Composite Objective Mirror Descent

John C. Duchi
UC Berkeley
jduchi@cs.berkeley.edu

Shai Shalev-Shwartz
Hebrew University
shais@cs.huji.ac.il

Yoram Singer
Google Research
singer@google.com

Ambuj Tewari
TTI Chicago
tewari@ttic.edu

Abstract

We present a new method for regularized convex optimization and analyze it under both online and stochastic optimization settings. In addition to unifying previously known first-order algorithms, such as the projected gradient method, mirror descent, and forward-backward splitting, our method yields new analysis and algorithms. We also derive specific instantiations of our method for commonly used regularization functions, such as ℓ_1, mixed norm, and trace-norm.

1 Introduction and Problem Statement

Regularized loss minimization is a common learning paradigm in which one jointly minimizes an empirical loss over a training set plus a regularization term. The paradigm yields an optimization problem of the form

$$\min_{\boldsymbol{w}\in\Omega} \frac{1}{n}\sum_{t=1}^{n} f_t(\boldsymbol{w}) + r(\boldsymbol{w}) \ , \tag{1}$$

where $\Omega \subset \mathbb{R}^d$ is the domain (a closed convex set), $f_t : \Omega \to \mathbb{R}$ is a (convex) loss function associated with a single example in a training set, and $r : \Omega \to \mathbb{R}$ is a (convex) regularization function. A few examples of famous learning problems that fall into this framework are least squares, ridge regression, support vector machines, support vector regression, lasso, and logistic regression.

In this paper, we describe and analyze a general framework for solving Eq. (1). The method we propose is a first-order approach, meaning that we access the functions f_t only by receiving subgradients. Recent work has shown that from the perspective of achieving good statistical performance on unseen data, first order methods are preferable to higher order approaches, especially when the number of training examples n is very large (Bottou and Bousquet, 2008; Shalev-Shwartz and Srebro, 2008). Furthermore, in large scale problems it is often prohibitively expensive to compute the gradient of the entire objective function (thus accessing all the examples in the training set), and randomly choosing a subset of the training set and computing the gradient over the subset (perhaps only a single example) can be significantly more efficient. This approach is very closely related to online learning. Our general framework handles both cases with ease—it applies to accessing a single example (or subset of the examples) at each iteration or accessing the entire training set at each iteration.

The method we describe is an adaptation of the Mirror Descent (MD) algorithm (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003), an iterative method for minimizing a convex function $\phi : \Omega \to \mathbb{R}$. If the dimension d is large enough, MD is optimal among first-order methods, and it has a close connection to online learning since it is possible to bound the regret

$$\sum_{t=1}^{T} \phi_t(\boldsymbol{w}_t) - \inf_{\boldsymbol{w}\in\Omega} \sum_{t=1}^{T} \phi_t(\boldsymbol{w}) \ ,$$

where $\{\boldsymbol{w}_t\}$ is the sequence generated by mirror descent and the ϕ_t are convex functions. In fact, one can view popular online learning algorithms, such as weighted majority (Littlestone and Warmuth, 1994) and online gradient descent (Zinkevich, 2003) as special cases of mirror descent. A guarantee on the online regret can be translated directly to a guarantee on the convergence rate of the algorithm to the optimum of Eq. (1), as we will show later.

Following Beck and Teboulle's exposition, a Mirror Descent update in the online setting can be written as

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\Omega} \; B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta \left\langle \phi_t'(\boldsymbol{w}_t), \boldsymbol{w} - \boldsymbol{w}_t \right\rangle , \tag{2}$$

where B_ψ is a Bregman divergence and ϕ_t' denotes an arbitrary subgradient of ϕ_t. Intuitively, MD minimizes a first-order approximation of the function ϕ_t at the current iterate $\boldsymbol{w}_t$ while forcing the next iterate $\boldsymbol{w}_{t+1}$ to lie close to $\boldsymbol{w}_t$. The step-size η controls the trade-off between these two.

Our focus in this paper is to generalize mirror descent to the case when the functions ϕ_t are composite, that is, they consist of two parts: $\phi_t = f_t + r$. Here the f_t change over time but the function r remains constant. Of course, one can ignore the composite structure of the ϕ_t and use MD. However, doing so can result in undesirable effects. For example, when $r(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$, applying MD directly does not lead to sparse updates. Since the sparsity inducing property of the ℓ_1-norm is a major reason for its use. The modification of mirror descent that we propose is simple:

$$\boldsymbol{w}_{t+1} \triangleq \operatorname*{argmin}_{\boldsymbol{w}\in\Omega} \; \eta \left\langle f_t'(\boldsymbol{w}_t), \boldsymbol{w} \right\rangle + B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta r(\boldsymbol{w}). \tag{3}$$

This is almost the same as the mirror descent update with an important difference: we *do not* linearize r. We call this algorithm Composite Objective MIrror Descent, or COMID. One of our contributions is to show that, in a variety of cases, the COMID update is no costlier than the usual mirror descent update. In these situations, each COMID update is efficient and benefits from the presence of the regularizer $r(\boldsymbol{w})$.

We now outline the remainder of the paper. We begin by reviewing related work, of which there is a copious amount, though we try to do some justice to prior research. We then give a general $O(\sqrt{T})$ regret bound for COMID in the online optimization setting, after which we give several extensions. We show $O(\log T)$ regret bounds for COMID when the composite functions $f_t + r$ are strongly convex, after which we show convergence rates and concentration results for stochastic optimization using COMID. The second focus of the paper is in the derived algorithms, where we outline step rules for several choices of Bregman function ψ and regularizer r, including ℓ_1, ℓ_∞, and mixed-norm regularization, as well as presenting new results on efficient matrix optimization with Schatten p-norms.

2 Related Work

Since the idea underlying COMID is simple, it is not surprising that similar algorithms have been proposed. One of our main contributions is to show that COMID generalizes much prior work and to give a clean unifying analysis. We do not have the space to thoroughly review the literature, though we try to do some small justice to what is known. We begin by reviewing work that we will show is a special case of COMID. Forward-backward splitting is a long-studied framework for minimizing composite objective functions (Lions and Mercier, 1979), though it has only recently been analyzed for the online and stochastic case (Duchi and Singer, 2009). Specializations of forward-backward splitting to the case where $r(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$ include iterative shrinkage and thresholding from the signal processing literature (Daubechies et al., 2004), and from machine learning, Truncated Gradient (Langford et al., 2009) and SMIDAS (Shalev-Shwartz and Tewari, 2009) are both special cases of COMID.

In the optimization community there has there has been significant recent interest—both applied and theoretical—on minimization of composite objective functions such as that in Eq. (1). Some notable examples include Wright et al. (2009); Nesterov (2007); Tseng (2009). These papers all assume that the objective $f + r$ to be minimized is fixed and that f is smooth, i.e. that it has Lipschitz continuous derivatives. The most related of these to COMID is probably Tseng (2009, see his Sec. 3.1 and the references therein), which proposes the same update as ours, but gives a Nesterov-like optimal method for the fixed f case. We do not have restrictions on f, though by going to stochastic, nondifferentiable f we naturally suffer in convergence rate. Nonetheless, we do answer in the affirmative a question posed by Tseng (2009), which is whether stochastic or incremental subgradient methods work for composite objectives.

Two recent papers for online and stochastic composite objective minimization are Xiao (2009) and Duchi and Singer (2009). The former extends Nesterov's 2009 analysis of primal-dual subgradient methods to the composite case, giving an algorithm which is similar to ours; however, our algorithms are different and the analysis for each is completely different. Duchi and Singer (2009) is simply a specialization of COMID to the case where the Euclidean Bregman divergence is used.

As a consequence of our general setting, we are able to give elegant new algorithms for minimization of functions on matrices, which include efficient and simple algorithms for trace-norm

minimization. Trace norm minimization has recently found strong applicability in matrix rank minimization (Recht et al., 2007), which has been shown to be very useful, for example, in collaborative filtering (Srebro et al., 2004). A special case of COMID has recently been developed for this task, which is very similar in spirit to fixed point and shrinkage methods from signal processing for ℓ_1-minimization (Ma et al., 2009). The authors of this paper note that the method is extremely efficient for rank-minimization problems but do not give rates of convergence, which we give as a corollary to our main convergence theorems.

3 Notation and Setting

Before continuing, we establish notation and our problem setting formally. Vectors are lower case bold italic letters, such as $\boldsymbol{x} \in \mathbb{R}^d$, and scalars are lower case italics such as $x \in \mathbb{R}$. We denote a sequence of vectors by subscripts, i.e. $\boldsymbol{w}_t, \boldsymbol{w}_{t+1}, \ldots$, and entries in a vector by non-bold subscripts as in w_j. Matrices are upper case bold italic letters, such as $\boldsymbol{W} \in \mathbb{R}^{d \times d}$. The subdifferential set of a function f evaluated at $\boldsymbol{w}$ is denoted $\partial f(\boldsymbol{w})$ and a particular subgradient by $f'(\boldsymbol{w}) \in \partial f(\boldsymbol{w})$. When a function is differentiable, we write $\nabla f(\boldsymbol{w})$.

