Journal of Machine Learning Research 11 (2010) 2543-2596 Submitted 5/09; Revised 3/10; Published 10/10

Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization

Lin Xiao LIN.XIAO@MICROSOFT.COM
Microsoft Research
1 Microsoft Way
Redmond, WA 98052, USA

Editor: Sham Kakade

Abstract

We consider regularized stochastic learning and online optimization problems, where the objective function is the sum of two convex terms: one is the loss function of the learning task, and the other is a simple regularization term such as ℓ_1-norm for promoting sparsity. We develop extensions of Nesterov's dual averaging method, that can exploit the regularization structure in an online setting. At each iteration of these methods, the learning variables are adjusted by solving a simple minimization problem that involves the running average of all past subgradients of the loss function and the whole regularization term, not just its subgradient. In the case of ℓ_1-regularization, our method is particularly effective in obtaining sparse solutions. We show that these methods achieve the optimal convergence rates or regret bounds that are standard in the literature on stochastic and online convex optimization. For stochastic learning problems in which the loss functions have Lipschitz continuous gradients, we also present an accelerated version of the dual averaging method.

Keywords: stochastic learning, online optimization, ℓ_1-regularization, structural convex optimization, dual averaging methods, accelerated gradient methods

1. Introduction

In machine learning, online algorithms operate by repetitively drawing random examples, one at a time, and adjusting the learning variables using simple calculations that are usually based on the single example only. The low computational complexity (per iteration) of online algorithms is often associated with their slow convergence and low accuracy in solving the underlying optimization problems. As argued by Bottou and Bousquet (2008), the combined low complexity and low accuracy, together with other tradeoffs in statistical learning theory, still make online algorithms favorite choices for solving large-scale learning problems. Nevertheless, traditional online algorithms, such as stochastic gradient descent, have limited capability of exploiting problem structure in solving *regularized* learning problems. As a result, their low accuracy often makes it hard to obtain the desired regularization effects, for example, sparsity under ℓ_1-regularization.

In this paper, we develop a new class of online algorithms, the *regularized dual averaging* (RDA) methods, that can exploit the regularization structure more effectively in an online setting. In this section, we describe the two types of problems that we consider, and explain the motivation of our work.

1.1 Regularized Stochastic Learning

The regularized stochastic learning problems we consider are of the following form:

$$\underset{w}{\text{minimize}} \quad \Big\{ \phi(w) \triangleq \mathbf{E}_z f(w,z) + \Psi(w) \Big\} \tag{1}$$

where $w \in \mathbf{R}^n$ is the optimization variable (often called *weights* in learning problems), $z = (x,y)$ is an input-output pair of data drawn from an (unknown) underlying distribution, $f(w,z)$ is the loss function of using w and x to predict y, and $\Psi(w)$ is a regularization term. We assume $\Psi(w)$ is a closed convex function (Rockafellar, 1970, Section 7), and its effective domain, $\operatorname{dom}\Psi = \{w \in \mathbf{R}^n \,|\, \Psi(w) < +\infty\}$, is closed. We also assume that $f(w,z)$ is convex in w for each z, and it is subdifferentiable (a subgradient always exists) on $\operatorname{dom}\Psi$. Examples of the loss function $f(w,z)$ include:

- Least-squares: $x \in \mathbf{R}^n$, $y \in \mathbf{R}$, and $f(w,(x,y)) = (y - w^T x)^2$.
- Hinge loss: $x \in \mathbf{R}^n$, $y \in \{+1,-1\}$, and $f(w,(x,y)) = \max\{0, 1 - y(w^T x)\}$.
- Logistic regression: $x \in \mathbf{R}^n$, $y \in \{+1,-1\}$, and $f(w,(x,y)) = \log\big(1 + \exp\big(-y(w^T x)\big)\big)$.

Examples of the regularization term $\Psi(w)$ include:

- ℓ_1-regularization: $\Psi(w) = \lambda \|w\|_1$ with $\lambda > 0$. With ℓ_1-regularization, we hope to get a relatively sparse solution, that is, with many entries of the weight vector w being zeroes.
- ℓ_2-regularization: $\Psi(w) = (\sigma/2)\|w\|_2^2$, with $\sigma > 0$. When ℓ_2-regularization is used with the hinge loss function, we have the standard setup of support vector machines.
- Convex constraints: $\Psi(w)$ is the indicator function of a closed convex set C, that is,

$$\Psi(w) = I_C(w) \triangleq \begin{cases} 0, & \text{if } w \in C, \\ +\infty, & \text{otherwise.} \end{cases}$$

We can also consider mixed regularizations such as $\Psi(w) = \lambda\|w\|_1 + (\sigma/2)\|w\|_2^2$. These examples cover a wide range of practical problems in machine learning.

A common approach for solving stochastic learning problems is to approximate the expected loss function $\phi(w)$ by using a finite set of independent observations $z_1, \ldots, z_T$, and solve the following problem to minimize the empirical loss:

$$\underset{w}{\text{minimize}} \quad \frac{1}{T} \sum_{t=1}^{T} f(w, z_t) + \Psi(w). \tag{2}$$

By our assumptions, this is a convex optimization problem. Depending on the structure of particular problems, they can be solved efficiently by interior-point methods (e.g., Ferris and Munson, 2003; Koh et al., 2007), quasi-Newton methods (e.g., Andrew and Gao, 2007), or accelerated first-order methods (Nesterov, 2007; Tseng, 2008; Beck and Teboulle, 2009). However, this *batch optimization* approach may not scale well for very large problems: even with first-order methods, evaluating one single gradient of the objective function in (2) requires going through the whole data set.

In this paper, we consider *online algorithms* that process samples sequentially as they become available. More specifically, we draw a sequence of i.i.d. samples $z_1, z_2, z_3, \ldots$, and use them to

calculate a sequence $w_1, w_2, w_3, \ldots$. Suppose at time t, we have the most up-to-date weight vector w_t. Whenever z_t is available, we can evaluate the loss $f(w_t, z_t)$, and also a subgradient $g_t \in \partial f(w_t, z_t)$ (here $\partial f(w,z)$ denotes the subdifferential of $f(w,z)$ with respect to w). Then we compute w_{t+1} based on these information.

