured as

$$R(T) = \sum_{t=1}^{T} f_t(x_t) - \inf_{x \in X} \sum_{t=1}^{T} f_t(x) \ .$$

At every timestep t, the learner receives the (sub)gradient information $g_t \in \partial f_t(x_t)$. Standard subgradient algorithms then move the predictor x_t in the opposite direction of g_t while maintaining $x_{t+1} \in \mathcal{X}$ via the projected gradient update (e.g. Zinkevich, 2003)

$$x_{t+1} = \Pi_{\mathcal{X}} \left(x_t - \eta g_t \right) = \operatorname*{argmin}_{x \in \mathcal{X}} \|x - (x_t - \eta g_t)\|_2^2 \ .$$

In contrast, let the Mahalanobis norm $\|\cdot\|_A = \sqrt{\langle \cdot, A, \cdot \rangle}$ and denote the projection of a point y onto $\mathcal{X}$ by $\Pi_{\mathcal{X}}^A(y) = \operatorname{argmin}_{x \in \mathcal{X}} \|x - y\|_A = \operatorname{argmin}_{x \in \mathcal{X}} \langle x - y, A(x - y) \rangle$. Using this notation, our generalization of standard gradient descent employs the update

$$x_{t+1} = \Pi_{\mathcal{X}}^{G_t^{1/2}} \Big(x_t - \eta G_t^{-1/2} g_t \Big) \ .$$

The above algorithm is computationally impractical in high dimensions since it requires computation of the root of the matrix G_t, the outer product matrix. Thus we specialize the update to

$$x_{t+1} = \Pi_{\mathcal{X}}^{\mathrm{diag}(G_t)^{1/2}}\left(x_t - \eta\,\mathrm{diag}(G_t)^{-1/2} g_t\right). \tag{1}$$

Both the inverse and root of $\mathrm{diag}(G_t)$ can be computed in linear time. Moreover, as we discuss later, when the gradient vectors are sparse the update above can often be performed in time proportional to the support of the gradient. We now elaborate and give a more formal discussion of our setting.

In this paper we consider several different online learning algorithms and their stochastic convex optimization counterparts. Formally, we consider online learning with a sequence of composite functions ϕ_t. Each function is of the form $\phi_t(x) = f_t(x) + \varphi(x)$ where f_t and φ are (closed) convex functions. In the learning settings we study, f_t is either an instantaneous loss or a stochastic estimate of the objective function in an optimization task. The function φ serves as a fixed regularization function and is typically used to control the complexity of x. At each round the algorithm makes a prediction $x_t \in \mathcal{X}$ and then receives the function f_t. We define the regret with respect to the fixed (optimal) predictor x^* as

$$R_\phi(T) \triangleq \sum_{t=1}^{T} \left[\phi_t(x_t) - \phi_t(x^*)\right] = \sum_{t=1}^{T} \left[f_t(x_t) + \varphi(x_t) - f_t(x^*) - \varphi(x^*)\right] \ . \tag{2}$$

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

Our analysis applies to related, yet different, methods for minimizing the regret defined in Eq. (2). The first is Nesterov's primal-dual subgradient method 2009, and in particular its specialized versions: regularized dual asent (RDA) (Xiao, 2009) and the follow-the-regularized-leader (FTRL) family of algorithms (see for instance Kalai and Vempala, 2003; Hazan et al., 2006). In the primal-dual subgradient method the algorithm makes 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_t is the *proximal* term. The update amounts to solving the problem

$$x_{t+1} = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{\eta\left\langle \bar{g}_t, x\right\rangle + \eta\varphi(x) + \frac{1}{t}\psi_t(x)\right\} \ , \tag{3}$$

where η is a step-size. The second method similarly has many names, including proximal gradient, forward-backward splitting, and composite mirror descent (Tseng, 2008; Duchi and Singer, 2009; Duchi et al., 2010). We use the term composite mirror descent. 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,

$$x_{t+1} = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{\eta\left\langle g_t, x\right\rangle + \eta\varphi(x) + B_{\psi_t}(x, x_t)\right\} \ . \tag{4}$$

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

$$H_t = \operatorname{diag}(G_t)^{1/2} \ \ \text{(Diagonal)} \quad \text{and} \quad H_t = G_t^{1/2} \ \ \text{(Full)} \ . \tag{5}$$

Plugging the appropriate matrix from the above equation into ψ_t in Eq. (3) or Eq. (4) 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 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 their non-adaptive version. In both examples we use the hinge loss, that is,

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

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 $\mathcal{X}$ 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. (4) with $\mathcal{X} = \{x : \|x\|_\infty \leq 1\}$. 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 $\mathcal{X}$ as in Eq. (1). 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 $y_t = \operatorname{sign}(\langle 1, z_t \rangle)$ so that there exists a perfect classifier $x^* = 1 \in \mathcal{X} \subset \mathbb{R}^d$. 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

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

After the first d rounds, the adaptive predictor has $x_{d+1} = x_{d+\tau} = 1$ for all $\tau \geq 1$ 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 component of the predictor is updated to $1/\sqrt{i} + 1/\sqrt{d+i}$ after the second set of d rounds (truncated so that $|x_i| \leq 1$). In general, the regret of adaptive gradient descent is d, while online gradient descent suffers

regret

$$d + \sum_{t=0}^{T} \sum_{i=1}^{d} \left[1 - \sum_{\tau=0}^{t} \frac{1}{\sqrt{i + \tau d}} \right]_+ . \tag{6}$$

The next proposition lower bounds Eq. (6).

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

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

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 stepsizes larger than $\eta/\sqrt{t}$, online gradient descent might suffer a lower loss in the above setting. However, an adversary could simply play $z_t = e_1$ indefinitely until $\eta/\sqrt{t} \leq \varepsilon$ 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. (5) gives substantially lower regret than stochastic gradient descent. For full divergences we set $\mathcal{X} = \{x : \|x\|_2 \leq \sqrt{d}\}$. 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 $y_t = \operatorname{sign}(\langle 1, V^\top z_t \rangle) = \operatorname{sign}\left(\sum_{i=1}^{d} \langle v_i, z_t \rangle\right)$. We provide an elaborated explanation in Appendix A. Intuitively, with $\psi_t(x) = \langle x, H_t x \rangle$ and H_t set to be the full matrix from Eq. (5), AdaGrad again needs to observe each orthonormal vector v_i only once while 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

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

We further show that

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

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

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

Similar to the above, we will show that

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

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

Following are a simple example and corollary to Theorem 6 to illustrate one regime in which we expect substantial improvements. Let $\varphi \equiv 0$ and consider Zinkevich's 2003 online gradient descent algorithm. Given a compact convex set $\mathcal{X} \subseteq \mathbb{R}^d$ and sequence of convex functions f_t, Zinkevich's algorithm makes the sequence of predictions $x_1, \ldots, x_T$ with $x_{t+1} = \Pi_{\mathcal{X}}(x_t - (\eta/\sqrt{t})g_t)$. If the diameter of $\mathcal{X}$ is bounded, thus $\sup_{x,y\in\mathcal{X}} \|x - y\|_2 \leq D_2$, then Zinkevich's algorithm, with the optimal choice in *hindsight* for the stepsize η (see Eq. (8)), achieves a regret bound of

$$\sum_{t=1}^T f_t(x_t) - \inf_{x\in\mathcal{X}} \sum_{t=1}^T f_t(x) \leq \sqrt{2} D_2 \sqrt{\sum_{t=1}^T \|g_t\|_2^2}. \tag{7}$$

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

Corollary 2 *Let the sequence $\{x_t\} \subset \mathbb{R}^d$ be generated by the update in Eq. (4) and let $\max_t \|x^* - x_t\|_\infty \leq D_\infty$. Then, using a stepsize $\eta = D_\infty/\sqrt{2}$, for any x^*, the following bound holds.*

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

The important feature of the bound above is the infimum under the square root, which allows us to perform better than simply using the identity matrix, and the fact that the stepsize is easy to set a priori. For example, if the set $\mathcal{X} = \{x : \|x\|_\infty \leq 1\}$, 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. (7) as the identity matrix belongs to the set over which we take the infimum. To conclude the outline of results we would like to note that, the regret bound by Zinkevich is tight and cannot be improved in a worst case setting (Abernethy et al., 2008). Therefore, we impose specific reasonable assumptions on the input space that yield improved bounds.

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 casts 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. Verbatim usage of the FTRL approach fails to achieve low regret, however, adding a proximal[1] 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 adjust the learning rates for online learning and stochastic gradient descent on a per-feature basis. The latter can be very useful when our gradient vectors g_t are sparse, for example, in a classification setting where examples may have only a small number of non-zero features. As we demonstrated 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 also extend a few well-known algorithms. In particular, we consider variants of the follow-the-regularized leader (FTRL) algorithms mentioned above, which are kin to Zinkevich's lazy projection algorithm. We use Xiao's recently analyzed regularized dual ascent (RDA) algorithm (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 (proximal gradient) generalizations (Tseng, 2008; Duchi et al., 2010), which in turn include as special cases projected gradients (Zinkevich, 2003) and mirror descent (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003).

The idea of adapting first order optimization methods is by no means new and can be traced back at least to the 1970s with the work on space dilation methods of Shor (1972) and variable metric methods, such as the BFGS family of algorithms (e.g. Fletcher, 1970). This older work often assumed that the function to be minimized was differentiable and, to our knowledge, did not consider stochastic, online, or composite optimization. More recently, Bottou and Gallinari (2009) proposed careful Quasi-Newton stochastic gradient descent, which is similar in spirit to our methods, but their convergence results assume a smooth objective with positive definite Hessian bounded away from 0. Our results apply more generally.

Prior to the analysis presented in this paper for online and stochastic optimization, the strongly convex function ψ in the update equations (3) and (4) 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)

1. 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.