We focus mostly on the problem of regularized online learning, in which the goal is to achieve low regret w.r.t. a static predictor $\boldsymbol{w}^* \in \Omega$ on a sequence of functions $\phi_t(\boldsymbol{w}) \triangleq f_t(\boldsymbol{w}) + r(\boldsymbol{w})$. Here, f_t and $r \geq 0$ are convex functions, and Ω is some convex set (which could be $\mathbb{R}^d$). Formally, at every round of the algorithm we make a prediction $\boldsymbol{w}_t \in \mathbb{R}^d$ and then receive the function f_t. We seek bounds on the *regularized regret* with respect to $\boldsymbol{w}^*$, defined as

$$R_\phi(T, \boldsymbol{w}^*) \triangleq \sum_{t=1}^{T} \left[f_t(\boldsymbol{w}_t) + r(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*) - r(\boldsymbol{w}^*) \right]. \tag{4}$$

In batch optimization we set $f_t = f$ for all t, while in stochastic optimization we choose f_t to be the average of some random subset of $\{f_1, \ldots, f_n\}$. As mentioned previously and as we will show, it is not difficult to transform regret bounds for Eq. (4) into convergence rates in expectation and with high probability for Eq. (1), which we do using techniques similar to Cesa-Bianchi et al. (2004).

Throughout, ψ designates a continuously differentiable function that is α-strongly convex w.r.t. a norm $\|\cdot\|$ on the set Ω. Recall that this means that the Bregman divergence associated with ψ,

$$B_\psi(\boldsymbol{w}, \boldsymbol{v}) = \psi(\boldsymbol{w}) - \psi(\boldsymbol{v}) - \langle \nabla\psi(\boldsymbol{v}), \boldsymbol{w} - \boldsymbol{v} \rangle ,$$

satisfies $B_\psi(\boldsymbol{w}, \boldsymbol{v}) \geq \frac{\alpha}{2}\|\boldsymbol{w} - \boldsymbol{v}\|^2$ for some $\alpha > 0$.

4 Composite Objective MIrror Descent

We use proof techniques similar to those in Beck and Teboulle (2003) to derive "progress" bounds on each step of the algorithm. We then use the bounds to straightforwardly prove convergence results for online and batch learning. We begin by bounding the progress made by each step of the algorithm in either an online or a batch setting. This lemma is the key to our later analysis, so we prove it in full here.

Lemma 1 *Let the sequence $\{\boldsymbol{w}_t\}$ be defined by the update in Eq. (3). Assume that $B_\psi(\cdot,\cdot)$ is α-strongly convex with respect to a norm $\|\cdot\|$, that is, $B_\psi(\boldsymbol{w}, \boldsymbol{v}) \geq \frac{\alpha}{2}\|\boldsymbol{w} - \boldsymbol{v}\|^2$. For any $\boldsymbol{w}^* \in \Omega$,*

$$\eta \left(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*) \right) + \eta \left(r(\boldsymbol{w}_{t+1}) - r(\boldsymbol{w}^*) \right) \leq B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) - B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}) + \frac{\eta^2}{2\alpha} \left\| f_t'(\boldsymbol{w}_t) \right\|_*^2 .$$

Proof: The optimality of $\boldsymbol{w}_{t+1}$ for Eq. (3) implies for all $\boldsymbol{w} \in \Omega$ and $r'(\boldsymbol{w}_{t+1}) \in \partial r(\boldsymbol{w}_{t+1})$,

$$\langle \boldsymbol{w} - \boldsymbol{w}_{t+1}, \eta f'(\boldsymbol{w}_t) + \nabla\psi(\boldsymbol{w}_{t+1}) - \nabla\psi(\boldsymbol{w}_t) + \eta r'(\boldsymbol{w}_{t+1}) \rangle \geq 0. \tag{5}$$

In particular, this obtains for $\boldsymbol{w} = \boldsymbol{w}^*$. From the subgradient inequality for convex functions, we have $f_t(\boldsymbol{w}^*) \geq f_t(\boldsymbol{w}_t) + \langle f_t'(\boldsymbol{w}_t), \boldsymbol{w}^* - \boldsymbol{w}_t \rangle$, or $f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*) \leq \langle f_t'(\boldsymbol{w}_t), \boldsymbol{w}_t - \boldsymbol{w}^* \rangle$, and likewise for $r(\boldsymbol{w}_{t+1})$. We thus have

$$\begin{aligned}
&\eta \left[f_t(\boldsymbol{w}_t) + r(\boldsymbol{w}_{t+1}) - f_t(\boldsymbol{w}^*) - r(\boldsymbol{w}^*) \right] \\
&\leq \eta \left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f_t'(\boldsymbol{w}_t) \right\rangle + \eta \left\langle \boldsymbol{w}_{t+1} - \boldsymbol{w}^*, r'(\boldsymbol{w}_{t+1}) \right\rangle \\
&= \eta \left\langle \boldsymbol{w}_{t+1} - \boldsymbol{w}^*, f_t'(\boldsymbol{w}_t) \right\rangle + \eta \left\langle \boldsymbol{w}_{t+1} - \boldsymbol{w}^*, r'(\boldsymbol{w}_{t+1}) \right\rangle + \eta \left\langle \boldsymbol{w}_t - \boldsymbol{w}_{t+1}, f_t'(\boldsymbol{w}_t) \right\rangle \\
&= \left\langle \boldsymbol{w}^* - \boldsymbol{w}_{t+1}, \nabla\psi(\boldsymbol{w}_t) - \nabla\psi(\boldsymbol{w}_{t+1}) - \eta f_t'(\boldsymbol{w}_t) - \eta r'(\boldsymbol{w}_{t+1}) \right\rangle + \left\langle \boldsymbol{w}^* - \boldsymbol{w}_{t+1}, \nabla\psi(\boldsymbol{w}_{t+1}) - \nabla\psi(\boldsymbol{w}_t) \right\rangle \\
&\quad + \eta \left\langle \boldsymbol{w}_t - \boldsymbol{w}_{t+1}, f_t'(\boldsymbol{w}_t) \right\rangle .
\end{aligned}$$

Now, by Eq. (5), the first term in the last equation is non-positive. Thus we have that

$$\begin{aligned}
&\eta\left[f_t(\boldsymbol{w}_t) + r(\boldsymbol{w}_{t+1}) - f_t(\boldsymbol{w}^*) - r(\boldsymbol{w}^*)\right] \\
&\quad\leq \left\langle \boldsymbol{w}^* - \boldsymbol{w}_{t+1}, \nabla\psi(\boldsymbol{w}_{t+1}) - \nabla\psi(\boldsymbol{w}_t)\right\rangle + \eta\left\langle \boldsymbol{w}_t - \boldsymbol{w}_{t+1}, f_t'(\boldsymbol{w}_t)\right\rangle \\
&\quad= B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) - B_\psi(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) - B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}) + \eta\left\langle \boldsymbol{w}_t - \boldsymbol{w}_{t+1}, f_t'(\boldsymbol{w}_t)\right\rangle \qquad (6) \\
&\quad= B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) - B_\psi(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) - B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}) + \eta\left\langle \sqrt{\frac{\alpha}{\eta}}(\boldsymbol{w}_t - \boldsymbol{w}_{t+1}), \sqrt{\frac{\eta}{\alpha}} f_t'(\boldsymbol{w}_t)\right\rangle \\
&\quad\leq B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) - B_\psi(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) - B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}) + \frac{\alpha}{2}\left\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\right\|^2 + \frac{\eta^2}{2\alpha}\left\|f_t'(\boldsymbol{w}_t)\right\|_*^2 \\
&\quad\leq B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) - B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}) + \frac{\eta^2}{2\alpha}\left\|f_t'(\boldsymbol{w}_t)\right\|_*^2.
\end{aligned}$$

In the above, the first equality follows from simple algebra of B_ψ, that is, $\langle\nabla\psi(\boldsymbol{b}) - \nabla\psi(\boldsymbol{a}), \boldsymbol{c} - \boldsymbol{a}\rangle = B_\psi(\boldsymbol{c}, \boldsymbol{a}) + B_\psi(\boldsymbol{a}, \boldsymbol{b}) - B_\psi(\boldsymbol{c}, \boldsymbol{b})$ and setting $\boldsymbol{c} = \boldsymbol{w}^*$, $\boldsymbol{a} = \boldsymbol{w}_{t+1}$, and $\boldsymbol{b} = \boldsymbol{w}_t$. The second to last inequality follows from the Fenchel-Young inequality applied to the conjugate pair $\frac{1}{2}\|\cdot\|^2$, $\frac{1}{2}\|\cdot\|_*^2$ (Boyd and Vandenberghe, 2004, Example 3.27). The last inequality follows from the strong convexity of B_ψ with respect to the norm $\|\cdot\|$. ■

The following theorem uses Lemma 1 to establish a general regret bound for the COMID framework.