The most widely used online algorithm is the *stochastic gradient descent* (SGD) method. Consider the general case $\Psi(w) = I_C(w) + \psi(w)$, where $I_C(w)$ is a "hard" set constraint and $\psi(w)$ is a "soft" regularization. The SGD method takes the form

$$w_{t+1} = \Pi_C\big(w_t - \alpha_t\,(g_t + \xi_t)\big), \tag{3}$$

where α_t is an appropriate stepsize, ξ_t is a subgradient of ψ at w_t, and $\Pi_C(\cdot)$ denotes Euclidean projection onto the set C. The SGD method belongs to the general scheme of *stochastic approximation*, which can be traced back to Robbins and Monro (1951) and Kiefer and Wolfowitz (1952). In general we are also allowed to use all previous information to compute w_{t+1}, and even second-order derivatives if the loss functions are smooth.

In a stochastic online setting, each weight vector w_t is a random variable that depends on $\{z_1, \ldots, z_{t-1}\}$, and so is the objective value $\phi(w_t)$. Assume an optimal solution $w^\star$ to the problem (1) exists, and let $\phi^\star = \phi(w^\star)$. The goal of online algorithms is to generate a sequence $\{w_t\}_{t=1}^\infty$ such that

$$\lim_{t\to\infty} \mathbf{E}\,\phi(w_t) = \phi^\star,$$

and hopefully with reasonable convergence rate. This is the case for the SGD method (3) if we choose the stepsize $\alpha_t = c/\sqrt{t}$, where c is a positive constant. The corresponding convergence rate is $O(1/\sqrt{t})$, which is indeed best possible for subgradient schemes with a *black-box* model, even in the case of deterministic optimization (Nemirovsky and Yudin, 1983). Despite such slow convergence and the associated low accuracy in the solutions (compared with batch optimization using, for example, interior-point methods), the SGD method has been very popular in the machine learning community due to its capability of scaling with very large data sets and good generalization performances observed in practice (e.g., Bottou and LeCun, 2004; Zhang, 2004; Shalev-Shwartz et al., 2007).

Nevertheless, a main drawback of the SGD method is its lack of capability in exploiting problem structure, especially for problems with explicit regularization. More specifically, the SGD method (3) treats the soft regularization $\psi(w)$ as a general convex function, and only uses its subgradient in computing the next weight vector. In this case, we can simply lump $\psi(w)$ into $f(w, z_t)$ and treat them as a single loss function. Although in theory the algorithm converges to an optimal solution (in expectation) as t goes to infinity, in practice it is usually stopped far before that. Even in the case of convergence in expectation, we still face (possibly big) variations in the solution due to the stochastic nature of the algorithm. Therefore, the regularization effect we hope to have by solving the problem (1) may be elusive for any particular solution generated by (3) based on finite random samples.

An important example and main motivation for this paper is ℓ_1-regularized stochastic learning, where $\Psi(w) = \lambda\|w\|_1$. In the case of batch learning, the empirical minimization problem (2) can be solved to very high precision, for example, by interior-point methods. Therefore simply rounding the weights with very small magnitudes toward zero is usually enough to produce desired sparsity. As a result, ℓ_1-regularization has been very effective in obtaining sparse solutions using the batch optimization approach in statistical learning (e.g., Tibshirani, 1996) and signal processing

(e.g., Chen et al., 1998). In contrast, the SGD method (3) hardly generates any sparse solution, and its inherent low accuracy makes the simple rounding approach very unreliable. Several principled soft-thresholding or truncation methods have been developed to address this problem (e.g., Langford et al., 2009; Duchi and Singer, 2009), but the levels of sparsity in their solutions are still unsatisfactory compared with the corresponding batch solutions.

In this paper, we develop *regularized dual averaging* (RDA) methods that can exploit the structure of (1) more effectively in a stochastic online setting. More specifically, each iteration of the RDA methods takes the form

$$w_{t+1} = \arg\min_{w} \left\{ \frac{1}{t} \sum_{\tau=1}^{t} \langle g_\tau, w \rangle + \Psi(w) + \frac{\beta_t}{t} h(w) \right\}, \tag{4}$$

where $h(w)$ is an auxiliary strongly convex function, and $\{\beta_t\}_{t\geq 1}$ is a nonnegative and nondecreasing input sequence, which determines the convergence properties of the algorithm. Essentially, at each iteration, this method minimizes the sum of three terms: a linear function obtained by averaging all previous subgradients (the dual average), the original regularization function $\Psi(w)$, and an additional strongly convex regularization term $(\beta_t/t)h(w)$. The RDA method is an extension of the *simple dual averaging* scheme of Nesterov (2009), which is equivalent to letting $\Psi(w)$ be the indicator function of a closed convex set.

For the RDA method to be practically efficient, we assume that the functions $\Psi(w)$ and $h(w)$ are *simple*, meaning that we are able to find a closed-form solution for the minimization problem in (4). Then the computational effort per iteration is only $O(n)$, the same as the SGD method. This assumption indeed holds in many cases. For example, if we let $\Psi(w) = \lambda\|w\|_1$ and $h(w) = (1/2)\|w\|_2^2$, then w_{t+1} has an entry-wise closed-from solution. This solution uses a much more aggressive truncation threshold than previous methods, thus results in significantly improved sparsity (see discussions in Section 5).

In terms of iteration complexity, we show that if $\beta_t = \Theta(\sqrt{t})$, that is, with order exactly $\sqrt{t}$, then the RDA method (4) has the standard convergence rate

$$\mathbf{E}\,\phi(\bar{w}_t) - \phi^\star \leq O\left(\frac{G}{\sqrt{t}}\right),$$

where $\bar{w}_t = (1/t)\sum_{\tau=1}^{t} w_\tau$ is the *primal average*, and G is a uniform upper bound on the norms of the subgradients g_t. If the regularization term $\Psi(w)$ is strongly convex, then setting $\beta_t \leq O(\ln t)$ gives a faster convergence rate $O(\ln t/t)$.

For stochastic optimization problems in which the loss functions $f(w,z)$ are all differentiable and have Lipschitz continuous gradients, we also develop an accelerated version of the RDA method that has the convergence rate

$$\mathbf{E}\,\phi(w_t) - \phi^\star \leq O(1)\left(\frac{L}{t^2} + \frac{Q}{\sqrt{t}}\right),$$

where L is the Lipschitz constant of the gradients, and Q^2 is an upper bound on the variances of the stochastic gradients. In addition to convergence in expectation, we show that the same orders of convergence rates hold with high probability.