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)

$$R_\phi(T) \leq \frac{1}{\eta} B_\psi(x^*, x_1) + \frac{\eta}{2} \sum_{t=1}^{T} \left\| f_t'(x_t) \right\|_*^2 \ . \tag{8}$$

Choosing $\eta \propto 1/\sqrt{T}$ gives $R_\phi(T) = O(\sqrt{T})$. When $B_\psi(x, x^*)$ is bounded for all $x \in \mathcal{X}$, 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, Theorem 3):

$$R_\phi(T) \leq \sqrt{T}\psi(x^*) + \frac{1}{2\sqrt{T}} \sum_{t=1}^{T} \left\| f_t'(x_t) \right\|_*^2 \ . \tag{9}$$

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 in the mistake-bound setting for online learning in the past. The framework that is most related to ours is probably confidence weighted learning scheme by 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). AROW maintains a mean prediction vector $\mu_t \in \mathbb{R}^d$ and a covariance matrix $\Sigma_t \in \mathbb{R}^{d \times d}$ over μ_t as well. At every step of the algorithm, the learner receives a pair (z_t, y_t) where $z_t \in \mathbb{R}^d$ is the tth example and $y_t \in \{-1, +1\}$ is the label. Whenever the predictor μ_t attains a margin value smaller than 1, AROW performs the update

$$\begin{aligned} \beta_t = \frac{1}{\langle z_t, \Sigma_t z_t \rangle + \lambda}, \quad \alpha_t = \left[1 - y_t \left\langle z_t, \mu_t \right\rangle\right]_+ , \\ \mu_{t+1} = \mu_t + \alpha_t \Sigma_t y_t z_t, \quad \Sigma_{t+1} = \Sigma_t - \beta_t \Sigma_t x_t x_t^\top \Sigma_t. \end{aligned} \tag{10}$$

In the above scheme, one can force Σ_t to be diagonal, which reduces the run-time and storage requirements of the algorithm but still gives good performance (Crammer et al., 2009). In contrast to AROW, the AdaGrad algorithm uses the *root* of the inverse covariance matrix, a consequence of our 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 AROW algorithm is $O(d^2)$, which is faster than ours as it requires computing the root of a matrix.

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 f_t were proposed by Cesa-Bianchi et al. (2007) and derived by Hazan and Kale (2008). Bartlett et al. (2007) explore another adaptation technique for η_t where they adapt the step size to accommodate both strongly and weakly convex functions.

Our approach differs from previous approaches as it does not focus on a particular loss function or mistake bound. Instead, we view the problem of adapting the proximal function as a meta-learning problem. We then 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. (8) and Eq. (9), 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 our 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, $\psi_{t+1}(x) \geq \psi_t(x)$. 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

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

Strong convexity is guaranteed if and only if $B_{\psi_t}(x, y) \geq \frac{1}{2}\|x - y\|^2_{\psi_t}$. 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 C, as they build straightforwardly on work by Duchi et al. (2010) and Xiao (2009). For the primal-dual subgradient update, the following bound holds.

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

$$R_\phi(T) \leq \frac{1}{\eta}\psi_T(x^*) + \frac{\eta}{2}\sum_{t=1}^{T}\left\|f'_t(x_t)\right\|^2_{\psi^*_{t-1}} \ . \tag{11}$$

For composite mirror descent algorithms we have a similar corollary.

Corollary 4 *Let the sequence $\{x_t\}$ be defined by the update in Eq. (4). Assume w.l.o.g. that $\varphi(x_1) = 0$. Then for any $x^* \in \mathcal{X}$, we have*

$$R_\phi(T) \leq \frac{1}{\eta}B_{\psi_1}(x^*, x_1) + \frac{1}{\eta}\sum_{t=1}^{T-1}\left[B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1})\right] + \frac{\eta}{2}\sum_{t=1}^{T}\left\|f'_t(x_t)\right\|^2_{\psi^*_t} . \tag{12}$$

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.

Input: $\eta > 0$, $\delta \geq 0$
Variables: $s \in \mathbb{R}^d, H \in \mathbb{R}^{d \times d}$, $g_{1:t,i} \in \mathbb{R}^t$ for $i \in \{1, \ldots, d\}$
Initialize $x_1 = 0$, $g_{1:0} = []$
For $t = 1$ to T
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 $H_t = \delta I + \mathrm{diag}(s_t)$, $\psi_t(x) = \frac{1}{2}\langle x, H_t\, x\rangle$

Primal-Dual Subgradient Update (Eq. (3)):

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \eta \left\langle \frac{1}{t} \sum_{\tau=1}^{t} g_\tau, x \right\rangle + \eta \varphi(x) + \frac{1}{t} \psi_t(x) \right\}.$$

Composite Mirror Descent Update (Eq. (4)):

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

Figure 1: AdaGrad with diagonal matrices

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

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

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 $\theta \geq 0$ 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 $-\|g_{1:T,i}\|_2^2 / s_i^2 - \lambda_i + \theta = 0$, and complimentarity conditions on $\lambda_i s_i$ (Boyd and Vandenberghe, 2004) imply that $\lambda_i = 0$. Thus we have $s_i = \theta^{-\frac{1}{2}} \|g_{1:T,i}\|_2$, 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 a final note, we can plug s_i into the objective above to see that

$$\inf_s \left\{ \sum_{t=1}^T \sum_{i=1}^d \frac{g_{t,i}^2}{s_i} \ : \ s \succeq 0, \langle 1, s \rangle \leq c \right\} = \frac{1}{c} \left(\sum_{i=1}^d \left\|g_{1:T,i}\right\|_2 \right)^2 \ . \tag{13}$$

Letting $\mathrm{diag}(v)$ denote the diagonal matrix with diagonal v, it is natural to suspect that if we use a proximal function similar to $\psi(x) = \langle x, \mathrm{diag}(s) x \rangle$ with associated squared dual norm $\|x\|_{\psi^*}^2 = \left\langle x, \mathrm{diag}(s)^{-1} x \right\rangle$, we should do well lowering the gradient terms in the regret in Eq. (11) and Eq. (12).

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*

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

Proof We prove the lemma by considering an arbitrary real-valued sequence $\{a_i\}$ and its vector representation $a_{1:i} = [a_1 \ \cdots \ a_i]$. We are next going to show that

$$\sum_{t=1}^T \frac{a_t^2}{\left\|a_{1:t}\right\|_2} \leq 2 \left\|a_{1:T}\right\|_2 \ , \tag{14}$$

where we define $\frac{0}{0} = 0$. We use induction on T to prove Eq. (14). For $T = 1$, the inequality trivially holds. Assume Eq. (14) holds true for $T - 1$, then

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

where the inequality follows from the inductive hypothesis. We define $b_T = \sum_{t=1}^T a_t^2$ and use concavity to obtain that $\sqrt{b_T - a_T^2} \leq \sqrt{b} - a_T^2 \frac{1}{2\sqrt{b_T}}$ so long as $b_T - a_T^2 \geq 0$.[2] Thus,

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

Having proved Eq. (14), we note that by construction $s_{t,i} = \left\|g_{1:t,i}\right\|_2$, so

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

■

2. We note that we use an identical technique in the full-matrix case. See Lemma 17 in the appendix

To get a regret bound, we need to consider the terms consisting of the dual-norm of the subgradient in Eq. (11) and Eq. (12), $\|f_t'(x_t)\|_{\psi_t^*}^2$. 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 $\|f_t'(x_t)\|_{\psi_t^*}^2 \leq \left\langle g_t, \operatorname{diag}(s_t)^{-1} g_t \right\rangle$. Note that if $s_{t,i} = 0$ then $g_{t,i} = 0$ by definition of $s_{t,i}$. Thus, Lemma 5 gives

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

To obtain a bound for a primal-dual subgradient method, we set $\delta \geq \max_t \|g_t\|_\infty$, in which case $\|g_t\|_{\psi_{t-1}^*}^2 \leq \left\langle g_t, \operatorname{diag}(s_t)^{-1} g_t \right\rangle$, 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 the composite mirror-descent update. Examing Eq. (12) and Alg. 1, we notice that

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

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

$$\begin{aligned} \sum_{t=1}^{T-1} B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1}) \;&\leq\; \frac{1}{2} \sum_{t=1}^{T-1} \|x^* - x_{t+1}\|_\infty^2 \left\langle s_{t+1} - s_t, 1 \right\rangle \\ &\leq\; \frac{1}{2} \max_{t \leq T} \|x^* - x_t\|_\infty^2 \sum_{i=1}^{d} \|g_{1:T,i}\|_2 - \frac{1}{2} \|x^* - x_1\|_\infty^2 \left\langle s_1, 1 \right\rangle \;. \end{aligned} \tag{15}$$

We also have

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

Combining the above arguments with Corollaries 3 and 4, and using Eq. (15) with the fact that $B_{\psi_1}(x^*, x_1) \leq \frac{1}{2} \|x^* - x_1\|_\infty^2 \langle 1, s_1\rangle$, we have proved the following theorem.