Theorem 2 *Let the sequence $\{\boldsymbol{w}_t\}$ be defined by the update in Eq. (3). Then for any $\boldsymbol{w}^* \in \Omega$,*

$$R_\phi(T, \boldsymbol{w}^*) \leq \frac{1}{\eta} B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1) + r(\boldsymbol{w}_1) + \frac{\eta}{2\alpha}\sum_{t=1}^T \left\|f_t'(\boldsymbol{w}_t)\right\|_*^2.$$

Proof: By Lemma 1,

$$\begin{aligned}
\eta\sum_{t=1}^T\left[f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*) + r(\boldsymbol{w}_{t+1}) - r(\boldsymbol{w}^*)\right] &\leq \sum_{t=1}^T B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) - B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}) + \frac{\eta^2}{2\alpha}\sum_{t=1}^T\left\|f_t'(\boldsymbol{w}_t)\right\|_*^2 \\
&= B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1) - B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{T+1}) + \frac{\eta^2}{2\alpha}\sum_{t=1}^T\left\|f_t'(\boldsymbol{w}_t)\right\|_*^2.
\end{aligned}$$

Noting that Bregman divergences are always non-negative, recall our assumption that $r(\boldsymbol{w}) \geq 0$. Adding $\eta r(\boldsymbol{w}_1)$ to both sides of the above equation and dropping the $r(\boldsymbol{w}_{t+1})$ term gives

$$\eta\sum_{t=1}^T\left[f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*) + r(\boldsymbol{w}_t) - r(\boldsymbol{w}^*)\right] \leq B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1) + \eta r(\boldsymbol{w}_1) + \frac{\eta^2}{2\alpha}\sum_{t=1}^T\left\|f_t'(\boldsymbol{w}_t)\right\|_*^2.$$

Dividing each side by η gives the result. ■

A few corollaries are immediate from the above result. First, suppose that the functions f_t are Lipschitz continuous. Then there is some G_* such that $\|f_t'(\boldsymbol{w}_t)\|_* \leq G_*$. In this case, we have

Corollary 3 *Let $\{\boldsymbol{w}_t\}$ be generated by the update Eq. (3) and assume that the functions f_t are Lipschitz with dual Lipschitz constant G_*. Then*

$$R_\phi(T) \leq \frac{1}{\eta} B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1) + r(\boldsymbol{w}_1) + \frac{T\eta}{2\alpha} G_*^2.$$

If we take $\eta \propto 1/\sqrt{T}$, then we have a regret which is $O(\sqrt{T})$ when the functions f_t are Lipschitz. If Ω is compact, the f_t are guaranteed to be Lipschitz continuous (Rockafellar, 1970).

Corollary 4 *Suppose that either Ω is compact or the functions f_t are Lipschitz so $\|f_t'\|_* \leq G_*$. Also assume $r(\boldsymbol{w}_1) = 0$. Then setting $\eta = \sqrt{2\,\alpha\, B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1)}/(G_*\sqrt{T})$,*

$$R_\phi(T) \leq \sqrt{2TB_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1)}\, G_*/\sqrt{\alpha}.$$

It is straightforward to prove results under the slightly different restriction that $\|f_t'(\boldsymbol{w})\|_*^2 \leq \rho f_t(\boldsymbol{w})$, which is similar to assuming a Lipschitz condition on the gradient of f_t. A common example in which this holds is linear regression, where $f_i(\boldsymbol{w}) = \frac{1}{2}(\langle\boldsymbol{w}, \boldsymbol{x}_i\rangle - y_i)^2$, so $\nabla f_i(\boldsymbol{w}) = (\langle\boldsymbol{w}, \boldsymbol{x}_i\rangle - y_i)\boldsymbol{x}_i$ and $\rho = \frac{1}{2}\|\boldsymbol{x}_i\|^2$. The proof essentially amounts to dividing out constants dependent on η and ρ from both sides of the regret.

Corollary 5 *Let $\|f_t'(\boldsymbol{w})\|_*^2 \le \rho f_t(\boldsymbol{w})$, $r \ge 0$, and assume $r(\boldsymbol{w}_1) = 0$. Setting $\eta \propto 1/\sqrt{T}$ gives*

$$R_\phi(T) = O(\rho\sqrt{T}B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1)/\alpha) \ .$$

Proof: From Theorem 2, the non-negativity of r, and that $r(\boldsymbol{w}_1) = 0$ we immediately have

$$\sum_{t=1}^T \left(1 - \frac{\rho\eta}{2\alpha}\right)[f_t(\boldsymbol{w}_t) + r(\boldsymbol{w}_t)] \le \sum_{t=1}^T \left(1 - \frac{\rho\eta}{2\alpha}\right) f_t(\boldsymbol{w}_t) + r(\boldsymbol{w}_t) \le \frac{1}{\eta}B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1) + \sum_{t=1}^T f_t(\boldsymbol{w}^*) + r(\boldsymbol{w}^*)$$

Setting $\eta = 2\alpha/(\rho\sqrt{T})$ gives $1 - \rho\eta/(2\alpha) = (\sqrt{T}-1)/\sqrt{T}$ so that

$$\sum_{t=1}^T f_t(\boldsymbol{w}_t) + r(\boldsymbol{w}_t) \le \frac{\rho T}{2\alpha(\sqrt{T}-1)}B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1) + \frac{\sqrt{T}}{\sqrt{T}-1}\sum_{t=1}^T f_t(\boldsymbol{w}^*) + r(\boldsymbol{w}^*).$$

■

5 Logarithmic Regret for Strongly Convex Functions

Following the vein of research begun in Hazan et al. (2006) and Shalev-Shwartz and Singer (2007), we show that COMID can get stronger regret guarantees when we assume curvature of the loss functions f_t or r. Similar to Shalev-Shwartz and Singer, we now assume that for all t, $f_t + r$ is λ-strongly convex with respect to a differentiable function ψ, that is, for any $\boldsymbol{w}, \boldsymbol{v} \in \Omega$,

$$f_t(\boldsymbol{v}) + r(\boldsymbol{v}) \ge f_t(\boldsymbol{w}) + r(\boldsymbol{w}) + \langle f_t'(\boldsymbol{w}) + r'(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w}\rangle + \lambda B_\psi(\boldsymbol{v}, \boldsymbol{w}). \tag{7}$$

For example, when $\psi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$, we recover the usual definition of strong convexity. For simplicity, we assume that we push all the strong convexity into the function r so that the f_t are simply convex (clearly, this is possible by redefining $\hat{f}_t(\boldsymbol{w}) = f_t(\boldsymbol{w}) - \lambda\psi(\boldsymbol{w})$ if the f_t are λ-strongly convex). In this case, a straightforward corollary to Lemma 1 follows.

Corollary 6 *Let the sequence $\{\boldsymbol{w}_t\}$ be defined by the update in Eq. (3) with step sizes η_t. Assume that $B_\psi(\cdot,\cdot)$ is α-strongly convex with respect to a norm $\|\cdot\|$ and that r is λ-strongly convex with respect to ψ. Then for any $\boldsymbol{w}^* \in \Omega$*

$$\begin{aligned}&\eta_t\left(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)\right) + \eta_t\left(r(\boldsymbol{w}_{t+1}) - r(\boldsymbol{w}^*)\right)\\ &\qquad \le B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) - B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}) + \frac{\eta_t^2}{2\alpha}\|f_t'(\boldsymbol{w}_t)\|_*^2 - \lambda\eta_t B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}).\end{aligned}$$

Proof: The proof is effectively identical to that of Lemma 1. We simply note that $r(\boldsymbol{w}_{t+1}) - r(\boldsymbol{w}^*) \le \langle r'(\boldsymbol{w}_{t+1}), \boldsymbol{w}_{t+1} - \boldsymbol{w}^*\rangle - \lambda B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1})$ so that

$$\begin{aligned}&\eta_t\left[f_t(\boldsymbol{w}_t) + r(\boldsymbol{w}_{t+1}) - f_t(\boldsymbol{w}^*) - r(\boldsymbol{w}^*)\right]\\ &\quad\le\quad \eta_t\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f_t'(\boldsymbol{w}_t)\rangle + \eta_t\langle \boldsymbol{w}_{t+1} - \boldsymbol{w}^*, r'(\boldsymbol{w}_{t+1})\rangle - \lambda\eta_t B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}).\end{aligned}$$

Now we simply proceed as in the proof of Lemma 1 following Eq. (5). ■

The above corollary almost immediately gives a logarithmic regret bound.

Theorem 7 *Let r be λ-strongly convex with respect to a differentiable function ψ and suppose ψ is α-strongly convex with respect to a norm $\|\cdot\|$. Assume that $r(\boldsymbol{w}_1) = 0$. If $\|f_t'(\boldsymbol{w}_t)\|_* \le G_*$ for all t,*

$$R_\phi(T) \le \lambda B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1) + \frac{G_*^2}{\lambda\alpha}(\log T + 1) = O\left(\frac{G_*^2}{\lambda\alpha}\log T\right).$$

Proof: Rearranging Corollary 6, we have

$$\begin{aligned}&\sum_{t=1}^T f_t(\boldsymbol{w}_t) + r(\boldsymbol{w}_{t+1}) - f_t(\boldsymbol{w}^*) - r(\boldsymbol{w}^*)\\ &\quad\le\quad \sum_{t=1}^T\left[\frac{1}{\eta_t}B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) - \frac{1}{\eta_t}B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}) - \lambda B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1})\right] + \sum_{t=1}^T\frac{\eta_t}{2\alpha}\|f_t'(\boldsymbol{w}_t)\|_*^2.\\ &\quad=\quad \frac{1}{\eta_1}B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1) - \frac{1}{\eta_T}B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}) + \sum_{t=1}^{T-1}\left[B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1})\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) - \lambda B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1})\right]\\ &\qquad + \sum_{t=1}^T\frac{\eta_t}{2\alpha}\|f_t'(\boldsymbol{w}_t)\|_*^2\end{aligned}$$

If we set $\eta_t = \frac{1}{\lambda t}$, then the first summation above is zero and

$$\sum_{t=1}^{T} f_t(\boldsymbol{w}_t) + r(\boldsymbol{w}_{t+1}) - f_t(\boldsymbol{w}^*) - r(\boldsymbol{w}^*) \leq \lambda B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1) + \frac{1}{\lambda\alpha} \sum_{t=1}^{T} \frac{1}{t} \left\| f_t'(\boldsymbol{w}_t) \right\|_*^2 .$$

Noting that $\sum_{t=1}^{T} \frac{1}{t} \leq \log T + 1$ completes the proof of the theorem. ■

An interesting point regarding the above theorem is that we do not require $B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t)$ to be bounded or the set Ω to be compact, which previous work assumed. When the functions f_t are Lipschitz, then whenever r is strongly convex COMID still attains logarithmic regret.