1.2 Regularized Online Optimization

In *online optimization*, we use an online algorithm to generate a sequence of decisions w_t, one at a time, for $t = 1, 2, 3, \ldots$. At each time t, a previously unknown cost function f_t is revealed, and we encounter a loss $f_t(w_t)$. We assume that the cost functions f_t are convex for all $t \geq 1$. The goal of the online algorithm is to ensure that the total cost up to each time t, $\sum_{\tau=1}^{t} f_\tau(w_\tau)$, is not much larger than $\min_w \sum_{\tau=1}^{t} f_\tau(w)$, the smallest total cost of any fixed decision w from hindsight. The difference between these two cost is called the *regret* of the online algorithm. Applications of online optimization include online prediction of time series and sequential investment (e.g., Cesa-Bianchi and Lugosi, 2006).

In regularized online optimization, we add a convex regularization term $\Psi(w)$ to each cost function. The regret with respect to any fixed decision $w \in \operatorname{dom}\Psi$ is

$$R_t(w) \triangleq \sum_{\tau=1}^{t} \big(f_\tau(w_\tau) + \Psi(w_\tau)\big) - \sum_{\tau=1}^{t} \big(f_\tau(w) + \Psi(w)\big). \tag{5}$$

As in the stochastic setting, the online algorithm can query a subgradient $g_t \in \partial f_t(w_t)$ at each step, and possibly use all previous information, to compute the next decision w_{t+1}. It turns out that the simple subgradient method (3) is well suited for online optimization: with a stepsize $\alpha_t = \Theta(1/\sqrt{t})$, it has a regret $R_t(w) \leq O(\sqrt{t})$ for all $w \in \operatorname{dom}\Psi$ (Zinkevich, 2003). This regret bound cannot be improved in general for convex cost functions. However, if the cost functions are strongly convex, say with convexity parameter σ, then the same algorithm with stepsize $\alpha_t = 1/(\sigma t)$ gives an $O(\ln t)$ regret bound (e.g., Hazan et al., 2006; Bartlett et al., 2008).

Similar to the discussions on regularized stochastic learning, the online subgradient method (3) in general lacks the capability of exploiting the regularization structure. In this paper, we show that the same RDA method (4) can effectively exploit such structure in an online setting, and ensure the $O(\sqrt{t})$ regret bound with $\beta_t = \Theta(\sqrt{t})$. For strongly convex regularizations, setting $\beta_t = O(\ln t)$ yields the improved regret bound $O(\ln t)$.

Since there is no specifications on the probability distribution of the sequence of functions, nor assumptions like mutual independence, online optimization can be considered as a more general framework than stochastic learning. In this paper, we will first establish regret bounds of the RDA method for solving online optimization problems, then use them to derive convergence rates for solving stochastic learning problems.

1.3 Outline of Contents

The methods we develop apply to more general settings than $\mathbf{R}^n$ with Euclidean geometry. In Section 1.4, we introduce the necessary notations and definitions associated with a general finite-dimensional real vector space.

In Section 2, we present the generic RDA method for solving both the stochastic learning and online optimization problems, and give several concrete examples of the method.

In Section 3, we present the precise regret bounds of the RDA method for solving regularized online optimization problems.

In Section 4, we derive convergence rates of the RDA method for solving regularized stochastic learning problems. In addition to the rates of convergence in expectation, we also give associated high probability bounds.

In Section 5, we explain the connections of the RDA method to several related work, and analyze its capability of generating better sparse solutions than other methods.

In Section 6, we give an enhanced version of the ℓ_1-RDA method, and present computational experiments on the MNIST handwritten data set (LeCun et al., 1998). Our experiments show that the RDA method is capable of generate sparse solutions that are comparable to those obtained by batch learning using interior-point methods.

In Section 7, we discuss the RDA methods in the context of *structural convex optimization* and their connections to incremental subgradient methods. As an extension, we develop an accelerated version of the RDA method for stochastic optimization problems with smooth loss functions. We also discuss in detail the *p*-norm based RDA methods.

Appendices A-D contain technical proofs of our main results.

1.4 Notations and Generalities

Let $\mathcal{E}$ be a finite-dimensional real vector space, endowed with a norm $\|\cdot\|$. This norm defines a systems of balls: $\mathcal{B}(w,r) = \{u \in \mathcal{E} \,|\, \|u-w\| \le r\}$. Let $\mathcal{E}^*$ be the vector space of all linear functions on $\mathcal{E}$, and let $\langle s, w\rangle$ denote the value of $s \in \mathcal{E}^*$ at $w \in \mathcal{E}$. The dual space $\mathcal{E}^*$ is endowed with the dual norm $\|s\|_* = \max_{\|w\|\le 1} \langle s, w\rangle$.

A function $h : \mathcal{E} \to \mathbf{R} \cup \{+\infty\}$ is called *strongly convex* with respect to the norm $\|\cdot\|$ if there exists a constant $\sigma > 0$ such that

$$h(\alpha w + (1-\alpha)u) \le \alpha h(w) + (1-\alpha)h(u) - \frac{\sigma}{2}\alpha(1-\alpha)\|w-u\|^2, \qquad \forall\, w, u \in \operatorname{dom} h.$$

The constant σ is called the *convexity parameter*, or the *modulus* of strong convexity. Let $\operatorname{rint} \mathcal{C}$ denote the *relative interior* of a convex set $\mathcal{C}$ (Rockafellar, 1970, Section 6). If h is strongly convex with modulus σ, then for any $w \in \operatorname{dom} h$ and $u \in \operatorname{rint}(\operatorname{dom} h)$,

$$h(w) \ge h(u) + \langle s, w-u\rangle + \frac{\sigma}{2}\|w-u\|^2, \qquad \forall\, s \in \partial h(u).$$

See, for example, Goebel and Rockafellar (2008) and Juditsky and Nemirovski (2008).

In the special case of the coordinate vector space $\mathcal{E} = \mathbf{R}^n$, we have $\mathcal{E} = \mathcal{E}^*$, and the standard inner product $\langle s, w\rangle = s^T w = \sum_{i=1}^n s^{(i)} w^{(i)}$, where $w^{(i)}$ denotes the i-th coordinate of w. For the standard Euclidean norm, $\|w\| = \|w\|_2 = \sqrt{\langle w, w\rangle}$ and $\|s\|_* = \|s\|_2$. For any $w_0 \in \mathbf{R}^n$, the function $h(w) = (\sigma/2)\|w - w_0\|_2^2$ is strongly convex with modulus σ.