Theorem 6 *Let the sequence $\{x_t\}$ be defined by Algorithm 1. For x_t generated using the primal-dual subgradient update of Eq. (3) with $\delta \geq \max_t \|g_t\|_\infty$, then for any $x^* \in \mathcal{X}$,*

$$R_\phi(T) \leq \frac{\delta}{\eta} \|x^*\|_2^2 + \frac{1}{\eta} \|x^*\|_\infty^2 \sum_{i=1}^{d} \|g_{1:T,i}\|_2 + \eta \sum_{i=1}^{d} \|g_{1:T,i}\|_2. \tag{16}$$

For x_t generated using the composite mirror-descent update of Eq. (4), then for any $x^* \in \mathcal{X}$

$$R_\phi(T) \leq \frac{1}{2\eta} \max_{t \leq T} \|x^* - x_t\|_\infty^2 \sum_{i=1}^{d} \|g_{1:T,i}\|_2 + \eta \sum_{i=1}^{d} \|g_{1:T,i}\|_2. \tag{17}$$

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

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

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

Corollary 7 *Assume that D_∞ and γ_T are defined as above. For $\{x_t\}$ generated by Algorithm 1 using the primal-dual subgradient update Eq. (3) with $\eta = \|x^*\|_\infty$, then for any $x^* \in \mathcal{X}$ we have*

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

Using the composite mirror descent update of Eq. (4) to generate $\{x_t\}$ and setting $\eta = D_\infty / \sqrt{2}$, we have

$$R_\phi(T) \leq \sqrt{2} D_\infty \sum_{i=1}^{d} \|g_{1:T,i}\|_2 = \sqrt{2} D_\infty \gamma_T \ .$$

Intuitively, as discussed in the introduction, Alg. 1 should have lower regret than non-adaptive algorithms on sparse data. For example, suppose that our learning problem is a logistic regression with 0/1-valued features. Then the gradient terms are likewise based on 0/1-valued features and sparse, so the gradient terms in the bound $\sum_{i=1}^{d} \|g_{1:t,i}\|_2$ should all be much smaller than $\sqrt{T}$. If 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 have significantly lower regret by using higher learning rates for infrequent features and lower learning rates on commonly appearing features. Further, if the optimal predictor is relatively dense,as is often the case in predictions problems with sparse inputs, 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 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:

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

The solution is obtained by defining $G_t = \sum_{\tau=1}^{t} g_\tau g_\tau^\top$ and setting S to be a normalized version of the root of G_T, that is, $S = c\, G_T^{1/2} / \operatorname{tr}(G_T^{1/2})$. For a proof, see Lemma 19 in

Input: $\eta > 0$, $\delta \geq 0$
Variables: $S_t \in \mathbb{R}^{d \times d}$, $H_t \in \mathbb{R}^{d \times d}$, $G_t \in \mathbb{R}^{d \times d}$
Initialize $x_1 = 0$, $S_0 = 0$, $H_0 = 0$, $G_0 = 0$
For $t = 1$ to T
Suffer loss $f_t(x_t)$
Receive subgradient $g_t \in \partial f_t(x_t)$ of f_t at x_t
Update $G_t = G_{t-1} + g_t g_t^\top$, $S_t = G_t^{\frac{1}{2}}$
Set $H_t = \delta I + S_t$, $\psi_t(x) = \frac{1}{2}\langle x, H_t\, x\rangle$

Primal-Dual Subgradient Update (Eq. (3)):

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \eta \left\langle \frac{1}{t} \sum_{\tau=1}^{t} g_\tau, x \right\rangle + \eta \varphi(x) + \frac{1}{t} \psi_t(x) \right\}.$$

Composite Mirror Descent Update (Eq. (4)):

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

Figure 2: AdaGrad with full matrices

Appendix B, 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 in the diagonal case to attain low regret by collecting gradient information. We 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. For x_t generated using the primal-dual subgradient update of Eq. (3) and $\delta \geq \max_t \|g_t\|_2$, for any $x^* \in \mathcal{X}$*

$$R_\phi(T) \leq \frac{\delta}{\eta} \|x^*\|_2^2 + \frac{1}{\eta} \|x^*\|_2^2 \operatorname{tr}(G_T^{1/2}) + \eta \operatorname{tr}(G_T^{1/2}). \tag{18}$$

For x_t generated with the composite mirror-descent update of Eq. (4), if $x^* \in \mathcal{X}$ and $\delta \geq 0$

$$R_\phi(T) \leq \frac{\delta}{\eta} \|x^*\|_2^2 + \frac{1}{2\eta} \max_{t \leq T} \|x^* - x_t\|_2^2 \operatorname{tr}(G_T^{1/2}) + \eta \operatorname{tr}(G_T^{1/2}). \tag{19}$$

Proof To begin, we consider the difference between the divergence terms at time $t+1$ and time t from Eq. (12) in Corollary 4. Let $\lambda_{\max}(M)$ denote the largest eigenvalue of a matrix M. We have

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

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 15) 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

$$\sum_{t=1}^{T-1} B_{\psi_{t+1}}(x^*, x_{t+1}) - B_{\psi_t}(x^*, x_{t+1}) \leq \frac{1}{2} \sum_{t=1}^{T-1} \|x^* - x_{t+1}\|_2^2 \left(\operatorname{tr}(G_{t+1}^{1/2}) - \operatorname{tr}(G_t^{1/2}) \right) .$$

Now we use the fact that G_1 is a rank 1 PSD matrix with non-negative trace to see that

$$\begin{aligned} &\sum_{t=1}^{T-1} \|x^* - x_{t+1}\|_2^2 \left(\operatorname{tr}(G_{t+1}^{1/2}) - \operatorname{tr}(G_t^{1/2}) \right) \\ &\qquad \leq \max_{t \leq T} \|x^* - x_t\|_2^2 \operatorname{tr}({G_T}^{1/2}) - \|x^* - x_1\|_2^2 \operatorname{tr}(G_1^{1/2}) . \end{aligned} \tag{20}$$

It remains 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 and $A^\dagger$ denote the pseudo-inverse of A. Then*

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

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

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

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

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

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

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

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

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

Using Lemma 17 in the appendix with the substitution $B = G_T$, $\nu = 1$, and $g = g_t$ lets us exploit the concavity of the function $\operatorname{tr}(A^{1/2})$ to bound the above sum by $2 \operatorname{tr}(G_T^{1/2})$. ▲

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 $\|g_t\|_{\psi_t^*}^2 \leq \left\langle g_t, S_t^{-1} g_t \right\rangle$. For the dual-ascent algorithms, we use Lemma 18 from the appendix to show that $\|g_t\|_{\psi_{t-1}^*}^2 \leq \left\langle g_t, S_t^{-1} g_t \right\rangle$ so long as $\delta \geq \|g_t\|_2$. Lemma 9's doubling inequality implies that $\sum_{t=1}^T \|f_t'(x_t)\|_{\psi_t^*}^2 \leq 2 \operatorname{tr}(G_T^{1/2})$ for the mirror-descent algorithms and that $\sum_{t=1}^T \|f_t'(x_t)\|_{\psi_{t-1}^*}^2 \leq 2 \operatorname{tr}(G_T^{1/2})$ for primal-dual subgradient algorithms.

Note that $B_{\psi_1}(x^*, x_1) \leq \frac{1}{2} \|x^* - x_1\|_2^2 \operatorname{tr}(G_1^{1/2})$ 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. (20), Corollary 4 gives the bound for the mirror-descent family of algorithms. Combining $\sum_{t=1}^T \|f_t'(x_t)\|_{\psi_{t-1}^*}^2 \leq 2 \operatorname{tr}(G_T^{1/2})$ and Eq. (20) 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 simplifies the bound implied by Theorem 8. The infimal equality in the corollary uses Lemma 19 in Appendix B. The corollary underscores that for learning problems in which there is a rotation U of the space for which the gradient vectors g_t have small inner products $\langle g_t, U g_t \rangle$ (essentially a sparse basis for the g_t) then using full-matrix proximal functions can attain significantly lower regret.

Corollary 10 *Assume 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*

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

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

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

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 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. For simplicity, we assume that each pair $f_t + \varphi$ has the same strong convexity parameter λ,

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

We focus on composite mirror descent algorithms here, as the analysis of strongly convex variants of primal-dual subgradient algorithms does not 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 of 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 composite mirror descent update in Eq. (4) and assume that φ is λ-strongly convex with respect to $\|\cdot\|$. For any $x^* \in \mathcal{X}$,*

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

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

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 algorithms, which use faster step rates to achieve logarithmic regret. The algorithm is identical to Alg. 1 except that the proximal function ψ_t grows faster.

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

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

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 $s_{t,i} = \|g_{1:t,i}\|_2^2$. Then*

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

Proof We begin by noting that for $a, b > 0$, the concavity of the logarithm 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

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

3. The tightest analysis of the strongly-convex primal-dual method keeps the function ψ constant rather than letting it grow. Allowing ψ to grow breaks the stronger regret.

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

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

■

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 $\mathcal{X}$. Assume further that $\|g\|_\infty \leq G_\infty$ for all $g \in \partial f_t(x)$ for $x \in \mathcal{X}$. 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*

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

6. Derived Algorithms

In this section, we derive updates using concrete regularization functions φ and settings of the domain $\mathcal{X}$ for the AdaGrad framework. We focus on showing how to solve Eqs. (3) and (4) with the diagonal matrix version of the algorithms we have presented. We focus on the diagonal case for two reasons. First, the updates often take closed-form in this case and carry some intuition. Second, the diagonal case is feasible to implement in very high dimensions, whereas the full matrix 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.

We begin by noting a simple but useful fact. Let G_t denote either the outer product matrix of gradients or its diagonal counterpart and let $H_t = \delta I + G_t^{1/2}$. Simple algebraic manipulations yield that each of the updates in the prequel, Eq. (3) and Eq. (4), can be written in the following form (omitting the stepsize η):

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle u, x \rangle + \varphi(x) + \frac{1}{2} \langle x, H_t x \rangle \right\} \ . \tag{21}$$

In particular, at time t for the RDA update, we have $u = \eta t \bar{g}_t$. For the composite gradient update of Eq. (4),

$$\eta \langle g_t, x \rangle + \frac{1}{2} \langle x - x_t, H_t(x - x_t) \rangle = \langle \eta g_t - H_t x_t, x \rangle + \frac{1}{2} \langle x, H_t x \rangle + \frac{1}{2} \langle x_t, H_t x_t \rangle$$

so that $u = \eta g_t - H_t x_t$. We now derive algorithms for solving Eq. (21).