Two notable examples attain the logarithmic bounds in the above theorem. It is clear that if r defines a valid Bregman divergence then that r is strongly convex with respect to itself in the sense of Eq. (7). First, consider optimization over the simplex with entropic regularization, that is, we set $r(\boldsymbol{w}) = \lambda \sum_i w_i \log w_i$ and $\Omega = \{\boldsymbol{w} : \boldsymbol{w} \succeq \boldsymbol{0}, \boldsymbol{1}^\top \boldsymbol{w} = 1\}$. In this case it is straightforward to see that $r(\boldsymbol{w}) = \sum_j w_i \log w_i$ is λ-strongly convex with respect to $\psi(\boldsymbol{w}) = r(\boldsymbol{w})$, which in turn is strongly convex with respect to the ℓ_1-norm $\|\cdot\|_1$ over Ω (see Shalev-Shwartz and Singer, 2007, Definition 2 and Example 2). Since the dual of the ℓ_1-norm is the ℓ_∞ norm, we have $R_\phi(T) = O\left(\frac{\log T}{\lambda} \max_t \left\| f_t'(\boldsymbol{w}_t) \right\|_\infty\right)$. We can also use $r(\boldsymbol{w}) = \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2$, in which case we recover the same bounds as those in Hazan et al. (2006).

6 Stochastic Convergence Results

In this section, we examine the application of COMID to solving stochastic optimization problems. The techniques we use have a long history in online algorithms and make connections between the regret of the algorithm and generalization performance using martingale concentration results (Littlestone, 1989). We build on known techniques for data-driven generalization bounds (Cesa-Bianchi et al., 2004) to give concentration results for COMID in the stochastic optimization setting. Further work on this subject for the strongly convex case can be found in Kakade and Tewari (2008), though we focus on the case when $f_t + r$ is weakly convex.

We let $f(\boldsymbol{w}) = \mathbb{E} f(\boldsymbol{w}; Z) = \int f(\boldsymbol{w}; z) dP(z)$, and at every step t the algorithm receives an independent random variable $Z_t \sim P$ that gives an unbiased estimate $f_t(\boldsymbol{w}_t) = f(\boldsymbol{w}_t; Z_t)$ of the function f evaluated at $\boldsymbol{w}_t$ and an unbiased estimate $f_t'(\boldsymbol{w}_t) = f'(\boldsymbol{w}_t; Z_t)$ of an arbitrary subgradient $f'(\boldsymbol{w}_t) \in \partial f(\boldsymbol{w}_t)$. We assume that $B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) \leq D^2$ for all t and for simplicity that $\|f_t'(\boldsymbol{w}_t)\|_* \leq G_*$ for all t, which are satisified when Ω is compact. We also assume without loss of generality that $r(\boldsymbol{w}_1) = 0$. For example, our original problem in which $f(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{w})$, where we randomly sample one f_i at each iteration, falls into this setup, resolving the question posed by Tseng (2009) on the existence of stochastic composite incremental subgradient methods.

Theorem 8 *Given the assumptions on f and Ω in the above paragraph, let $\{\boldsymbol{w}_t\}$ be the sequence generated by Eq. (3). In addition, let $\bar{\boldsymbol{w}}_T = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{w}_t$ and $\eta_t = \frac{D}{G_* \sqrt{\alpha t}}$. Then*

$$P\left(f(\bar{\boldsymbol{w}}_T) + r(\bar{\boldsymbol{w}}_T) \geq f(\boldsymbol{w}^*) + r(\boldsymbol{w}^*) + \frac{DG_*}{\sqrt{\alpha T}} + \varepsilon\right) \leq \exp\left(-\frac{T\alpha\varepsilon^2}{16 D^2 G_*^2}\right) .$$

Alternatively, with probability at least $1 - \delta$

$$f(\bar{\boldsymbol{w}}_T) + r(\bar{\boldsymbol{w}}_T) \leq f(\boldsymbol{w}^*) + r(\boldsymbol{w}^*) + \frac{DG_*}{\sqrt{\alpha T}} \left(1 + 4\sqrt{\log \frac{1}{\delta}}\right) .$$

Proof: We begin our derivation by recalling Lemma 1. Convexity of f and r imply

$$\begin{aligned}
&\eta_t \left[f(\boldsymbol{w}_t) + r(\boldsymbol{w}_{t+1}) - f(\boldsymbol{w}^*) - r(\boldsymbol{w}^*)\right] \\
&\leq \eta_t \left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f'(\boldsymbol{w}_t) \right\rangle + \eta_t \left\langle \boldsymbol{w}_{t+1} - \boldsymbol{w}^*, r'(\boldsymbol{w}_{t+1}) \right\rangle \\
&= \eta_t \left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f_t'(\boldsymbol{w}_t) \right\rangle + \eta_t \left\langle \boldsymbol{w}_{t+1} - \boldsymbol{w}^*, r'(\boldsymbol{w}_{t+1}) \right\rangle + \eta_t \left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f'(\boldsymbol{w}_t) - f_t'(\boldsymbol{w}_t) \right\rangle
\end{aligned}$$

We now follow the same derivation as Lemma 1, leaving $\eta_t \left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f'(\boldsymbol{w}_t) - f_t'(\boldsymbol{w}_t) \right\rangle$ intact, thus

$$\begin{aligned}
&\eta_t \left[f(\boldsymbol{w}_t) + r(\boldsymbol{w}_{t+1}) - f(\boldsymbol{w}^*) - r(\boldsymbol{w}^*)\right] \\
&\leq B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) - B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}) + \frac{\eta_t^2}{2\alpha} \left\| f_t'(\boldsymbol{w}_t) \right\|_*^2 + \eta_t \left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f'(\boldsymbol{w}_t) - f_t'(\boldsymbol{w}_t) \right\rangle . \qquad (8)
\end{aligned}$$

Now we subtract $r(\boldsymbol{w}_{T+1}) \geq 0$ from both sides, use the assumption that $r(\boldsymbol{w}_1) = 0$, and sum to get

$$\begin{aligned}
&\sum_{t=1}^{T} \left[f(\boldsymbol{w}_t) + r(\boldsymbol{w}_t) - f(\boldsymbol{w}^*) - r(\boldsymbol{w}^*) \right] \\
&\leq \sum_{t=1}^{T} \frac{1}{\eta_t} \left[B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) - B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_{t+1}) \right] + \frac{1}{2\alpha} \sum_{t=1}^{T} \eta_t \left\| f'_t(\boldsymbol{w}_t) \right\|_*^2 + \sum_{t=1}^{T} \left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f'(\boldsymbol{w}_t) - f'_t(\boldsymbol{w}_t) \right\rangle \\
&\leq \frac{1}{\eta_1} B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_1) + \sum_{t=2}^{T} B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) \left[\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right] + \frac{G_*^2}{2\alpha} \sum_{t=1}^{T} \eta_t + \sum_{t=1}^{T} \left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f'(\boldsymbol{w}_t) - f'_t(\boldsymbol{w}_t) \right\rangle \ . \qquad (9)
\end{aligned}$$

Let $\mathcal{F}_t$ be a filtration with $Z_\tau \in \mathcal{F}_t$ for $\tau \leq t$. Since $\boldsymbol{w}_t \in \mathcal{F}_{t-1}$,

$$\mathbb{E}\left[\left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f'(\boldsymbol{w}_t) - f'(\boldsymbol{w}_t; Z_t) \right\rangle \mid \mathcal{F}_{t-1} \right] = \left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f'(\boldsymbol{w}_t) - \mathbb{E}[f'(\boldsymbol{w}_t; Z_t) \mid \mathcal{F}_{t-1}] \right\rangle = 0 \ ,$$

and thus the last sum in Eq. (9) is a martingale difference sequence. We next use our assumptions that $B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) \leq D^2$ and $\frac{\alpha}{2} \left\| \boldsymbol{w}^* - \boldsymbol{w}_t \right\|^2 \leq B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t)$, therefore $\left\| \boldsymbol{w}^* - \boldsymbol{w}_t \right\| \leq \sqrt{2/\alpha} D$. Then

$$\left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f'(\boldsymbol{w}_t) - f'_t(\boldsymbol{w}_t) \right\rangle \leq \left\| \boldsymbol{w}_t - \boldsymbol{w}^* \right\| \left\| f'(\boldsymbol{w}_t) - f'_t(\boldsymbol{w}_t) \right\|_* \leq 2\sqrt{2/\alpha} D G_* \ .$$