For another example, consider the ℓ_1-norm $\|w\| = \|w\|_1 = \sum_{i=1}^n |w^{(i)}|$ and its associated dual norm $\|w\|_* = \|w\|_\infty = \max_{1\le i\le n} |w^{(i)}|$. Let $\mathcal{S}_n$ be the standard simplex in $\mathbf{R}^n$, that is, $\mathcal{S}_n = \left\{w \in \mathbf{R}^n_+ \,|\, \sum_{i=1}^n w^{(i)} = 1\right\}$. Then the negative entropy function

$$h(w) = \sum_{i=1}^n w^{(i)} \ln w^{(i)} + \ln n, \tag{6}$$

with $\operatorname{dom} h = \mathcal{S}_n$, is strongly convex with respect to $\|\cdot\|_1$ with modulus 1 (see, e.g., Nesterov, 2005, Lemma 3). In this case, the unique minimizer of h is $w_0 = (1/n, \ldots, 1/n)$.

For a closed proper convex function Ψ, we use $\operatorname{Arg\,min}_w \Psi(w)$ to denote the (convex) set of minimizing solutions. If a convex function h has a unique minimizer, for example, when h is strongly convex, then we use $\arg\min_w h(w)$ to denote that single point.

Algorithm 1 Regularized dual averaging (RDA) method

input:

- an auxiliary function $h(w)$ that is strongly convex on $\operatorname{dom}\Psi$ and also satisfies

$$\arg\min_w h(w) \in \operatorname{Arg}\min_w \Psi(w). \tag{7}$$

- a nonnegative and nondecreasing sequence $\{\beta_t\}_{t\geq 1}$.

initialize: set $w_1 = \arg\min_w h(w)$ and $\bar{g}_0 = 0$.

for $t = 1, 2, 3, \ldots$ **do**

1. Given the function f_t, compute a subgradient $g_t \in \partial f_t(w_t)$.
2. Update the average subgradient:

$$\bar{g}_t = \frac{t-1}{t}\bar{g}_{t-1} + \frac{1}{t}g_t.$$

3. Compute the next weight vector:

$$w_{t+1} = \arg\min_w \left\{ \langle \bar{g}_t, w\rangle + \Psi(w) + \frac{\beta_t}{t}h(w) \right\}. \tag{8}$$

end for

2. Regularized Dual Averaging Method

In this section, we present the generic RDA method (Algorithm 1) for solving regularized stochastic learning and online optimization problems, and give several concrete examples. To unify notation, we use $f_t(w)$ to denote the cost function at each step t. For stochastic learning problems, we simply let $f_t(w) = f(w, z_t)$.

At the input to the RDA method, we need an auxiliary function h that is strongly convex on $\operatorname{dom}\Psi$. The condition (7) requires that its unique minimizer must also minimize the regularization function Ψ. This can be done, for example, by first choosing a starting point $w_0 \in \operatorname{Arg}\min_w \Psi(w)$ and an arbitrary strongly convex function $h'(w)$, then letting

$$h(w) = h'(w) - h'(w_0) - \langle \nabla h'(w_0), w - w_0\rangle.$$

In other words, $h(w)$ is the *Bregman divergence* from w_0 induced by $h'(w)$. If h' is not differentiable, but subdifferentiable at w_0, we can replace $\nabla h'(w_0)$ with a subgradient. The input sequence $\{\beta_t\}_{t\geq 1}$ determines the convergence rate, or regret bound, of the algorithm.

There are three steps in each iteration of the RDA method. Step 1 is to compute a subgradient of f_t at w_t, which is standard for all subgradient or gradient based methods. Step 2 is the online version of computing the average subgradient:

$$\bar{g}_t = \frac{1}{t}\sum_{\tau=1}^{t} g_\tau.$$

The name *dual averaging* comes from the fact that the subgradients live in the dual space $\mathcal{E}^*$.

Step 3 is most interesting and worth further explanation. In particular, the efficiency in computing w_{t+1} determines how useful the method is in practice. For this reason, we assume the regularization functions $\Psi(w)$ and $h(w)$ are *simple*. This means the minimization problem in (8) can be solved with little effort, especially if we are able to find a closed-form solution for w_{t+1}. At first sight, this assumption seems to be quite restrictive. However, the examples below show that this indeed is the case for many important learning problems in practice.

2.1 RDA Methods with General Convex Regularization

For a general convex regularization Ψ, we can choose any positive sequence $\{\beta_t\}_{t\geq 1}$ that is order exactly $\sqrt{t}$, to obtain an $O(1/\sqrt{t})$ convergence rate for stochastic learning, or an $O(\sqrt{t})$ regret bound for online optimization. We will state the formal convergence theorems in Sections 3 and 4. Here, we give several concrete examples. To be more specific, we choose a parameter $\gamma > 0$ and use the sequence

$$\beta_t = \gamma\sqrt{t}, \qquad t = 1, 2, 3, \ldots.$$

- *Nesterov's dual averaging method.* Let $\Psi(w)$ be the indicator function of a closed convex set $\mathcal{C}$. This recovers the *simple dual averaging* scheme in Nesterov (2009). If we choose $h(w) = (1/2)\|w\|_2^2$, then the Equation (8) yields

$$w_{t+1} = \Pi_{\mathcal{C}}\left(-\frac{\sqrt{t}}{\gamma}\bar{g}_t\right) = \Pi_{\mathcal{C}}\left(-\frac{1}{\gamma\sqrt{t}}\sum_{\tau=1}^{t} g_\tau\right). \tag{9}$$

  When $\mathcal{C} = \{w \in \mathbf{R}^n \,|\, \|w\|_1 \leq \delta\}$ for some $\delta > 0$, we have "hard" ℓ_1-regularization. In this case, although there is no closed-form solution for w_{t+1}, efficient algorithms for projection onto the ℓ_1-ball can be found, for example, in Duchi et al. (2008).