6.1 ℓ_1-regularization

We begin by considering how to solve the minimization problems necessary for Alg. 1 with diagonal matrix divergences and $\varphi(x) = \lambda \|x\|_1$. 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 $H_{t,ii} = \delta + \|g_{1:t,i}\|_2$. For the primal-dual subgradient update, we need to solve Eq. (3), which amounts to the following:

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

Let $\hat{x}$ denote the optimal solution of the above optimization problem. Standard subgradient calculus implies that when $|\bar{g}_{t,i}| \leq \lambda$ 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:

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

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

$$x_{t+1,i} = \mathrm{sign}\left(-\bar{g}_{t,i}\right) \frac{\eta t}{H_{t,ii}} \left[|\bar{g}_{t,i}| - \lambda\right]_+ \ . \tag{22}$$

We can now compare Eq. (22) to the dual averaging update (Xiao, 2009), which is

$$x_{t+1,i} = \mathrm{sign}\left(-\bar{g}_{t,i}\right) \eta\sqrt{t} \left[|\bar{g}_{t,i}| - \lambda\right]_+ \ . \tag{23}$$

The difference between Eq. (22) and Eq. (23) 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. 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. (4). 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

$$x_{t+1,i} = \mathrm{sign}\left(x_{t,i} - \frac{\eta}{H_{t,ii}} g_{t,i}\right) \left[\left|x_{t,i} - \frac{\eta}{H_{t,ii}} g_{t,i}\right| - \frac{\lambda\eta}{H_{t,ii}}\right]_+ \ . \tag{24}$$

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 $H_{t,ii}$ remains intact. At time t when $x_{t,i}$ is needed, we rewrite the composite mirror-descent update,

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

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 $u_t = t\bar{g}_t = \sum_{\tau=1}^{t} g_\tau = u_{t-1} + g_t$ and rewrite the update in terms of u_t. We then use Eq. (22):

$$x_{t,i} = \operatorname{sign}(-u_{t,i}) \frac{\eta t}{H_{t,ii}} \left[\frac{|u_{t,i}|}{t} - \lambda \right]_+ ,$$

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

6.2 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. (4) we face the following optimization problem:

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

Assuming for simplicity that $\mathcal{X} = \mathbb{R}^d$, 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

$$x_{t+1} = ((\eta\lambda + \delta) I + \operatorname{diag}(s_t))^{-1} (\operatorname{diag}(s_t) x_t - \eta g_t) \ . \tag{25}$$

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. (25) $t - t_0$ 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. (25) to update the regularized predictors x, and if b is the bias term with associated gradients $g_{1:t,b}$, then $b_{t+1} = b_t - \frac{\eta}{\delta + \|g_{1:t,b}\|_2} g_{t,b}$.

6.3 ℓ_1-ball Projections

We next consider the setting in which $\varphi \equiv 0$ and $\mathcal{X} = \{x : \|x\|_1 \leq c\}$, which has been the topic of recent research (Duchi et al., 2008; Daubechies et al., 2008). We use the matrix $H_t = \delta I + \operatorname{diag}(G_t)^{1/2}$ from Alg. 1. We begin by converting the problem in Eq. (21) to a

projection task onto a scaled ℓ_1-ball. First, we make the substitution $z = H^{1/2}x$ so that the original problem in Eq. (21) is the same as solving

$$\min_z \left\langle u, H^{-1/2}z \right\rangle + \frac{1}{2}\|z\|_2^2 \quad \text{s.t.} \quad \left\|H^{-1/2}z\right\|_1 \leq c \ .$$

Further algebraic manipulations yield that the above problem is equivalent to the following:

$$\min_z \left\|z + H^{-1/2}u\right\|_2^2 \quad \text{s.t.} \quad \left\|H^{-1/2}z\right\|_1 \leq c \ .$$

We arrive at the minimization problem

$$\min_z \frac{1}{2}\|z - v\|_2^2 \quad \text{s.t.} \quad \|Az\|_1 \leq c \ , \tag{26}$$

where $v = -H^{-1/2}u = -\eta t H_t^{-1/2}\bar{g}_t$ for the primal-dual update of Eq. (3) and $v = H_t^{1/2}x_t - \eta H_t^{-1/2}g_t$ for the composite gradient update of Eq. (4). In both cases, $A = H_t^{-1/2}$ and x_{t+1} is recovered from the solution z^* through the equality $x_{t+1} = H_t^{-1/2}z^*$.

We derive an efficient algorithm for solving Eq. (26) when A is diagonal by building on the special case in which the constraint is $\|z\|_1 \leq c$ (Duchi et al., 2008). Consider

$$\min_z \frac{1}{2}\|z - v\|_2^2 \quad \text{s.t.} \quad \sum_{i=1}^d a_i|z_i| \leq c$$

for $a \succeq 0$. Due to the symmetry of the objective and constraints we assume without loss of generality that $v \succeq 0$. Otherwise, we can reverse the sign of each negative component in v, and once the solution is obtained reverse the sign of the corresponding component in the solution vector. We thus obtain the following problem:

$$\min_z \frac{1}{2}\|z - v\|_2^2 \quad \text{s.t.} \quad \langle a, z\rangle \leq c, \ z \succeq 0 \ . \tag{27}$$

Clearly, if $\langle a, v\rangle \leq c$ the optimal $z^* = v$, hence we assume that $\langle a, v\rangle > c$. We also assume without loss of generality that $v_i/a_i \geq v_{i+1}/a_{i+1}$ for simplicity of our derivation. (We revisit this assumption at the end of the derivation.) Introducing Lagrange multipliers $\theta \in \mathbb{R}_+$ for the constraint that $\langle a, z\rangle \leq c$ and $\alpha \in \mathbb{R}_+^d$ for the positivity constraint on z, we get

$$\mathcal{L}(z, \alpha, \theta) = \frac{1}{2}\|z - v\|_2^2 + \theta(\langle a, z\rangle - c) - \langle \alpha, z\rangle \ .$$

Computing the gradient of $\mathcal{L}$, we have $\nabla_z\mathcal{L}(z, \alpha, \theta) = z - v + \theta a - \alpha$. Suppose that we knew the optimal $\theta^* \geq 0$. Using the complimentarity conditions on z and α for optimality of z (Boyd and Vandenberghe, 2004), we see that the solution z_i^* satisfies

$$z_i^* = \begin{cases} v_i - \theta^* a_i & \text{if } v_i \geq \theta^* a_i \\ 0 & \text{otherwise} \ . \end{cases}$$

Analogously, the complimentary conditions on $\langle a, z\rangle \leq c$ show that given θ^*, we have

$$\sum_{i=1}^d a_i \left[v_i - \theta^* a_i\right]_+ = c \quad \text{or} \quad \sum_{i=1}^d a_i^2 \left[\frac{v_i}{a_i} - \theta^*\right]_+ = c \ .$$

Input: $v \succeq 0$, $a \succeq 0$, $c \geq 0$.
If $\sum_i v_i \leq c$ return $z^* = v$
Sort v_i/a_i into $\mu = \left[v_{i_j}/a_{i_j}\right]$ s.t. $v_{i_j}/a_{i_j} \geq v_{i_{j+1}}/a_{i_{j+1}}$
Set $\rho := \max\left\{\rho : \sum_{j=1}^{\rho} a_{i_j} v_{i_j} - \frac{v_{i_\rho}}{a_{i_\rho}} \sum_{j=1}^{\rho} a_{i_j}^2 < c\right\}$
Set $\theta = \frac{\sum_{j=1}^{\rho} a_{i_j} v_{i_j} - c}{\sum_{j=1}^{\rho} a_{i_j}^2}$
Return z^* where $z_i^* = [v_i - \theta a_i]_+$.

Figure 3: Project $v \succeq 0$ to $\langle a, z\rangle \leq c$ and $z \succeq 0$.

Conversely, had we obtained a value $\theta \geq 0$ satisfying the above equation, then θ would evidently induce the optimal z^* through the equation $z_i = [v_i - \theta a_i]_+$.

Now, let ρ be the largest index in $\{1, \ldots, d\}$ such that $v_i - \theta^* a_i > 0$ for $i \leq \rho$ and $v_i - \theta^* a_i \leq 0$ for $i > \rho$. From the assumption that $v_i/a_i \leq v_{i+1}/a_{i+1}$, we have $v_{\rho+1}/a_{\rho+1} \leq \theta^* < v_\rho/a_\rho$. Thus, had we known the last non-zero index ρ, we would have obtained

$$\sum_{i=1}^{\rho} a_i v_i - \frac{v_\rho}{a_\rho} \sum_{i=1}^{\rho} a_i^2 = \sum_{i=1}^{\rho} a_i^2 \left(\frac{v_i}{a_i} - \frac{v_\rho}{a_\rho}\right) < c \ ,$$
$$\sum_{i=1}^{\rho} a_i v_i - \frac{v_{\rho+1}}{a_{\rho+1}} \sum_{i=1}^{\rho} a_i^2 = \sum_{i=1}^{\rho+1} a_i^2 \left(\frac{v_i}{a_i} - \frac{v_{\rho+1}}{a_{\rho+1}}\right) \geq c \ .$$

Given ρ satisfying the above inequalities, we can reconstruct the optimal θ^* by noting that the latter inequality should equal c exactly when we replace v_ρ/a_ρ with θ, that is,

$$\theta^* = \frac{\sum_{i=1}^{\rho} a_i v_i - c}{\sum_{i=1}^{\rho} a_i^2} \ . \tag{28}$$

The above derivation results in the following procedure (when $\langle a, v\rangle > c$). We sort v in descending order of v_i/a_i and find the largest index ρ such that $\sum_{i=1}^{\rho} a_i v_i - (v_\rho/a_\rho) \sum_{i=1}^{\rho-1} a_i^2 < c$. We then reconstruct θ^* using Eq. (28) and return the soft-thresholded values of v_i. We provide the pseudo-code in Alg. 3. It is easy to verify that the algorithm can be implemented in $O(d \log d)$ time. Furthermore, a randomized search with bookkeeping as suggested by Duchi et al. (2008) can be straightforwardly used to derive a linear time algorithm.