Thus Eq. (9) consists of a bounded difference martingale, and we can use standard concentration techniques to get strong convergence guarantees. Applying Azuma's inequality,

$$P\left(\sum_{t=1}^{T} \left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f'(\boldsymbol{w}_t) - f'_t(\boldsymbol{w}_t) \right\rangle \geq \varepsilon \right) \leq \exp\left(-\frac{\alpha \varepsilon^2}{16 T D^2 G_*^2} \right) \ . \qquad (10)$$

Define $\gamma_T = \sum_{t=1}^{T} \left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, f'_t(\boldsymbol{w}_t) - f'(\boldsymbol{w}_t) \right\rangle$ and recall that $B_\psi(\boldsymbol{w}^*, \boldsymbol{w}_t) \leq D^2$. The convexity of f and r give $T[f(\bar{\boldsymbol{w}}_T) + r(\bar{\boldsymbol{w}}_T)] \leq \sum_{t=1}^{T} f(\boldsymbol{w}_t) + r(\boldsymbol{w}_t)$, so that

$$\begin{aligned}
T\left[f(\bar{\boldsymbol{w}}_T) + r(\bar{\boldsymbol{w}}_T) \right] &\leq T\left[f(\boldsymbol{w}^*) + r(\boldsymbol{w}^*) \right] + D^2 \left[\frac{1}{\eta_1} + \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \right] + \frac{G_*^2}{2\alpha} \sum_{t=1}^{T} \eta_t + \gamma_T \\
&= T\left[f(\boldsymbol{w}^*) + r(\boldsymbol{w}^*) \right] + \frac{D^2}{\eta_T} + \frac{G_*^2}{2\alpha} \sum_{t=1}^{T} \eta_t + \gamma_T .
\end{aligned}$$

Setting $\eta_t = \frac{D\sqrt{\alpha}}{G_* \sqrt{t}}$ we have $f(\bar{\boldsymbol{w}}_T) + r(\bar{\boldsymbol{w}}_T) \leq f(\boldsymbol{w}^*) + r(\boldsymbol{w}^*) + D G_*^2 \sqrt{T} / \sqrt{\alpha} + \frac{1}{T} \gamma_T$, and we can immediately apply Azuma's inequality from Eq. (10) to complete the theorem. ■

7 Special Cases and Derived Algorithms

In this section, we show specific instantiations of our framework for different regularization functions r, and we also show that some previously developed algorithms are special cases of the framework for optimization presented here. We also give results on learning matrices with Schatten p-norm divergences that generalize some recent interesting work on trace norm regularization.

7.1 Fobos

The recently proposed Fobos algorithm of Duchi and Singer (2009) is comprised, at each iteration, of the following two steps:

$$\tilde{\boldsymbol{w}}_{t+1} = \boldsymbol{w}_t - \eta f'_t(\boldsymbol{w}_t) \quad \text{and} \quad \boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \ \frac{1}{2} \left\| \boldsymbol{w} - \tilde{\boldsymbol{w}}_{t+1} \right\|_2 + \eta r(\boldsymbol{w}) \ .$$

It is straightforward to verify that the update

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \ \frac{1}{2} \left\| \boldsymbol{w} - \boldsymbol{w}_t \right\|_2^2 + \eta \left\langle f'_t(\boldsymbol{w}_t), \boldsymbol{w} - \boldsymbol{w}_t \right\rangle + \eta r(\boldsymbol{w})$$

is equivalent to the two step update above. Thus, Comid reduces to Fobos when we take $\psi(\boldsymbol{w}) = \frac{1}{2} \left\| \boldsymbol{w} \right\|_2^2$ and $\Omega = \mathbb{R}^d$ (with constant learning rate η). This also shows that we can run Fobos by restricting to a convex set $\Omega \neq \mathbb{R}^d$. Further, our results give tighter convergence guarantees than Fobos, in particular, they do not depend in any negative way on the regularization function r.

It is also not difficult to show in general that the two step process of setting

$$\widetilde{\boldsymbol{w}}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \; B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta \left\langle f_t'(\boldsymbol{w}_t), \boldsymbol{w} \right\rangle \quad \text{and} \quad \boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \; B_\psi(\boldsymbol{w}, \widetilde{\boldsymbol{w}}_t) + \eta r(\boldsymbol{w})$$

is equivalent to the original COMID update of Eq. (3) when $\Omega = \mathbb{R}^d$. Indeed, the optimal solution to the first step satisfies

$$\nabla\psi(\tilde{\boldsymbol{w}}_{t+1}) - \nabla\psi(\boldsymbol{w}_t) + \eta f_t'(\boldsymbol{w}_t) = 0 \quad \text{so that} \quad \tilde{\boldsymbol{w}}_{t+1} = \nabla\psi^{-1}(\nabla\psi(\boldsymbol{w}_t) - \eta f_t'(\boldsymbol{w}_t)).$$

Then looking at the optimal solution for the second step, for some $r'(\boldsymbol{w}_{t+1}) \in \partial r(\boldsymbol{w}_{t+1})$ we have

$$\nabla\psi(\boldsymbol{w}_{t+1}) - \nabla\psi(\tilde{\boldsymbol{w}}_{t+1}) + \eta r'(\boldsymbol{w}_{t+1}) = 0 \quad \text{i.e.} \quad \nabla\psi(\boldsymbol{w}_{t+1}) - \nabla\psi(\boldsymbol{w}_t) + \eta f_t'(\boldsymbol{w}_t) + \eta r'(\boldsymbol{w}_{t+1}) = 0.$$

This is clearly the solution to the one-step update of Eq. (3).

7.2 p-norm divergences

Now we consider divergence functions ψ which are the ℓ_p-norms squared. $\frac{1}{2}\|\boldsymbol{w}\|_p^2$ is $(p-1)$-strongly convex over $\mathbb{R}^d$ with respect to the ℓ_p-norm for any $p \in (1, 2]$ (Ball et al., 1994). We see that if we choose $\psi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_p^2$ to be the divergence function, we have a corollary to Theorem 2.

Corollary 9 *Suppose that $r(\mathbf{0}) = 0$ and that $\boldsymbol{w}_1 = \mathbf{0}$. Let $p = 1 + 1/\log d$ and use the Bregman function $\psi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_p^2$. Further suppose that either Ω is compact or the f_t are Lipschitz so that for $q = \log d + 1$, $\max_t \|f_t'(\boldsymbol{w}_t)\|_q \le G_q$. Setting $\eta = \frac{\|\boldsymbol{w}^*\|_p}{G_q}\sqrt{\frac{1}{T\log d}}$, the regret of* COMID *satisfies*

$$R_\phi(T) \le \|\boldsymbol{w}^*\|_p G_q \sqrt{T\log d} \asymp \|\boldsymbol{w}^*\|_1 G_\infty \sqrt{T\log d}\ .$$

Proof: Recall that the dual norm for an ℓ_p-norm is an ℓ_q-norm, where $q = p/(p-1)$. From Thm. 2, we immediately have that when $\boldsymbol{w}_1 = \mathbf{0}$ and $\psi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_p^2$

$$R(T) \le \frac{1}{2\eta}\|\boldsymbol{w}^*\|_p^2 + \frac{\eta}{2(p-1)}\sum_{t=1}^T \|f_t'(\boldsymbol{w}_t)\|_q^2 .$$

Now use the assumption that $\max_t \|f_t'(\boldsymbol{w}_t)\|_q \le G_q$, replace p with $1 + 1/\log d$ (so $q = \log d + 1$), and set $\eta = c\sqrt{\frac{1}{T\log d}}$, which results in

$$R(T) \le \frac{\sqrt{T\log d}}{2c}\|\boldsymbol{w}^*\|_p^2 + c\frac{\sqrt{T\log d}}{2}G_q^2.$$

Setting $c = \|\boldsymbol{w}^*\|_p / G_q$ gives us our desired result. ■

From the above, we see that COMID is a good candidate for (dense) problems in high dimensions, especially when we use ℓ_1-regularization. For high dimensions when $\boldsymbol{w} \in \mathbb{R}^d$, taking $p = 1+1/\log d \approx 1$ means our bounds depend roughly on the ℓ_1-norm of the optimal predictor and the infinity norm of the function gradients f_t. Shalev-Shwartz and Tewari (2009) recently proposed the "Stochastic Mirror Descent made Sparse" algorithm (SMIDAS) using this intuition. We recover SMIDAS by taking the divergence in COMID to be $\psi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_p^2$ and $r(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|_1$. The COMID update is

$$\nabla\psi(\tilde{\boldsymbol{w}}_{t+1}) = \nabla\psi(\boldsymbol{w}_t) - \eta f_t'(\boldsymbol{w}_t), \quad \boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \; B_\psi(\boldsymbol{w}, \tilde{\boldsymbol{w}}_{t+1}) + \eta\lambda\|\boldsymbol{w}\|_1 \ .$$

The SMIDAS update, on the other hand, is

$$\nabla\psi(\tilde{\boldsymbol{w}}_{t+1}) = \nabla\psi(\boldsymbol{w}) - \eta f_t'(\boldsymbol{w}_t), \quad \nabla\psi(\boldsymbol{w}_{t+1}) = \mathcal{S}_{\eta\lambda}\left(\nabla\psi(\tilde{\boldsymbol{w}}_{t+1})\right) \ ,$$

where $\mathcal{S}_\tau$ is the shrinkage/thresholding operator defined by

$$[\mathcal{S}_\tau(\boldsymbol{x})]_j = \operatorname{sign}(x_j)\left[|x_j| - \tau\right]_+ \ . \tag{11}$$

The following lemma proves that the two updates are identical in cases including p-norm divergences.