- *"Soft" ℓ_1-regularization.* Let $\Psi(w) = \lambda\|w\|_1$ for some $\lambda > 0$, and $h(w) = (1/2)\|w\|_2^2$. In this case, w_{t+1} has a closed-form solution (see Appendix A for the derivation):

$$w_{t+1}^{(i)} = \begin{cases} 0 & \text{if } \left|\bar{g}_t^{(i)}\right| \leq \lambda, \\ -\dfrac{\sqrt{t}}{\gamma}\left(\bar{g}_t^{(i)} - \lambda\,\mathrm{sgn}\left(\bar{g}_t^{(i)}\right)\right) & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n. \tag{10}$$

  Here $\mathrm{sgn}(\cdot)$ is the *sign* or *signum* function, that is, $\mathrm{sgn}(\omega)$ equals 1 if $\omega > 0$, -1 if $\omega < 0$, and 0 if $\omega = 0$. Whenever a component of $\bar{g}_t$ is less than λ in magnitude, the corresponding component of w_{t+1} is set to zero. Further extensions of the ℓ_1-RDA method, and associated computational experiments, are given in Section 6.

- *Exponentiated dual averaging method.* Let $\Psi(w)$ be the indicator function of the standard simplex $\mathcal{S}_n$, and $h(w)$ be the negative entropy function defined in (6). In this case,

$$w_{t+1}^{(i)} = \frac{1}{Z_{t+1}}\exp\left(-\frac{\sqrt{t}}{\gamma}\bar{g}_t^{(i)}\right), \qquad i = 1, \ldots, n,$$

  where Z_{t+1} is a normalization parameter such that $\sum_{i=1}^n w_{t+1}^{(i)} = 1$. This is the dual averaging version of the exponentiated gradient algorithm (Kivinen and Warmuth, 1997); see also Tseng and Bertsekas (1993) and Juditsky et al. (2005). We note that this example is also covered by Nesterov's dual averaging method.

We discuss in detail the special case of p-norm RDA method in Section 7.2. Several other examples, including ℓ_∞-norm and a hybrid ℓ_1/ℓ_2-norm (*Berhu*) regularization, also admit closed-form solutions for w_{t+1}. Their solutions are similar in form to those obtained in the context of the FOBOS algorithm in Duchi and Singer (2009).

2.2 RDA Methods with Strongly Convex Regularization

If the regularization term $\Psi(w)$ is strongly convex, we can use any nonnegative and nondecreasing sequence $\{\beta_t\}_{t\geq 1}$ that grows no faster than $O(\ln t)$, to obtain an $O(\ln t/t)$ convergence rate for stochastic learning, or an $O(\ln t)$ regret bound for online optimization. For simplicity, in the following examples, we use the zero sequence $\beta_t = 0$ for all $t \geq 1$. In this case, we do not need the auxiliary function $h(w)$, and the Equation (8) becomes

$$w_{t+1} = \arg\min_w \left\{ \langle \bar{g}_t, w \rangle + \Psi(w) \right\}.$$

- *ℓ_2^2-regularization.* Let $\Psi(w) = (\sigma/2)\|w\|_2^2$ for some $\sigma > 0$. In this case,

$$w_{t+1} = -\frac{1}{\sigma}\bar{g}_t = -\frac{1}{\sigma t}\sum_{\tau=1}^{t} g_\tau.$$

- *Mixed ℓ_1/ℓ_2^2-regularization.* Let $\Psi(w) = \lambda\|w\|_1 + (\sigma/2)\|w\|_2^2$ with $\lambda > 0$ and $\sigma > 0$. In this case, we have

$$w_{t+1}^{(i)} = \begin{cases} 0 & \text{if } |\bar{g}_t^{(i)}| \leq \lambda, \\ -\frac{1}{\sigma}\left(\bar{g}_t^{(i)} - \lambda \operatorname{sgn}\left(\bar{g}_t^{(i)}\right)\right) & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n.$$

- *Kullback-Leibler (KL) divergence regularization.* Let $\Psi(w) = \sigma D_{\mathrm{KL}}(w\|p)$, where the given probability distribution $p \in \operatorname{rint} \mathcal{S}_n$, and

$$D_{\mathrm{KL}}(w\|p) \triangleq \sum_{i=1}^{n} w^{(i)} \ln\left(\frac{w^{(i)}}{p^{(i)}}\right).$$

  Here $D_{\mathrm{KL}}(w\|p)$ is strongly convex with respect to $\|w\|_1$ with modulus 1. In this case,

$$w_{t+1}^{(i)} = \frac{1}{Z_{t+1}} p^{(i)} \exp\left(-\frac{1}{\sigma}\bar{g}_t^{(i)}\right),$$

  where Z_{t+1} is a normalization parameter such that $\sum_{i=1}^n w_{t+1}^{(i)} = 1$. KL divergence regularization has the *pseudo-sparsity* effect, meaning that most elements in w can be replaced by elements in the constant vector p without significantly increasing the loss function (e.g., Bradley and Bagnell, 2009).

3. Regret Bounds for Online Optimization

In this section, we give the precise regret bounds of the RDA method for solving regularized online optimization problems. The convergence rates for stochastic learning problems can be established based on these regret bounds, and will be given in the next section. For clarity, we gather here the general assumptions used throughout this paper:

- The regularization term $\Psi(w)$ is a closed proper convex function, and $\operatorname{dom}\Psi$ is closed. The symbol σ is dedicated to the convexity parameter of Ψ. Without loss of generality, we assume $\min_w \Psi(w) = 0$.
- For each $t \geq 1$, the function $f_t(w)$ is convex and subdifferentiable on $\operatorname{dom}\Psi$.
- The function $h(w)$ is strongly convex on $\operatorname{dom}\Psi$, and subdifferentiable on $\operatorname{rint}(\operatorname{dom}\Psi)$. Without loss of generality, assume $h(w)$ has convexity parameter 1 and $\min_w h(w) = 0$.

We will not repeat these general assumptions when stating our formal results later.

To facilitate regret analysis, we first give a few definitions. For any constant $D > 0$, we define the set

$$\mathcal{F}_D \triangleq \left\{ w \in \operatorname{dom}\Psi \;\middle|\; h(w) \leq D^2 \right\},$$

and let

$$\Gamma_D = \sup_{w \in \mathcal{F}_D} \inf_{g \in \partial\Psi(w)} \|g\|_*. \tag{11}$$

We use the convention $\inf_{g\in\emptyset} \|g\|_* = +\infty$, where $\emptyset$ denotes the empty set. As a result, if Ψ is not subdifferentiable everywhere on $\mathcal{F}_D$, that is, if $\partial\Psi(w) = \emptyset$ at some $w \in \mathcal{F}_D$, then we have $\Gamma_D = +\infty$. Note that Γ_D is not a Lipschitz-type constant which would be required to be an upper bound on all the subgradients; instead, we only require that at least one subgradient is bounded in norm by Γ_D at every point in the set $\mathcal{F}_D$.