6.4 ℓ_2 Regularization

We now turn to the case where $\varphi(x) = \lambda \|x\|_2$ while $\mathcal{X} = \mathbb{R}^d$. This type of regularization is useful for zeroing multiple weights in a group, for example in multi-task or multiclass learning (Obozinski et al., 2007). Recalling Eq. (21), we need to solve

$$\min_x \ \langle u, x\rangle + \frac{1}{2}\langle x, Hx\rangle + \lambda \|x\|_2 \ . \tag{29}$$

There is no closed form solution for this problem, but we give an efficient bisection-based procedure for solving Eq. (29). We start by deriving the dual. Introducing a variable $z = x$,

Input: $u \in \mathbb{R}^d$, $H \succeq 0$, $\lambda > 0$.
If $\|u\|_2 \leq \lambda$
  Return $x = 0$
Set $v = H^{-1}u$, $\theta_{\max} = \|v\|_2 / \lambda - 1/\sigma_{\min}(H)$
    $\theta_{\min} = \|v\|_2 / \lambda - 1/\sigma_{\max}(H)$
While $\theta_{\max} - \theta_{\min} > \varepsilon$
  Set $\theta = (\theta_{\max} + \theta_{\min})/2$, $\alpha(\theta) = -(H^{-1} + \theta I)^{-1} v$
  If $\|\alpha(\theta)\|_2 > \lambda$
    Set $\theta_{\min} = \theta$
  Else
    Set $\theta_{\max} = \theta$
Return $x = -H^{-1}(u + \alpha(\theta))$

Figure 4: Minimize $\langle u, x\rangle + \frac{1}{2}\langle x, Hx\rangle + \lambda \|x\|_2$

we get the equivalent problem of minimizing $\langle u, x\rangle + \frac{1}{2}\langle x, Hx\rangle + \lambda \|z\|_2$ subject to $x = z$. With Lagrange multipliers α for the equality constraint, we obtain the Lagrangian

$$\mathcal{L}(x, z, \alpha) = \langle u, x\rangle + \frac{1}{2}\langle x, Hx\rangle + \lambda \|z\|_2 + \langle \alpha, x - z\rangle \ .$$

Taking the infimum of $\mathcal{L}$ with respect to the primal variables x and z, we see that the infimum is attained at $x = -H^{-1}(u + \alpha)$. Coupled with the fact that $\inf_z \lambda \|z\|_2 - \langle \alpha, z\rangle = -\infty$ unless $\|\alpha\|_2 \leq \lambda$, in which case the infimum is 0, we arrive at the dual form

$$\inf_{x,z} \mathcal{L}(x, z, \alpha) = \begin{cases} -\frac{1}{2}\left\langle u + \alpha, H^{-1}(u + \alpha)\right\rangle & \text{if } \|\alpha\|_2 \leq \lambda \\ -\infty & \text{otherwise.} \end{cases}$$

Setting $v = H^{-1}u$, we further distill the dual to

$$\min_{\alpha} \ \langle v, \alpha\rangle + \frac{1}{2}\left\langle \alpha, H^{-1}\alpha\right\rangle \ \text{ s.t. } \|\alpha\|_2 \leq \lambda \ . \tag{30}$$

We can solve Eq. (30) efficiently using a bisection search of its equivalent representation in Lagrange form,

$$\min_{\alpha} \ \langle v, \alpha\rangle + \frac{1}{2}\left\langle \alpha, H^{-1}\alpha\right\rangle + \frac{\theta}{2}\|\alpha\|_2^2 \ ,$$

where $\theta > 0$ is an unknown scalar. The solution to the latter as a function of θ is clearly $\alpha(\theta) = -(H^{-1} + \theta I)^{-1} v = -(H^{-1} + \theta I)^{-1} H^{-1} u$. Since $\|\alpha(\theta)\|_2$ is monotonically decreasing in θ (consider the the eigen-decomposition of the positive definite H^{-1}), we can simply perform a bisection search over θ, checking at each point whether $\|\alpha(\theta)\|_2 \gtrless \lambda$.

To find initial upper and lower bounds on θ, we note that

$$(1/\sigma_{\max}(H) + \theta)^{-1} \|v\|_2 \leq \|\alpha(\theta)\|_2 \leq (1/\sigma_{\min}(H) + \theta)^{-1} \|v\|_2$$

where $\sigma_{\max}(H)$ denotes the maximum singular value of H and $\sigma_{\min}(H)$ the minimum. To guarantee $\|\alpha(\theta_{\max})\|_2 \leq \lambda$, we thus set $\theta_{\max} = \|v\|_2 / \lambda - 1/\sigma_{\max}(H)$. Similarly, for $\theta_{\min}$ we see that so long as $\theta \geq \|v\|_2 / \lambda - 1/\sigma_{\min}(H)$ we have $\|\alpha(\theta)\|_2 \geq \lambda$. The fact that $\partial \|x\|_2 = \{z : \|z\|_2 \leq 1\}$ when $x = 0$ implies that the optimal x for Eq. (29) is $x = 0$ if and only if $\|u\|_2 \leq \lambda$. We provide pseudocode for solving Eq. (29) in Alg. 4.

6.5 ℓ_∞ Regularization

We again let $\mathcal{X} = \mathbb{R}^d$ but now choose $\varphi(x) = \lambda \|x\|_\infty$. This type of update, similarly to ℓ_2, zeroes groups of variables, which is handy in finding structurally sparse solutions for multitask or multiclass problems. Solving the ℓ_∞ regularized problem amounts to

$$\min_x \ \langle u, x \rangle + \frac{1}{2} \langle x, Hx \rangle + \lambda \|x\|_\infty \ . \tag{31}$$

The dual of this problem is a modified ℓ_1-projection problem. As in the case of ℓ_2 regularization, we introduce an equality constrained variable $z = x$ with associated Lagrange multipliers $\alpha \in \mathbb{R}^d$ to obtain

$$\mathcal{L}(x, z, \alpha) = \langle u, x \rangle + \frac{1}{2} \langle x, Hx \rangle + \lambda \|z\|_\infty + \langle \alpha, x - z \rangle \ .$$

Performing identical manipulations to the ℓ_2 case, we take derivatives and get that $x = -H^{-1}(u + \alpha)$ and, similarly, unless $\|\alpha\|_1 \leq \lambda$, $\inf_z \mathcal{L}(x, z, \alpha) = -\infty$. Thus the dual problem for Eq. (31) is

$$\max_\alpha \ -\frac{1}{2}(u + \alpha) H^{-1}(u + \alpha) \ \text{ s.t. } \ \|\alpha\|_1 \leq \lambda \ .$$

When H is diagonal we can find the optimal α^* using the generalized ℓ_1-projection in Alg. 3, then reconstruct the optimal x via $x = -H^{-1}(u + \alpha^*)$.

6.6 Mixed-norm Regularization

Finally, we combine the above results to show how to solve problems with matrix-valued inputs $X \in \mathbb{R}^{d \times k}$, where $X = [\overline{x}_1 \ \cdots \ \overline{x}_d]^\top$. We consider mixed-norm regularization, which is very useful for encouraging sparsity across several tasks (Obozinski et al., 2007). Now φ is an ℓ_1/ℓ_p norm, that is, $\varphi(X) = \lambda \sum_{i=1}^d \|\overline{x}_i\|_p$. By imposing an ℓ_1-norm over p-norms of the rows of X, entire rows are nulled at once.

When $p \in \{2, \infty\}$ and the proximal H in Eq. (21) is diagonal, the previous algorithms can be readily used to solve the mixed norm problems. We simply maintain diagonal matrix information for each of the rows $\bar{x}_i$ of X separately, then solve one of the previous updates for each row independently. We use this form of regularization in our experiments with multiclass prediction problems in the next section.

7. Experiments

We performed experiments with several real world datasets with different characteristics: the ImageNet image database (Deng et al., 2009), the Reuters RCV1 text classification dataset (Lewis et al., 2004), the MNIST multiclass digit recognition problem, and the census income dataset from the UCI repository (Asuncion and Newman, 2007). For uniformity across experiments, we focus on the completely online (fully stochastic) optimization setting, in which at each iteration the learning algorithm receives a single example. We measure performance using two metrics: the online loss or error and the test set performance of the predictor the learning algorithm outputs at the end of a single pass through the training data. We also give some results that show how imposing sparsity constraints (in the

	RDA	FB	AdaGrad-RDA	AdaGrad-FB	PA	AROW
ECAT	.051 (.099)	.058 (.194)	**.044** (**.086**)	**.044** (.238)	.059	.049
CCAT	.064 (.123)	.111 (.226)	**.053** (**.105**)	**.053** (.276)	.107	.061
GCAT	.046 (.092)	.056 (.183)	**.040** (**.080**)	**.040** (.225)	.066	.044
MCAT	.037 (.074)	.056 (.146)	.035 (**.063**)	**.034** (.176)	.053	.039

Table 1: Test set error rates and proportion non-zero (in paranthesis) on Reuters RCV1.

form of ℓ_1 and mixed-norm regularization) affects the learning algorithm's performance. One benefit of the AdaGrad framework is its ability to straightforwardly generalize to domain constraints $\mathcal{X} \neq \mathbb{R}^d$ and arbitrary regularization functions φ, in contrast to previous adaptive online algorithms.