Lemma 10 *Suppose ψ is strongly convex and its gradient satisfies*

$$\operatorname{sign}([\nabla\psi(\boldsymbol{w})]_j) = \operatorname{sign}(w_j) \ . \tag{12}$$

Then the unique solution $\boldsymbol{v}$ of $\boldsymbol{v} = \operatorname{argmin}_{\boldsymbol{w}}\{B_\psi(\boldsymbol{w}, \boldsymbol{u}) + \tau\|\boldsymbol{w}\|_1\}$ is given by

$$\nabla\psi(\boldsymbol{v}) = \mathcal{S}_\tau(\nabla\psi(\boldsymbol{u})) \ . \tag{13}$$

Proof: Since ψ is strongly convex, the solution is unique. We will show that if $\boldsymbol{v}$ satisfies Eq. (13) then it is a solution to the problem. Therefore, suppose Eq. (13) holds. The proof proceeds by considering three cases.

Case I: $[\nabla\psi(\boldsymbol{u})]_j > \tau$. In this case, $[\nabla\psi(\boldsymbol{v})]_j = [\nabla\psi(\boldsymbol{u})]_j - \tau > 0$ and by Eq. (12), $v_j > 0$. Thus

$$[\nabla\psi(\boldsymbol{v})]_j - [\nabla\psi(\boldsymbol{u})]_j + \tau\,\mathrm{sign}(v_j) = 0\ .$$

Case II: $[\nabla\psi(\boldsymbol{u})]_j < -\tau$. In this case, $[\nabla\psi(\boldsymbol{v})]_j = [\nabla\psi(\boldsymbol{u})]_j + \tau < 0$ and Eq. (12) implies $v_j < 0$. So

$$[\nabla\psi(\boldsymbol{v})]_j - [\nabla\psi(\boldsymbol{u})]_j + \tau\,\mathrm{sign}(v_j) = 0\ .$$

Case III: $[\nabla\psi(\boldsymbol{u})]_j \in [-\tau, \tau]$. Here, we can take $v_j = 0$ and Eq. (12) will give $[\nabla\psi(\boldsymbol{v})]_j = 0$. Thus

$$0 \in [\nabla\psi(\boldsymbol{v})]_j - [\nabla\psi(\boldsymbol{u})]_j + \tau[-1, 1]\ .$$

Combining the three cases, $\boldsymbol{v}$ satisfies $\mathbf{0} \in \nabla\psi(\boldsymbol{v}) - \nabla\psi(\boldsymbol{u}) + \tau\partial\left\|\boldsymbol{v}\right\|_1$, which is the optimality condition for $\boldsymbol{v} \in \mathrm{argmin}_{\boldsymbol{w}}\{B_\psi(\boldsymbol{w}, \boldsymbol{u}) + \tau\left\|\boldsymbol{w}\right\|_1\}$. We thus have $\nabla\psi(\boldsymbol{v}) = \mathcal{S}_\tau(\nabla\psi(\boldsymbol{u}))$ as desired. ■

Rewriting the above lemma slightly gives the following result. The solution to

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}} \left\{B_\psi(\boldsymbol{w}, \boldsymbol{w}_t) + \eta\left\langle f_t'(\boldsymbol{w}_t), \boldsymbol{w} - \boldsymbol{w}_t\right\rangle + \eta r(\boldsymbol{w})\right\}$$

when ψ satisfies the gradient condition of Eq. (12) is

$$\begin{aligned}\boldsymbol{w}_{t+1} &= (\nabla\psi)^{-1}\left[\mathrm{sign}\left(\nabla\psi(\boldsymbol{w}_t) - \eta f_t'(\boldsymbol{w}_t)\right) \odot \max\left\{\left|\nabla\psi(\boldsymbol{w}_t) - \eta f_t'(\boldsymbol{w}_t)\right| - \eta\lambda, 0\right\}\right] \\ &= (\nabla\psi)^{-1}\left[\mathcal{S}_{\eta\lambda}(\nabla\psi(\boldsymbol{w}_t) - \eta f_t'(\boldsymbol{w}_t))\right]. \qquad (14)\end{aligned}$$

Note that when $\psi(\cdot) = \left\|\cdot\right\|_p^2$ we recover Shalev-Shwartz and Tewari's SMIDAS, while with $p = 2$ we get Langford et al.'s 2009 truncated gradient method. See Shalev-Shwartz and Tewari (2009) or Gentile and Littlestone (1999) for the simple formulae to compute $(\nabla\psi)^{-1} \equiv \nabla\psi^*$.

ℓ_∞-regularization Let us now consider the problem of setting $r(\boldsymbol{w})$ to be a general ℓ_p-norm (we will specialize this to ℓ_∞ shortly). We describe the dual function and then use it to derive a few particular updates, mentioning an open problem. First, let $p_1 > 1$ be the norm associated with the Bregman function ψ and p_2 be the norm for $r(\boldsymbol{w}) = \left\|\boldsymbol{w}\right\|_{p_2}$. Let $q_i = p_i/(p_i - 1)$ be the associated dual norm. Then, ignoring constants, the minimization problem from Eq. (3) becomes

$$\min_{\boldsymbol{w}}\ \left\langle \boldsymbol{v}, \boldsymbol{w}\right\rangle + \frac{1}{2}\left\|\boldsymbol{w}\right\|_{p_1}^2 + \lambda\left\|\boldsymbol{w}\right\|_{p_2}.$$

We introduce a variable $\boldsymbol{z} = \boldsymbol{w}$ and get the equivalent problem $\min_{\boldsymbol{w}=\boldsymbol{z}} \left\langle \boldsymbol{v}, \boldsymbol{w}\right\rangle + \frac{1}{2}\left\|\boldsymbol{w}\right\|_{p_1}^2 + \lambda\left\|\boldsymbol{z}\right\|_{p_2}$. To derive the dual of the problem, we introduce Lagrange multiplier $\boldsymbol{\theta}$ and find the Lagrangian

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}) = \left\langle \boldsymbol{v} - \boldsymbol{\theta}, \boldsymbol{w}\right\rangle + \frac{1}{2}\left\|\boldsymbol{w}\right\|_{p_1}^2 + \lambda\left\|\boldsymbol{z}\right\|_{p_2} + \left\langle \boldsymbol{\theta}, \boldsymbol{z}\right\rangle.$$

Taking the infimum over $\boldsymbol{w}$ and $\boldsymbol{z}$ in the Lagrangian, since the conjugate of $\frac{1}{2}\left\|\cdot\right\|_p^2$ is $\frac{1}{2}\left\|\cdot\right\|_q^2$ when $1/p + 1/q = 1$ (Boyd and Vandenberghe, 2004, Example 3.27) we have

$$\inf_{\boldsymbol{w}}\left[\left\langle \boldsymbol{v} - \boldsymbol{\theta}, \boldsymbol{w}\right\rangle + \tfrac{1}{2}\left\|\boldsymbol{w}\right\|_{p_1}^2\right] = -\frac{1}{2}\left\|\boldsymbol{v} - \boldsymbol{\theta}\right\|_{q_1}^2 \qquad \inf_{\boldsymbol{z}}\left[\lambda\left\|\boldsymbol{z}\right\|_{p_2} + \left\langle \boldsymbol{\theta}, \boldsymbol{z}\right\rangle\right] = \begin{cases} 0 & \text{if } \left\|\boldsymbol{\theta}\right\|_{q_2} \le \lambda \\ -\infty & \text{otherwise.}\end{cases}$$

Thus, our dual is the non-Euclidean projection problem

$$\min_{\boldsymbol{\theta}}\ \frac{1}{2}\left\|\boldsymbol{v} - \boldsymbol{\theta}\right\|_{q_1} \quad \text{s.t. } \left\|\boldsymbol{\theta}\right\|_{q_2} \le \lambda.$$

The Lagrangian earlier is differentiable with respect to $\boldsymbol{w}$, so we can recover the optimal $\boldsymbol{w}$ from $\boldsymbol{\theta}$ by noting that when $\psi(\boldsymbol{w}) = \frac{1}{2}\left\|\boldsymbol{w}\right\|_{p_1}^2$, $\nabla\psi(\boldsymbol{w}) + \boldsymbol{v} - \boldsymbol{\theta} = 0$ at optimum or $\boldsymbol{w} = (\nabla\psi)^{-1}(\boldsymbol{\theta} - \boldsymbol{v})$. When $p_2 = 1$, we easily recover Eq. (14) as our update. However, the case $p_2 = \infty$ is more interesting, as it can be a building block for group-sparsity (Obozinski et al., 2007). In this case our problem is

$$\min_{\boldsymbol{\theta}}\ \frac{1}{2}\left\|\boldsymbol{v} - \boldsymbol{\theta}\right\|_q \quad \text{s.t. } \left\|\boldsymbol{\theta}\right\|_1 \le \lambda.$$