We assume that the sequence of subgradients $\{g_t\}_{t\geq 1}$ generated by Algorithm 1 is bounded, that is, there exist a constant G such that

$$\|g_t\|_* \leq G, \quad \forall t \geq 1. \tag{12}$$

This is true, for example, if $\operatorname{dom}\Psi$ is compact and each f_t has Lipschitz-continuous gradient on $\operatorname{dom}\Psi$. We require that the input sequence $\{\beta_t\}_{t\geq 1}$ be chosen such that

$$\max\{\sigma, \beta_1\} > 0, \tag{13}$$

where σ is the convexity parameter of $\Psi(w)$. For convenience, we let $\beta_0 = \max\{\sigma, \beta_1\}$ and define the sequence of *regret bounds*

$$\Delta_t \triangleq \beta_t D^2 + \frac{G^2}{2} \sum_{\tau=0}^{t-1} \frac{1}{\sigma\tau + \beta_\tau} + \frac{2(\beta_0 - \beta_1)G^2}{(\beta_1 + \sigma)^2}, \qquad t = 1, 2, 3, \ldots, \tag{14}$$

where D is the constant used in the definition of $\mathcal{F}_D$. We could always set $\beta_1 \geq \sigma$, so that $\beta_0 = \beta_1$ and therefore the term $2(\beta_0 - \beta_1)G^2/(\beta_1 + \sigma)^2$ vanishes in the definition (14). However, when $\sigma > 0$, we would like to keep the flexibility of setting $\beta_t = 0$ for all $t \geq 1$, as we did in Section 2.2.

Theorem 1 *Let the sequences $\{w_t\}_{t\geq 1}$ and $\{g_t\}_{t\geq 1}$ be generated by Algorithm 1, and assume (12) and (13) hold. Then for any $t \geq 1$ and any $w \in \mathcal{F}_D$, we have:*

(a) The regret defined in (5) is bounded by Δ_t, that is,

$$R_t(w) \leq \Delta_t. \tag{15}$$

(b) The primal variables are bounded as

$$\|w_{t+1} - w\|^2 \leq \frac{2}{\sigma t + \beta_t}\big(\Delta_t - R_t(w)\big). \tag{16}$$

(c) If w is an interior point, that is, $\mathcal{B}(w,r) \subset \mathcal{F}_D$ for some $r > 0$, then

$$\|\bar{g}_t\|_* \leq \Gamma_D - \frac{1}{2}\sigma r + \frac{1}{rt}\big(\Delta_t - R_t(w)\big). \tag{17}$$

In Theorem 1, the bounds on $\|w_{t+1} - w\|^2$ and $\|\bar{g}_t\|_*$ depend on the regret $R_t(w)$. More precisely, they depend on $\Delta_t - R_t(w)$, which is the *slack* of the regret bound in (15). A smaller slack is equivalent to a larger regret $R_t(w)$, which means w is a better *fixed* solution for the online optimization problem (the best one gives the largest regret); correspondingly, the inequality (16) gives a tighter bound on $\|w_{t+1} - w\|^2$. In (17), the left-hand side $\|\bar{g}_t\|_*$ does not depend on any particular interior point w to compare with, but the right-hand side depends on both $R_t(w)$ and how far w is from the boundary of $\mathcal{F}_D$. The tightest bound on $\|\bar{g}_t\|_*$ can be obtained by taking the infimum of the right-hand side over all $w \in \operatorname{int}\mathcal{F}_D$. We further elaborate on part (c) through the following two examples:

- Consider the case when Ψ is the indicator function of a closed convex set $\mathcal{C}$. In this case, $\sigma = 0$ and $\partial\Psi(w)$ is the *normal cone* to $\mathcal{C}$ at w (Rockafellar, 1970, Section 23). By the definition (11), we have $\Gamma_D = 0$ because the zero vector is a subgradient at every $w \in \mathcal{C}$, even though the normal cones can be unbounded at the boundary of $\mathcal{C}$. In this case, if $\mathcal{B}(w,r) \subset \mathcal{F}_D$ for some $r > 0$, then (17) simplifies to

$$\|\bar{g}_t\|_* \leq \frac{1}{rt}\big(\Delta_t - R_t(w)\big).$$

- Consider the function $\Psi(w) = \sigma D_{\mathrm{KL}}(w\|p)$ with $\operatorname{dom}\Psi = \mathcal{S}_n$ (assuming $p \in \operatorname{rint}\mathcal{S}_n$). In this case, $\operatorname{dom}\Psi$, and hence $\mathcal{F}_D$, have empty interior. Therefore the bound in part (c) does not apply. In fact, the quantity Γ_D can be unbounded anyway. In particular, the subdifferentials of Ψ at the relative boundary of $\mathcal{S}_n$ are all empty. In the relative interior of $\mathcal{S}_n$, the subgradients (actually gradients) of Ψ always exist, but can become unbounded for points approaching the relative boundary. Nevertheless, the bounds in parts (a) and (b) still hold.

The proof of Theorem 1 is given in Appendix B. In the rest of this section, we discuss more concrete regret bounds depending on whether or not Ψ is strongly convex.

3.1 Regret Bound with General Convex Regularization

For a general convex regularization term Ψ, any nonnegative and nondecreasing sequence $\beta_t = \Theta(\sqrt{t})$ gives an $O(\sqrt{t})$ regret bound. Here we give detailed analysis for the sequence used in Section 2.1. More specifically, we choose a constant $\gamma > 0$ and let

$$\beta_t = \gamma\sqrt{t}, \qquad \forall\, t \geq 1. \tag{18}$$

We have the following corollary of Theorem 1.