We experiment with RDA (Xiao, 2009), Fobos (Duchi and Singer, 2009), adaptive RDA, adaptive Fobos, the Passive-Aggressive (PA) algorithm (Crammer et al., 2006), and AROW (Crammer et al., 2009). To remind the reader, PA is an online learning procedure with the update

$$x_{t+1} = \operatorname*{argmin}_x \left[1 - y_t \left\langle z_t, x \right\rangle\right]_+ + \frac{\lambda}{2} \left\|x - x_t\right\|_2^2 \ ,$$

where λ is a regularization parameter. PA's update is similar to the update employed by AROW (see Eq. (10)), but the latter maintains second order information on x. By using a representer theorem it is also possible to derive efficient updates for PA and AROW when the loss is the logistic loss, $\log(1 + \exp(-y_t \left\langle z_t, x_t \right\rangle))$. We thus we compare the above six algorithms using both hinge and logistic loss.

7.1 Text Classification

The Reuters RCV1 dataset consists of a collection of approximately 800,000 text articles, each of which is assigned multiple labels. There are 4 high-level categories, Economics, Commerce, Medical, and Government (ECAT, CCAT, MCAT, GCAT), and multiple more specific categories. We focus on training binary classifiers for each of the four major categories. The input features we use are 0/1 bigram features, which, post word stemming, give data of approximately 2 million dimensions. The feature vectors are very sparse, however, and most examples have fewer than 5000 non-zero features.

We compare the twelve different algorithms mentioned in the prequel as well as variants of Fobos and RDA with ℓ_1-regularization. We summarize the results of the ℓ_1-regularized runs as well as AROW and PA in Table 1. The results for both hinge and logistic losses are qualitatively and quantitatively very similar, so we report results only for training with the hinge loss in Table 1. Each row in the table represents the average of four different experiments in which we hold out 25% of the data for a test set and perform an online pass on the remaining 75% of the data. For RDA and Fobos, we cross-validate the stepsize parameter η by simply running multiple passes and then choosing the output of the learner that had the fewest mistakes during training. For PA and AROW we choose λ using the same approach. We use the same regularization multiplier on the ℓ_1 term for RDA and Fobos, selected so that RDA achieved approximately 10% non-zero predictors.

It is evident from the results presented in Table 1 that the adaptive algorithms (AROW and AdaGrad) are far superior to non-adaptive algorithms in terms of error rate on test

Alg.	Avg. Prec.	P@1	P@3	P@5	P@10	Prop. nonzero
AdaGrad RDA	**0.6022**	0.8502	0.8307	0.8130	0.7811	**0.7267**
AROW	0.5813	**0.8597**	**0.8369**	**0.8165**	**0.7816**	1.0000
PA	0.5581	0.8455	0.8184	0.7957	0.7576	1.0000
RDA	0.5042	0.7496	0.7185	0.6950	0.6545	0.8996

Table 2: Test set precision for ImageNet

data. The AdaGrad algorithms naturally incorporate sparsity as well since they are run with ℓ_1-regularization, though RDA has significantly higher sparsity levels (PA and AROW do not have any sparsity). Furthermore, although omitted from the table to avoid clutter, in *every* test with the RCV1 corpus, the adaptive algorithms outperformed the non-adaptive algorithms. Moreover, both AdaGrad-RDA and AdaGrad-Fobos outperform AROW on all the classification tasks. Unregularized RDA and Fobos attained similar results as did the ℓ_1-regularized variants (of course without sparsity), but we omit the results to avoid clutter and because they do not give much more understanding.

7.2 Image Ranking

ImageNet (Deng et al., 2009) consists of images organized according to the nouns in the WordNet hierarchy, where each noun is associated on average with more than 500 images collected from the web. We selected 15,000 important nouns from the hierarchy and conducted a large scale image ranking task for *each* noun. This approach is identical to the task tackled by Grangier and Bengio (2008) using the Passive-Aggressive algorithm. To solve this problem, we train 15,000 ranking machines using Grangier and Bengio's visterms features, which represent patches in an image with 79-dimensional sparse vectors. There are approximately 120 patches per image, resulting in a 10,000-dimensional feature space.

Based on the results in the previous section, we focus on four algorithms for solving this task: AROW, AdaGrad with RDA updates and ℓ_1-regularization, vanilla RDA with ℓ_1, and Passive-Aggressive. We use the ranking hinge loss, which is $[1 - \langle x, z_1 - z_2 \rangle]_+$ when z_1 is ranked above z_2. We train a ranker x_c for each of the image classes individually, cross-validating the choice of initial stepsize for each algorithm on a small held-out set. To train an individual ranker for class c, at each step of the algorithm we randomly sample a positive image z_1 for the category c and an image z_2 from the training set (which with high probability is a negative example for class c) and perform an update on the example $z_1 - z_2$. We let each algorithm take 100,000 such steps for each image category, we train four sets of rankers with each algorithm, and the training set includes approximately 2 million images.

For evaluation, we use a distinct test set of approximately 1 million images. To evaluate a set of rankers, we iterate through all 15,000 classes in the dataset. For each class we take all the positive image examples in the test set and sample 10 times as many negative image examples. Following Grangier and Bengio, we then rank the set of positive and negative images and compute precision-at-k for $k = \{1, \ldots, 10\}$ and the average precision for each category. The precision-at-k is defined as the proportion of examples ranked in the top k for a category c that actually belong to c, and the average precision is the average of the precisions at each position in which a relevant picture appears. Letting $\text{Pos}(c)$ denote the positive examples for category c and $p(i)$ denote the position of the ith returned picture in

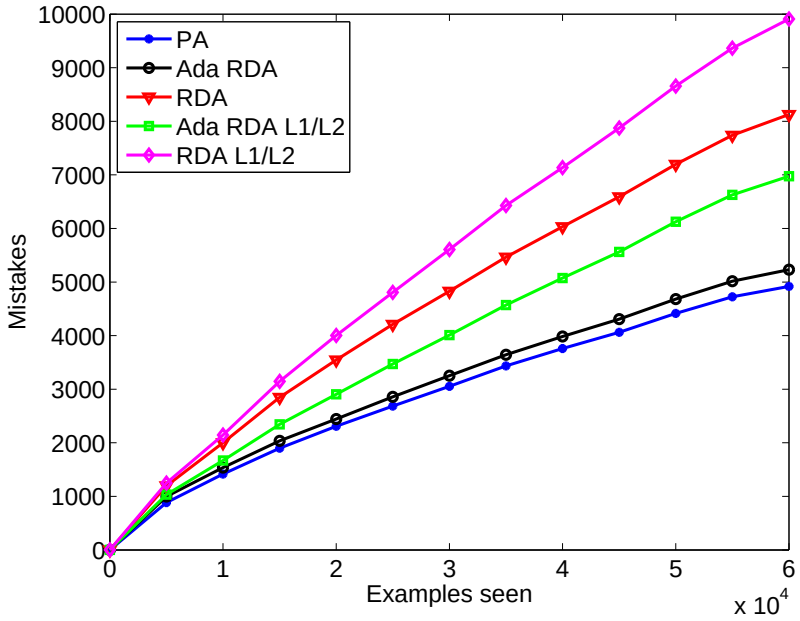

Figure 5: Learning curves on MNIST

list of images sorted by inner product with x_c, the average precision is

$$\frac{1}{|\text{Pos}(c)|} \sum_{i=1}^{|\text{Pos}(c)|} \frac{i}{p(i)} \ .$$

We compute the mean of each measurement across all classes, performing this twelve times for each of the sets of rankers trained. Table 2 summarizes our results. We do not report variance as the variance was on the order of 10^{-5} for each algorithm. One apparent characteristic to note from the table is that AdaGrad RDA achieves higher levels of sparsity than the other algorithms—using only 73% of the input features it achieves very high performance. Moreover, it outperforms all the algorithms in average precision. AROW has better results than the other algorithms in terms of precision-at-k for $k \leq 10$, though AdaGrad's performance catches up to and eventually surpasses AROW's as k grows.

7.3 Multiclass Optical Character Recognition

In the well-known MNIST multiclass classification dataset, we are given 28×28 pixel images a_i, and the learner's task is to classify each image as a digit in $\{0, \ldots, 9\}$. Linear classifiers do not work well on a simple pixel-based representation. Thus we learn classifiers built on top of a kernel machine with Gaussian kernels, as do Duchi and Singer (2009), which gives a different (and non-sparse) structure to the feature space in contrast to our previous experiments. In particular, for the ith example and jth feature, the feature value is $z_{ij} = K(a_i, a_j) \triangleq \exp\left(-\frac{1}{2\sigma^2} \|a_i - a_j\|_2^2\right)$. We use a support set of approximately 3000 images to compute the kernels and trained multiclass predictors, which consist of one vector $x_c \in \mathbb{R}^{3000}$ for each class c, giving a 30,000 dimensional problem. There is no known multiclass AROW

	Test error rate	Prop. nonzero
PA	0.062	1.000
Ada-RDA	0.066	1.000
RDA	0.108	1.000
Ada-RDA $\lambda = 5 \cdot 10^{-4}$	0.100	0.569
RDA $\lambda = 5 \cdot 10^{-4}$	0.138	0.878
Ada-RDA $\lambda = 10^{-3}$	0.137	0.144
RDA $\lambda = 10^{-3}$	0.192	0.532

Table 3: Test set error rates and sparsity proportions on MNIST. λ is the multiplier on the ℓ_1/ℓ_2 norm.

algorithm. We therefore compare adaptive RDA with and without mixed-norm ℓ_1/ℓ_2 and ℓ_1/ℓ_∞ regularization (see Sec. 6.6), RDA, and multiclass Passive Aggressive to one another using the multiclass hinge loss (Crammer et al., 2006). For each algorithm we used the first 5000 of 60,000 training examples to choose the stepsize η (for RDA) and λ (for PA).