It is clear by symmetry in the above that we can assume $\boldsymbol{v} \succeq \mathbf{0}$ with no loss of generality. We can raise the ℓ_q-norm to a power greater than 1 and maintain convexity, so our equivalent problem is

$$\min_{\theta}\ \frac{1}{q}\left\|\boldsymbol{v} - \boldsymbol{\theta}\right\|_q^q \quad \text{s.t. } \left\langle \mathbf{1}, \boldsymbol{\theta}\right\rangle \le \lambda, \boldsymbol{\theta} \succeq \mathbf{0}. \qquad (15)$$

Let $\widehat{\boldsymbol{\theta}}$ be the solution of Eq. (15). Clearly, at optimum we will have $\widehat{\theta}_i \leq v_i$, though we use this only for clarity in the derivation and omit constraints as they do not affect the optimization problem. Introducing Lagrange multipliers ν and $\boldsymbol{\alpha} \succeq \mathbf{0}$ we get the Lagrangian

$$\mathcal{L}(\boldsymbol{\theta}, \nu, \boldsymbol{\alpha}) = \frac{1}{q}\sum_{i=1}^{d}(v_i - \theta_i)^q + \nu(\langle \mathbf{1}, \boldsymbol{\theta}\rangle - \lambda) - \langle \boldsymbol{\alpha}, \boldsymbol{\theta}\rangle$$

Taking the derivative of the above, we have

$$-(v_i - \theta_i)^{q-1} + \nu - \alpha_i = 0 \quad \Rightarrow \quad \theta_i = v_i - (\nu - \alpha_i)^{1/(q-1)}$$

Now suppose we knew the optimal ν. If an index i satisfies $\nu \geq v_i^{q-1}$, then we will have $\widehat{\theta}_i = 0$. To see this, suppose for the sake of contradiction that $\widehat{\theta}_i > 0$. The KKT conditions for optimality of Eq. (15) (Boyd and Vandenberghe, 2004) imply that for such i we have

$$\widehat{\theta}_i = v_i - (\nu - \alpha_i)^{1/(q-1)} = v_i - \nu^{1/(q-1)} \leq 0,$$

a contradiction. Similarly, if $\nu < v_i^{q-1}$ and $\alpha_i \geq 0$, then $\widehat{\theta}_i > 0$. Ineed, since $\alpha_i \geq 0$,

$$\widehat{\theta}_i = v_i - (\nu - \alpha_i)^{1/(q-1)} > v_i - (v_i^{q-1} - \alpha_i)^{1/(q-1)} \geq v_i - v_i = 0,$$

so that $\widehat{\theta}_i > 0$ and the KKT conditions imply $\alpha_i = 0$. Had we known ν, the optimal $\widehat{\theta}_i$ would have been easy to attain as $\widehat{\theta}_i(\nu) = v_i - (\min\{\nu, v_i^{q-1}\})^{1/(q-1)}$ (note that this satisfies $0 \leq \widehat{\theta}_i \leq v_i$). Since we know that the structure of the optimal $\widehat{\theta}$ must obey the above equation, we can boil our problem down to finding $\nu \geq 0$ so that

$$\sum_{i=1}^{d}\widehat{\theta}_i(\nu) = \sum_{i=1}^{d} v_i - \min\{\nu, v_i^{q-1}\}^{1/(q-1)} = \lambda. \tag{16}$$

Interestingly, this reduces to *exactly* the same root-finding problem as that for solving Euclidean projection to an ℓ_1-ball (Duchi et al., 2008). As shown by Duchi et al., it is straightforward to find the optimal ν in time linear in the dimension d.

An open problem is to find an efficient algorithm for solving the generalized projections above when using the 2 rather than ∞ norm.

Mixed-norm regularization Now we consider the problem of mixed-norm regularization, in which we wish to minimize functions $f_t(\boldsymbol{W}) + r(\boldsymbol{W})$ of a matrix $\boldsymbol{W} \in \mathbb{R}^{d\times k}$. In particular, we define $\overline{\boldsymbol{w}}_i \in \mathbb{R}^k$ to be the i^{th} row of $\boldsymbol{W}$, and we set $r(\boldsymbol{W}) = \lambda\sum_{i=1}^{d}\left\|\overline{\boldsymbol{w}}_i\right\|_{p_2}$. We also use the p-norm Bregman functions as above with $\psi(\boldsymbol{W}) = \frac{1}{2}\left\|\boldsymbol{W}\right\|_p^2 = \frac{1}{2}\left(\sum_{i,j}\boldsymbol{W}_{ij}^p\right)^{1/p}$, which are $(p-1)$-strongly convex with respect to the ℓ_p-norm squared. As earlier, our minimization problem becomes

$$\min_{\boldsymbol{W}} \ \langle \boldsymbol{V}, \boldsymbol{W}\rangle + \frac{1}{2}\left\|\boldsymbol{W}\right\|_{p_1}^2 + \lambda\left\|\boldsymbol{W}\right\|_{\ell_1/\ell_{p_2}},$$

whose dual problem is

$$\min_{\boldsymbol{\Theta}} \ \left\|\boldsymbol{V} - \boldsymbol{\Theta}\right\|_{q_1} \quad \text{s.t.} \ \left\|\overline{\boldsymbol{\theta}}_i\right\|_{q_2} \leq \lambda$$

Raising the first norm to the q_1-power, we see that the problem is separable, and we can solve it using the techniques in the prequel.

7.3 Matrix Composite Mirror Descent

We now consider a setting that generalizes the previous discussions in which our variables $\boldsymbol{W}_t$ are matrices $\boldsymbol{W}_t \in \Omega = \mathbb{R}^{d_1\times d_2}$. We use Bregman functions based on Schatten p-norms (e.g. Horn and Johnson, 1985, Section 7.4). Schatten p-norms are the family of unitarily invariant matrix norms arising out of applying p-norms to the singular values of the matrix $\boldsymbol{W}$. That is, letting $\sigma(\boldsymbol{W})$ denote the vector of singular values of $\boldsymbol{W}$, we set $\left\|\boldsymbol{W}\right\|_p = \left\|\sigma(\boldsymbol{W})\right\|_p$. We use Bregman functions $\psi(\boldsymbol{W}) = \frac{1}{2}\left\|\boldsymbol{W}\right\|_p^2$, which, similar to the p-norm on vectors, are $(p-1)$-strongly convex over $\Omega = \mathbb{R}^{d_1\times d_2}$ with respect to the norm $\left\|\cdot\right\|_p$ (Ball et al., 1994).

As in the previous subsection, we mainly consider two values for p, $p = 2$ and a value very near 1, namely $p = 1 + 1/\log d$. For $p = 2$, ψ is 1-strongly convex with respect to $\left\|\cdot\right\|_2 = \left\|\cdot\right\|_{\text{Fr}}$, the

Frobenius norm. For the second value, ψ is $1/\log d$-strongly convex with respect to $\|\cdot\|_{1+1/\log d}$, or, with a bit more work, ψ is $1/(3\log d)$-strongly convex w.r.t. $\|\cdot\|_1$, the trace or nuclear norm.

We focus on the specific setting of trace-norm regularization, or $r(\boldsymbol{W}) = \lambda \|\boldsymbol{W}\|_1$. This norm, similar to the ℓ_1-norm on vectors, gives sparsity in the singular value spectrum of $\boldsymbol{W}$ and hence is useful for rank-minimization (Recht et al., 2007). The generic COMID update with the above choice of ψ gives a "Schatten p-norm" COMID algorithm for matrix applications:

$$\boldsymbol{W}_{t+1} = \operatorname*{argmin}_{\boldsymbol{W}\in\Omega} \ \eta \langle f_t'(\boldsymbol{W}_t), \boldsymbol{W}\rangle + B_\psi(\boldsymbol{W}, \boldsymbol{W}_t) + \eta\lambda \|\boldsymbol{W}\|_1 \ . \tag{17}$$

The update is well defined since ψ is strongly convex as per the above discussion. We also have defined $\langle \boldsymbol{W}, \boldsymbol{V}\rangle = \operatorname{tr}(\boldsymbol{W}^\top \boldsymbol{V})$ as the usual matrix inner product. The generic COMID convergence result in Thm. 2 immediately yields the following two corollaries.

Corollary 11 *Let the sequence $\{\boldsymbol{W}_t\}$ be defined by the update in Eq. (17) with $p = 2$. If each f_t satisfies $\|f_t'(\boldsymbol{W}_t)\|_2 \le G_2$, then there is a stepsize η for which the regret against $\boldsymbol{W}^* \in \Omega$ is*

$$R_\phi(T) \le G_2 \|\boldsymbol{W}^*\|_2 \sqrt{T}$$

Corollary 12 *Let $p = 1 + 1/\log d$ in the Schatten* COMID *update of Eq. (17). Let $q = 1 + \log d$. If each f_t satisfies $\|f_t'(\boldsymbol{W}_t)\|_q \le G_q$ then*

$$R_\phi(T) \le G_q \|\boldsymbol{W}^*\|_p \sqrt{T\log d} \asymp G_\infty \|\boldsymbol{W}^*\|_1 \sqrt{T\log d}$$

where $G_\infty = \max_t \|f_t'(\boldsymbol{W}_t)\|_\infty = \max_t \sigma_{\max}(f_t'(\boldsymbol{W}_t))$.