Corollary 2 *Let the sequences $\{w_t\}_{t\geq 1}$ and $\{g_t\}_{t\geq 1}$ be generated by Algorithm 1 using $\{\beta_t\}_{t\geq 1}$ defined in (18), and assume (12) holds. Then for any $t \geq 1$ and any $w \in \mathcal{F}_D$:*

(a) The regret is bounded as

$$R_t(w) \leq \left(\gamma D^2 + \frac{G^2}{\gamma}\right)\sqrt{t}.$$

(b) The primal variables are bounded as

$$\frac{1}{2}\|w_{t+1} - w\|^2 \leq D^2 + \frac{G^2}{\gamma^2} - \frac{1}{\gamma\sqrt{t}}R_t(w).$$

(c) If w is an interior point, that is, $\mathcal{B}(w,r) \subset \mathcal{F}_D$ for some $r > 0$, then

$$\|\bar{g}_t\|_* \leq \Gamma_D + \left(\gamma D^2 + \frac{G^2}{\gamma}\right)\frac{1}{r\sqrt{t}} - \frac{1}{rt}R_t(w).$$

Proof To simplify regret analysis, let $\gamma \geq \sigma$. Therefore $\beta_0 = \beta_1 = \gamma$. Then Δ_t defined in (14) becomes

$$\Delta_t = \gamma\sqrt{t}D^2 + \frac{G^2}{2\gamma}\left(1 + \sum_{\tau=1}^{t-1}\frac{1}{\sqrt{\tau}}\right).$$

Next using the inequality

$$\sum_{\tau=1}^{t-1}\frac{1}{\sqrt{\tau}} \leq 1 + \int_1^t \frac{1}{\sqrt{\tau}}d\tau = 2\sqrt{t} - 1,$$

we get

$$\Delta_t \leq \gamma\sqrt{t}D^2 + \frac{G^2}{2\gamma}\left(1 + \left(2\sqrt{t} - 1\right)\right) = \left(\gamma D^2 + \frac{G^2}{\gamma}\right)\sqrt{t}.$$

Combining the above inequality and the conclusions of Theorem 1 proves the corollary. ■

The regret bound in Corollary 2 is essentially the same as the *online gradient descent* method of Zinkevich (2003), which has the form (3), with the stepsize $\alpha_t = 1/(\gamma\sqrt{t})$. The main advantage of the RDA method is its capability of exploiting the regularization structure, as shown in Section 2. The parameters D and G are not used explicitly in the algorithm. However, we need good estimates of them for choosing a reasonable value for γ. The best γ that minimizes the expression $\gamma D^2 + G^2/\gamma$ is

$$\gamma^\star = \frac{G}{D},$$

which leads to the simplified regret bound

$$R_t(w) \le 2GD\sqrt{t}.$$

If the total number of online iterations T is known in advance, then using a constant stepsize in the classical gradient method (3), say

$$\alpha_t = \frac{1}{\gamma^\star}\sqrt{\frac{2}{T}} = \frac{D}{G}\sqrt{\frac{2}{T}}, \qquad \forall t = 1,\ldots,T, \tag{19}$$

gives a slightly improved bound $R_T(w) \le \sqrt{2}GD\sqrt{T}$ (see, e.g., Nemirovski et al., 2009).

The bound in part (b) does not converge to zero. This result is still interesting because there is no special caution taken in the RDA method, more specifically in (8), to ensure the boundedness of the sequence w_t. In the case $\Psi(w) = 0$, as pointed out by Nesterov (2009), this may even look surprising since we are minimizing over $\mathcal{E}$ the sum of a linear function and a regularization term $(\gamma/\sqrt{t})h(w)$ that eventually goes to zero.

Part (c) gives a bound on the norm of the dual average. If $\Psi(w)$ is the indicator function of a closed convex set, then $\Gamma_D = 0$ and part (c) shows that $\bar{g}_t$ actually converges to zero if there exist an interior w in $\mathcal{F}_D$ such that $R_t(w) \ge 0$. However, a properly scaled version of $\bar{g}_t$, $-(\sqrt{t}/\gamma)\bar{g}_t$, tracks the optimal solution; see the examples in Section 2.1.

3.2 Regret Bounds with Strongly Convex Regularization

If the regularization function $\Psi(w)$ is strongly convex, that is, with a convexity parameter $\sigma > 0$, then any nonnegative, nondecreasing sequence that satisfies $\beta_t \le O(\ln t)$ will give an $O(\ln t)$ regret bound. If $\{\beta_t\}_{t\ge1}$ is not the all zero sequence, we can simply choose the auxiliary function $h(w) = (1/\sigma)\Psi(w)$. Here are several possibilities:

- *Positive constant sequences.* For simplicity, let $\beta_t = \sigma$ for $t \ge 0$. In this case,

$$\Delta_t = \sigma D^2 + \frac{G^2}{2\sigma}\sum_{\tau=0}^{t-1}\frac{1}{\tau+1} \le \sigma D^2 + \frac{G^2}{2\sigma}(1+\ln t).$$

- *Logarithmic sequences.* Let $\beta_t = \sigma(1+\ln t)$ for $t \ge 1$. In this case, $\beta_0 = \beta_1 = \sigma$ and

$$\Delta_t = \sigma(1+\ln t)D^2 + \frac{G^2}{2\sigma}\left(1+\sum_{\tau=1}^{t-1}\frac{1}{\tau+1+\ln\tau}\right) \le \left(\sigma D^2 + \frac{G^2}{2\sigma}\right)(1+\ln t).$$

- *The zero sequence.* Let $\beta_t = 0$ for $t \ge 1$. In this case, $\beta_0 = \sigma$ and

$$\Delta_t = \frac{G^2}{2\sigma}\left(1+\sum_{\tau=1}^{t-1}\frac{1}{\tau}\right) + \frac{2G^2}{\sigma} \le \frac{G^2}{2\sigma}(6+\ln t). \tag{20}$$

  Notice that in this last case, the regret bound does not depend on D.

When Ψ is strongly convex, we also conclude that, given two different points u and v, the regrets $R_t(u)$ and $R_t(v)$ cannot be nonnegative simultaneously if t is large enough. To see this, we notice that if $R_t(u)$ and $R_t(v)$ are nonnegative simultaneously for some t, then part (b) of Theorem 1 implies

$$\|w_{t+1}-u\|^2 \leq O\left(\frac{\ln t}{t}\right), \qquad \text{and} \qquad \|w_{t+1}-v\|^2 \leq O\left(\frac{\ln t}{t}\right),$$

which again implies

$$\|u-v\|^2 \leq (\|w_{t+1}-u\|+\|w_{t+1}-v\|)^2 \leq O\left(\frac{\ln t}{t}\right).$$

Therefore, if the event $R_t(u) \geq 0$ and $R_t(v) \geq 0$ happens for infinitely many t, we must have $u = v$. If $u \neq v$, then eventually at least one of the regrets associated with them will become negative. However, it is possible to construct sequences of functions f_t such that the points with nonnegative regrets do not converge to a fixed point.