In Fig. 5, we plot the learning curves (cumulative mistakes made) of multiclass PA, RDA, RDA with ℓ_1/ℓ_2 regularization, adaptive RDA, and adaptive RDA with ℓ_1/ℓ_2 regularization (ℓ_1/ℓ_∞ is similar). From the curves, we see that Adaptive RDA seems to have similar performance to PA, and the adaptive versions of RDA are vastly superior to their non-adaptive counterparts. Table 3 further supports this, where we see that the adaptive RDA algorithms outperform their non-adaptive counterparts both in terms of sparsity (the proportion of non-zero rows) and test set error rates.

7.4 Income Prediction

The KDD census income dataset from the UCI repository (Asuncion and Newman, 2007) contains census data extracted from 1994 and 1995 population surveys conducted by the U.S. Census Bureau. The data consists of 40 demographic and employment related variables which are used to predict whether a respondent has income above or below $50,000. We quantize each feature into bins (5 per feature for continuous features) and take products of features to give a 4001 dimensional feature space with 0/1 features. The data is divided into a training set of 199,523 instances and test set of 99,762 test instances.

As in the prequel, we compare AROW, PA, RDA, and adaptive RDA with and without ℓ_1-regularization on this dataset. We use the first 10,000 examples of the training set to select the step size parameters λ for AROW and PA and η for RDA. We perform ten experiments on random shuffles of the training data. Each experiment consists of a training pass through some proportion of the data (.05, .1, .25, .5, or the entire training set) and computing the test set error rate of the learned predictor. Table 4 and Fig. 6 summarize the results of these experiments. The variance of the test error rates is on the order of 10^{-6} so we do not report it. As earlier, the table and figure make it clear that the adaptive methods (AROW and AdaGrad-RDA) give better performance than non-adaptive methods. Further, as detailed in the table, the AdaGrad methods can give extremely sparse predictors that still give excellent test set performance. This is consistent with the experiments we have seen to this point, where AdaGrad gives sparse but highly accurate predictors.

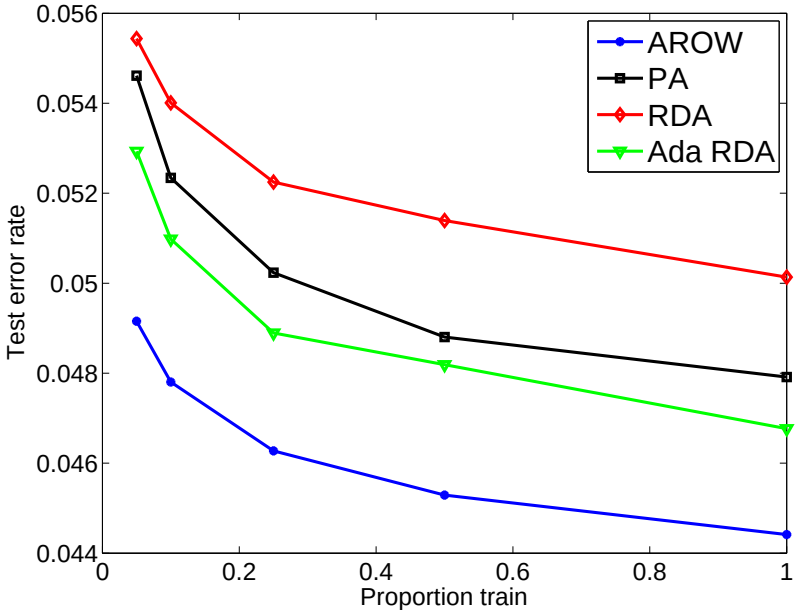

Figure 6: Test set error rates as function of proportion of training data seen on Census Income dataset.

Prop. Train	0.05	0.10	0.25	0.50	1.00
AROW	0.049	0.048	0.046	0.045	0.044
PA	0.055	0.052	0.050	0.049	0.048
RDA	0.055	0.054	0.052	0.051	0.050
Ada-RDA	0.053	0.051	0.049	0.048	0.047
ℓ_1 RDA	0.056 (0.075)	0.054 (0.066)	0.053 (0.058)	0.052 (0.053)	0.051 (0.050)
ℓ_1 Ada-RDA	0.052 (0.062)	0.051 (0.053)	0.050 (0.044)	0.050 (0.040)	0.049 (0.037)

Table 4: Test set error rates as function of proportion of training data seen (proportion of non-zeros in paranthesis where appropriate) on Census Income dataset.

8. Conclusions

We 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 regret bounds can be naturally converted into rate of convergence and generalization bounds (Cesa-Bianchi et al., 2004). Our experiments show that adaptive methods, specifically AdaGrad-Fobos, AdaGrad-RDA, and AROW clearly outperform their non-adaptive counterparts. Furthermore, the AdaGrad family of algorithms naturally incorporates regularization and gives very sparse solutions with similar performance to dense solutions. Our experiments with adaptive methods use a diagonal approximation to the matrix obtained by taking outer products of subgradients computed along the run of the

algorithm. It remains to be tested whether using the full outer product matrix can further improve performance.

To conclude we would like to underscore a possible elegant generalization that interpolates between full-matrix proximal functions and diagonal approximations using block diagonal matrices. Specifically, for $v \in \mathbb{R}^d$ let $v = [v_{[1]}^\top \ \cdots \ v_{[k]}^\top]^\top$ where $v_{[i]} \in \mathbb{R}^{d_i}$ are subvectors of v with $\sum_{i=1}^k d_i = d$. We can define the associated block-diagonal approximation to the outer product matrix $\sum_{\tau=1}^t g_\tau g_\tau^\top$ by

$$G_t = \sum_{\tau=1}^t \begin{bmatrix} g_{\tau,[1]} g_{\tau,[1]}^\top & 0 & \cdots & 0 \\ 0 & g_{\tau,[2]} g_{\tau,[2]}^\top & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & g_{\tau,[k]} g_{\tau,[k]}^\top \end{bmatrix}.$$

In this case, a combination of Theorems 6 and 8 gives the next corollary.

Corollary 14 *Let G_t be the block-diagonal outer product matrix defined above and the sequence $\{x_t\}$ be defined by the RDA update of Eq. (3) with $\psi_t(x) = \left\langle x, G_t^{1/2} x \right\rangle$. Then, for any $x^* \in \mathcal{X}$,*

$$R_\phi(T) \leq \frac{1}{\eta} \max_i \left\| x_{[i]}^* \right\|_2^2 \operatorname{tr}(G_T^{1/2}) + \eta \operatorname{tr}(G_T^{1/2}).$$

A similar bound holds for composite mirror-descent updates, and it is straightforward to get infimal equalities similar to those in Corollary 10 with the infimum taken over block-diagonal matrices. Such an algorithm can interpolate between the computational simplicity of the diagonal proximal functions and the ability of full matrices to capture correlation in the gradient vectors.

A few open questions stem from this line of research. The first is whether we can *efficiently* use full matrices in the proximal functions, as in Section 4. A second open issue is whether non-Euclidean proximal functions, such as the relative entropy, can be used. We hope to investigate both empirical and formal extensions of this work in the near future.

Acknowledgments

There are many people to whom we owe our sincere thanks for this research. Fernando Pereira helped push us in the direction of working on adaptive online methods and has been a constant source of discussion and helpful feedback. Samy Bengio provided us with a processed version of the ImageNet dataset and was instrumental in helping to get our experiments running, and Adam Sadovsky gave many indispensable coding suggestions. Lastly, Sam Roweis was a sounding board for some of our earlier ideas on the subject, and we will miss him dearly.

Appendix 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. (6) is lower bounded by

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

From the concavity of $\sqrt{\cdot}$,

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

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

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

However, $\sum_{i=1}^{d}\sqrt{i} \ge d\sqrt{d}/2$, thus we have that Eq. (6) 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 $\mathcal{X}$. 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 the projection step simply shrinks the multipliers α for x_t, the loss suffered by gradient descent suffers is at least as large as the value given by Eq. (6). Adaptive descent, for $t \le d$, constructs 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. The above properties are easiest to verify using an inductive argument. Clearly, since $g_1 = v_1$ and $||v_1|| = 1$, $G_1^2 = v_1 v_1^\top v_1 v_1^\top = v_1 I v_1^\top = G_1$ and we get, $G_1 = G_1^\dagger = G_1^{\frac{1}{2}}$. To construct x_2, adaptive descent sets

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

thus $\langle x_2, z_2\rangle = \pm\langle x_2, v_2\rangle = 0$ and $G_2 = G_1 + v_2 v_2^\top = v_1 v_1^\top + v_2 v_2^\top$. The same argument applies applied to G_1 is now also applicable to G_2 and also to all future constructions of G_t for $t \le d$. 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 $x_1, \ldots, x_{d+1}$ are all inside the domain $\mathcal{X} = \{x : \|x\|_2 \le \sqrt{d}\}$ no further projection is performed on the vectors. For adaptive descent, $\langle x_{d+1}, z_t\rangle = \left\langle 1, V^\top z_t\right\rangle = \left\langle \sum_{i=1}^{d} v_i, z_t\right\rangle$, and all future predictions achieve a margin of 1 and suffer zero losses. To recap, adaptive gradient descent suffers loss d, while projected gradient descent suffers loss at least $d\sqrt{d}$.