Let us consider the actual implementation of the COMID update. Similar convergence rates (with worse constants and a negative dependence on the spectrum of r) to those above can be achieved via simple mirror descent, i.e. by linearizing $r(\boldsymbol{W})$. The advantage of COMID is that it achieves sparsity in the spectrum, as the following proposition demonstrates.

Proposition 13 *Let $\Psi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_p^2$ and $\mathcal{S}_\tau(v) = \operatorname{sign}(v)[|v| - \tau]_+$ as in the prequel. For $p \in (1, 2]$, the update in Eq. (17) can be implemented as follows.*

$$\text{Compute SVD: } \boldsymbol{W}_t = \boldsymbol{U}_t \operatorname{diag}(\sigma(\boldsymbol{W}_t))\boldsymbol{V}_t^\top \tag{18a}$$

$$\text{Gradient step: } \boldsymbol{\Theta}_t = \boldsymbol{U}_t \operatorname{diag}(\nabla\Psi(\sigma(\boldsymbol{W}_t)))\boldsymbol{V}_t^\top - \eta f_t'(\boldsymbol{W}_t)$$

$$\text{Compute SVD: } \boldsymbol{\Theta}_t = \widetilde{\boldsymbol{U}}_t \operatorname{diag}(\sigma(\boldsymbol{\Theta}_t))\widetilde{\boldsymbol{V}}_t^\top$$

$$\text{Splitting update: } \boldsymbol{W}_{t+1} = \widetilde{\boldsymbol{U}}_t \operatorname{diag}\left((\nabla\Psi)^{-1}(\mathcal{S}_{\eta\lambda}(\sigma(\boldsymbol{\Theta}_t)))\right)\widetilde{\boldsymbol{V}}_t^\top \tag{18b}$$

Note that the first SVD, Eq. (18a), is used for notational convenience only and need not be computed at each iteration, since it is maintained at the end of iteration $t-1$ via Eq. (18b). The computational requirements for COMID are thus the same as standard mirror descent, which also requires an SVD computation on each step to compute $\partial\|\boldsymbol{W}\|_1$. The last step of the update, Eq. (18b) applies the shrinkage/thresholding operator $\mathcal{S}_{\eta\lambda}$ to the *spectrum* of $\boldsymbol{\Theta}_t$, which introduces sparsity. Furthermore, due to the sign (and hence sparsity) preserving nature of the map $(\nabla\Psi)^{-1}$, the sparsity in the spectrum is maintained in $\boldsymbol{W}_{t+1}$. Lastly, the special case for $p = 2$, the standard Frobenius norm update, was derived (but without rates or allowing stochastic gradients) by Ma et al. (2009), who report good empirical results for their algorithm. In trace-norm applications, we expect $\|\boldsymbol{W}^*\|_1$ to be small. Therefore, in such applications, our new Schatten-p COMID algorithm with $p \approx 1$ should give strong performance since G_∞ can be much smaller than G_2.

Proof of Proposition 13: We know from the prequel that the COMID step is equivalent to

$$\nabla\psi(\widetilde{\boldsymbol{W}}_t) = \nabla\psi(\boldsymbol{W}_t) - \eta f_t'(\boldsymbol{W}_t) \quad \text{and} \quad \boldsymbol{W}_{t+1} = \operatorname*{argmin}_{\boldsymbol{W}} \left\{ B_\psi(\boldsymbol{W}, \widetilde{\boldsymbol{W}}_t) + \eta r(\widetilde{\boldsymbol{W}}_t)\right\} \ .$$

Since $\boldsymbol{W}_t$ has singular value decomposition $\boldsymbol{U}_t \operatorname{diag}(\sigma(\boldsymbol{W}))\boldsymbol{V}_t$ and $\psi(\boldsymbol{W}) = \Psi(\sigma(\boldsymbol{W}))$ is unitarily invariant, $\nabla\psi(\boldsymbol{W}_t) = \boldsymbol{U}_t \operatorname{diag}(\nabla\Psi(\sigma(\boldsymbol{W}_t)))\boldsymbol{V}_t$ (Lewis, 1995, Corollary 2.5). This means that the $\boldsymbol{\Theta}_t$ computed in step 2 above is simply $\nabla\psi(\widetilde{\boldsymbol{W}}_t)$. The proof essentially amounts to a reduction to the vector case, since the norms are unitarily invariant, and will be complete if we prove that

$$\boldsymbol{V} = \operatorname*{argmin}_{\boldsymbol{W}} \left\{ B_\psi(\boldsymbol{W}, \widetilde{\boldsymbol{W}}_t) + \tau\|\boldsymbol{W}\|_1\right\}$$

has the unique solution

$$\boldsymbol{V} = \widetilde{\boldsymbol{U}}_t \operatorname{diag}\Big(\underbrace{(\nabla\Psi)^{-1}(\mathcal{S}_\tau(\sigma(\nabla\psi(\widetilde{\boldsymbol{W}}_t))))}_{\widetilde{\boldsymbol{w}}} \Big) \widetilde{\boldsymbol{V}}_t \ , \tag{19}$$

where $\widetilde{\boldsymbol{U}}_t \operatorname{diag}(\sigma(\widetilde{\boldsymbol{W}}_t))\widetilde{\boldsymbol{V}}_t^\top$ is the SVD of $\widetilde{\boldsymbol{W}}_t$. By subgradient optimality conditions, it is sufficient that the proposed solution $\boldsymbol{V}$ satisfy

$$\mathbf{0}_{d_1 \times d_2} \in \nabla\psi(\boldsymbol{V}) - \nabla\psi(\widetilde{\boldsymbol{W}}_t) + \tau\partial \left\|\boldsymbol{V}\right\|_1 \ .$$

Applying Lewis's Corollary 2.5, we can continue to use the orthonormal matrices $\widetilde{\boldsymbol{U}}_t$ and $\widetilde{\boldsymbol{V}}_t$, and we see that the proposed $\boldsymbol{V}$ in Eq. (19) satisfies

$$\nabla\psi(\boldsymbol{V}) = \widetilde{\boldsymbol{U}}_t \operatorname{diag}(\nabla\Psi(\widetilde{\boldsymbol{w}}))\widetilde{\boldsymbol{V}}_t^\top \quad \text{and} \quad \partial \left\|\boldsymbol{V}\right\|_1 = \widetilde{\boldsymbol{U}}_t \operatorname{diag}(\partial \left\|\widetilde{\boldsymbol{w}}\right\|_1)\widetilde{\boldsymbol{V}}_t^\top .$$

We have thus reduced the problem to showing that $\mathbf{0}_d \in \nabla\Psi(\widetilde{\boldsymbol{w}}) + \nabla\Psi(\sigma(\widetilde{\boldsymbol{W}}_t)) + \tau\partial \left\|\widetilde{\boldsymbol{w}}\right\|_1$, since we chose the matrices to have the same singular vectors by construction. From Lemma 10 presented earlier, we already know that $\widetilde{\boldsymbol{w}}$ satisfies this equation if and only if $\nabla\Psi(\widetilde{\boldsymbol{w}}) = \mathcal{S}_\tau(\nabla\Psi(\sigma(\widetilde{\boldsymbol{W}}_t)))$, which is indeed the case by definition of $\widetilde{\boldsymbol{w}}$ (noting of course that $\sigma(\nabla\psi(\widetilde{\boldsymbol{W}}_t)) = \nabla\Psi(\sigma(\widetilde{\boldsymbol{W}}_t))$). ■

Acknowledgments

JCD was supported by a National Defense Science and Engineering Graduate (NDSEG) fellowship, and some of this work was completed while JCD was an intern at Google Research.

References

K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.

K. Ball, E. Carlen, and E. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones mathematicae*, 115:463–482, 1994.

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

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems 20*, pages 161–168, 2008.

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.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communication on Pure and Applied Mathematics*, 57(11): 1413–1457, 2004.

J. Duchi and Y. Singer. Online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 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.

C. Gentile and N. Littlestone. The robustness of the p-norm algorithms. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, 1999.

E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. In *Proceedings of the 19th Annual Conference on Computational Learning Theory*, 2006.

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

S. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems 22*. MIT Press, 2008.

J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:747–776, 2009.

A. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2:173–183, 1995.

P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16:964–979, 1979.

N. Littlestone. From on-line to batch learning. In *Proceedings of the Second Annual Workshop on Computational Learning Theory*, pages 269–284, July 1989.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

S. Ma, D. Goldfarb, and L. Chen. Fixed point and Bregman iterative methods for matrix rank minimization. *Submitted*, 2009. URL `http://arxiv.org/abs/0905.1643`.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, 1983.

Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 76, Center for Operations Research and Econometrics, Catholic University of Louvain (UCL), 2007.

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.

B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *Submitted*, 2007.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

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`.

S. Shalev-Shwartz and N. Srebro. SVM optimization: Inverse dependence on training set size. In *Proceedings of the 25th International Conference on Machine Learning*, pages 928–935, 2008.

S. Shalev-Shwartz and A. Tewari. Stochastic methods for ℓ_1-regularized loss minimization. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, 2004.

P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Submitted*, 2009.

S. Wright, R. Nowak, and M. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

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 20th International Conference on Machine Learning*, 2003.