Appendix B. Technical Lemmas

Lemma 15 *Let $A \succeq B \succeq 0$ be symmetric $d \times d$ PSD matrices. Then $A^{1/2} \succeq B^{1/2}$.*

Proof This is Example 3 of Davis (1963). We include a proof for convenience of the reader. Let λ be any eigenvalue (with corresponding eigenvector x) of $A^{1/2} - B^{1/2}$; we show that $\lambda \ge 0$. Clearly $A^{1/2}x - \lambda x = B^{1/2}x$. Taking the inner product of both sides with $A^{1/2}x$, we have $\left\|A^{1/2}x\right\|_2^2 - \lambda\left\langle A^{1/2}x, x\right\rangle = \left\langle A^{1/2}x, B^{1/2}x\right\rangle$. We use the Cauchy-Schwarz inequality:

$$\left|\left\|A^{1/2}x\right\|_2^2 - \lambda\left\langle A^{1/2}x, x\right\rangle\right| \le \left\|A^{1/2}x\right\|_2 \left\|B^{1/2}x\right\|_2 = \sqrt{\langle Ax, x\rangle \langle Bx, x\rangle} \le \langle Ax, x\rangle = \left\|A^{1/2}x\right\|_2^2$$

where the last inequality follows from the assumption that $A \succeq B$. Thus we must have $\lambda\left\langle A^{1/2}x, x\right\rangle \ge 0$, which implies $\lambda \ge 0$. ■

The gradient of the function $\mathrm{tr}(X^p)$ 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 16 *Let $p \in \mathbb{R}$ and $X \succ 0$. Then $\nabla_X \mathrm{tr}(X^p) = pX^{p-1}$.*

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

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

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 have

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

which completes the proof. ■

Lemma 17 *Let $B \succeq 0$ and $B^{-1/2}$ denote the root of the inverse of B when $B \succ 0$ and the root of the pseudo-inverse of B otherwise. For any ν such that $B - \nu gg^\top \succeq 0$ the following inequality holds.*

$$2\operatorname{tr}((B - \nu gg^\top)^{1/2}) \leq 2\operatorname{tr}(B^{1/2}) - \nu \operatorname{tr}(B^{-1/2} gg^\top) \ .$$

Proof The core of the proof is based on the concavity of the function $\operatorname{tr}(A^{1/2})$. 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 $\operatorname{tr}(A^{1/2})$ is $\frac{1}{2}A^{-1/2}$ when $A \succ 0$.

First, A^p is matrix-concave for $A \succ 0$ and $0 \leq p \leq 1$ (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

$$(\alpha A + (1-\alpha)B)^p \succeq \alpha A^p + (1-\alpha)B^p \ . \tag{32}$$

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 15 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. (32) holds for $A, B \succeq 0$. Thus, we proved that

$$\operatorname{tr}((\alpha A + (1-\alpha)B)^p) \geq \alpha \operatorname{tr}(A^p) + (1-\alpha)\operatorname{tr}(B^p) \quad \text{for } 0 \leq p \leq 1 \ .$$

Recall now that Lemma 16 implies that the gradient of $\operatorname{tr}(A^{1/2})$ 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 $A, B \succ 0$,

$$\operatorname{tr}(A^{1/2}) \leq \operatorname{tr}(B^{1/2}) + \frac{1}{2}\operatorname{tr}(B^{-1/2}(A-B)) \ . \tag{33}$$

Let $A = B - \nu gg^\top \succeq 0$ 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 $2\nabla_B \operatorname{tr}((B+\delta I)^{1/2}) = (B+\delta I)^{-1/2}$, and $A - B = -\nu gg^\top$. From Eq. (33) and Lemma 15, we have

$$\begin{aligned} 2\operatorname{tr}(B - tgg^\top)^{1/2} &= 2\operatorname{tr}(A^{1/2}) \ \leq \ 2\operatorname{tr}((A+\delta I)^{1/2}) \\ &\leq 2\operatorname{tr}(B+\delta I)^{1/2} - \nu \operatorname{tr}((B+\delta I)^{-1/2} gg^\top) \ . \end{aligned} \tag{34}$$

Note that $g \in \operatorname{Range}(B)$, because if it were not, we could choose some u with $Bu = 0$ and $\langle g, u\rangle \neq 0$, which would give $\langle u, (B - cgg^\top)u\rangle = -c\langle g, u\rangle^2 < 0$, a contradiction. Now let $B = V \operatorname{diag}(\lambda) V^\top$ be the eigen-decomposition of B. Since $g \in \operatorname{Range}(B)$,

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

Thus, by taking $\delta \downarrow 0$ in Eq. (34), and since both $\operatorname{tr}(B+\delta I)^{1/2}$ and $\operatorname{tr}((B+\delta I)^{-1/2} gg^\top)$ are evidently continuous in δ, we complete the proof. ■

Lemma 18 *Let $\delta \geq \|g\|_2$ and $A \succeq 0$, then $\left\langle g, (\delta I + A^{1/2})^{-1} g \right\rangle \leq \left\langle g, \left((A + gg^\top)^\dagger\right)^{1/2} g \right\rangle$.*

Proof We begin by noting that $\delta^2 I \succeq gg^\top$, so from Lemma 15 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 $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, which holds for $a, b \geq 0$, 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 $g \in \text{Range}((A + gg^\top)^{1/2})$, we can apply an analogous limiting argument to the one used in the proof of Lemma 17 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

$$\min_S \ \operatorname{tr}(S^{-1}A) \ \ \text{subject to} \ \ S \succeq 0, \ \operatorname{tr}(S) \leq c \ \ \text{where} \ \ A \succeq 0 \ . \tag{35}$$

Lemma 19 *If A is of full rank, then the minimizer of Eq. (35) 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. (35). 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 derivative of $\mathcal{L}$ with respect to S is

$$-S^{-1}AS^{-1} + \theta I - Z. \tag{36}$$

If S is full rank, then to satisfy the generalized complimentarity conditions for the problem (Boyd and Vandenberghe, 2004), we must have $Z = 0$. Therefore, we get $S^{-1}AS^{-1} = \theta I$. 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. (35). From the above analysis and Eq. (36), 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. ■

Appendix C. 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. (13) 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 2.3 from Duchi et al. (2010). We provide the proof for completeness.

Corollary 20 *Let $\{x_t\}$ be the sequence defined by the update in Eq. (4) 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^*,*

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

Proof The optimality of x_{t+1} for Eq. (4) implies for all $x \in \mathcal{X}$ and $\varphi'(x_{t+1}) \in \partial\varphi(x_{t+1})$

$$\left\langle x - x_{t+1}, \eta f'(x_t) + \nabla\psi_t(x_{t+1}) - \nabla\psi_t(x_t) + \eta\varphi'(x_{t+1})\right\rangle \ge 0. \tag{37}$$

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

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

Now, by Eq. (37), the first term in the last equation is non-positive. Thus we have that

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

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 $\frac{1}{2}\left\|\cdot\right\|_{\psi_t}^2$ and $\frac{1}{2}\left\|\cdot\right\|_{\psi_t^*}^2$ (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 $\left\|\cdot\right\|_{\psi_t}$. ■

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. (3) now that ψ_t changes, however the step-size η is a constant and not time-dependent. We start by defining conjugate-like functions

$$U_t(g) = \sup_{x\in\mathcal{X}:\psi_t(x)\leq\psi_t(x^*)}\left[\langle g, x\rangle - t\varphi(x)\right] \tag{38}$$

$$V_t(g) = \sup_{x\in\mathcal{X}}\left[\langle g, x\rangle - t\varphi(x) - \frac{1}{\eta}\psi_t(x)\right]. \tag{39}$$

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

Proof We have

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

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

$$\pi_t(-g) = \operatorname*{argmin}_{x \in \mathcal{X}} \left[\langle g, x\rangle + t\varphi(x) + \frac{1}{\eta}\psi_t(x)\right] \;. \tag{40}$$

We next use the following standard properties of $V_t(g)$ (see, e.g. Nesterov, 2005, Theorem 1). 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):

$$V_t(g + h) \leq V_t(g) + \langle h, \nabla V_t(g)\rangle + \frac{\eta}{2}\left\|h\right\|_{\psi_t^*}^2 \;. \tag{41}$$

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 22 *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

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

▲

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

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

The above definition along with the convexity of f_τ implies

$$\delta_t \geq \sum_{\tau=1}^{t} \left[\langle g_\tau, x_\tau - x^* \rangle + \varphi(x_\tau) \right] - t\varphi(x^*) \geq \sum_{\tau=1}^{t} \left[f_\tau(x_\tau) - f_\tau(x^*) + \varphi(x_\tau) - \varphi(x^*) \right] . \quad (42)$$

Using Lemma 21, we upper bound δ_t by

$$\delta_t \ = \ \sum_{\tau=1}^{t} \left[\langle g_\tau, x_\tau \rangle + \varphi(x_\tau) \right] + U_t(-t\bar{g}_t) \ \leq \ \sum_{\tau=1}^{t} \left[\langle g_\tau, x_\tau \rangle + \varphi(x_\tau) \right] + V_t(-t\bar{g}_t) + \frac{1}{\eta}\psi_t(x^*) . \quad (43)$$

Finally, we upper bound V_t, and rearrange terms to obtain our desired inequality. First, Lemma 22 and Eq. (41) 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. (43), 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. (42) finishes the proof. ■

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*, 2008.

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

A. Asuncion and D. J. Newman. UCI machine learning repository, 2007. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

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.

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

A. Bordes L. Bottou and P. Gallinari. Sgd-qn: Careful quasi-newton stochastic gradient descent. *Journal of Machine Learning Research*, 10:1737–1754, 2009.

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, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

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.

I. Daubechies, M. Fornasier, and I. Loris. Accelerated projected gradient method for linear inverse problems with sparsity constraints. *Fourier Analysis and Applications*, 14(5): 764–792, 2008.

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. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

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 T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

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

R. Fletcher. A new approach to variable metric algorithms. *Computer Journal*, 13:317–322, 1970.

D. Grangier and S. Bengio. A discriminative kernel-based model to rank images from text queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1371–1384, 2008.

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.

David Lewis, Yiming Yang, Tony Rose, and Fan Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

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.

G. Obozinski, B. Taskar, and M. Jordan. Joint covariate selection for grouped classification. Technical Report 743, Dept. of Statistics, University of California Berkeley, 2007.

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`.

N. Z. Shor. Utilization of the operation of space dilation in the minimization of convex functions. *Cybernetics and Systems Analysis*, 6(1):7–15, 1972. Translated from *Kibernetika*.

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, Department of Mathematics, University of Washington, 2008.

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.