4. Convergence Rates for Stochastic Learning

In this section, we give convergence rates of the RDA method when it is used to solve the regularized stochastic learning problem (1), and also the related high probability bounds. These rates and bounds are established not for the individual w_t's generated by the RDA method, but rather for the *primal average*

$$\bar{w}_t = \frac{1}{t}\sum_{\tau=1}^{t} w_\tau, \qquad t \geq 1.$$

4.1 Rate of Convergence in Expectation

Theorem 3 *Assume there exists an optimal solution $w^\star$ to the problem (1) that satisfies $h(w^\star) \leq D^2$ for some $D > 0$, and let $\phi^\star = \phi(w^\star)$. Let the sequences $\{w_t\}_{t\geq 1}$ and $\{g_t\}_{t\geq 1}$ be generated by Algorithm 1, and assume (12) holds. Then for any $t \geq 1$, we have:*

(a) The expected cost associated with the random variable $\bar{w}_t$ is bounded as

$$\mathbf{E}\,\phi(\bar{w}_t) - \phi^\star \leq \frac{1}{t}\Delta_t.$$

(b) The primal variables are bounded as

$$\mathbf{E}\,\|w_{t+1}-w^\star\|^2 \leq \frac{2}{\sigma t+\beta_t}\Delta_t.$$

(c) If $w^\star$ is an interior point, that is, $\mathcal{B}(w^\star, r) \subset \mathcal{F}_D$ for some $r > 0$, then

$$\mathbf{E}\|\bar{g}_t\|_* \leq \Gamma_D - \frac{1}{2}\sigma r + \frac{1}{rt}\Delta_t.$$

Proof First, we substitute all $f_\tau(\cdot)$ by $f(\cdot, z_\tau)$ in the definition of the regret

$$R_t(w^\star) = \sum_{\tau=1}^{t} \big(f(w_\tau, z_\tau) + \Psi(w_\tau) \big) - \sum_{\tau=1}^{t} \big(f(w^\star, z_\tau) + \Psi(w^\star) \big).$$

Let $\mathbf{z}[t]$ denote the collection of i.i.d. random variables $(z_1, \ldots, z_t)$. All the expectations in Theorem 3 are taken with respect to $\mathbf{z}[t]$, that is, the symbol $\mathbf{E}$ can be written more explicitly as $\mathbf{E}_{\mathbf{z}[t]}$. We note that the random variable w_τ, where $1 \le \tau \le t$, is a function of $(z_1, \ldots, z_{\tau-1})$, and is independent of $(z_\tau, \ldots, z_t)$. Therefore

$$\mathbf{E}_{\mathbf{z}[t]} \big(f(w_\tau, z_\tau) + \Psi(w_\tau) \big) = \mathbf{E}_{\mathbf{z}[\tau-1]} \big(\mathbf{E}_{z_\tau} f(w_\tau, z_\tau) + \Psi(w_\tau) \big) = \mathbf{E}_{\mathbf{z}[\tau-1]} \phi(w_\tau) = \mathbf{E}_{\mathbf{z}[t]} \phi(w_\tau),$$

and

$$\mathbf{E}_{\mathbf{z}[t]} \big(f(w^\star, z_\tau) + \Psi(w^\star) \big) = \mathbf{E}_{z_\tau} f(w^\star, z_\tau) + \Psi(w^\star) = \phi(w^\star) = \phi^\star.$$

Since $\phi^\star = \phi(w^\star) = \min_w \phi(w)$, we have

$$\mathbf{E}_{\mathbf{z}[t]} R_t(w^\star) = \sum_{\tau=1}^{t} \mathbf{E}_{\mathbf{z}[t]} \phi(w_\tau) - t\phi^\star \ge 0. \tag{21}$$

By convexity of ϕ, we have

$$\phi(\bar{w}_t) = \phi \left(\frac{1}{t} \sum_{\tau=1}^{t} w_\tau \right) \le \frac{1}{t} \sum_{\tau=1}^{t} \phi(w_\tau)$$

Taking expectation with respect to $\mathbf{z}[t]$ and subtracting $\phi^\star$, we have

$$\mathbf{E}_{\mathbf{z}[t]} \phi(\bar{w}_t) - \phi^\star \le \frac{1}{t} \left(\sum_{\tau=1}^{t} \mathbf{E}_{\mathbf{z}[t]} \phi(\bar{w}_\tau) - t\phi^\star \right) = \frac{1}{t} \mathbf{E}_{\mathbf{z}[t]} R_t(w^\star).$$

Then part (a) follows from that of Theorem 1, which states that $R_t(w^\star) \le \Delta_t$ for all realizations of $\mathbf{z}[t]$. Similarly, parts (b) and (c) follow from those of Theorem 1 and (21). ■

Specific convergence rates can be obtained in parallel with the regret bounds discussed in Sections 3.1 and 3.2. We only need to divide every regret bound by t to obtain the corresponding rate of convergence in expectation. More specifically, using appropriate sequences $\{\beta_t\}_{t \ge 1}$, we have $\mathbf{E}\phi(\bar{w}_t)$ converging to $\phi^\star$ with rate $O(1/\sqrt{t})$ for general convex regularization, and $O(\ln t/t)$ for strongly convex regularization.

The bound in part (b) applies to both the case $\sigma = 0$ and the case $\sigma > 0$. For the latter, we can derive a slightly different and more specific bound. When Ψ has convexity parameter $\sigma > 0$, so is the function ϕ. Therefore,

$$\phi(w_t) \ge \phi(w^\star) + \langle s, w_t - w^\star \rangle + \frac{\sigma}{2} \|w_t - w^\star\|^2, \qquad \forall s \in \partial\phi(w^\star).$$

Since $w^\star$ is the minimizer of ϕ, we must have $0 \in \partial\phi(w^\star)$ (Rockafellar, 1970, Section 27). Setting $s = 0$ in the above inequality and rearranging terms, we have

$$\|w_t - w^\star\|^2 \le \frac{2}{\sigma} \left(\phi(w_t) - \phi^\star \right).$$

Taking expectation of both sides of the above inequality leads to

$$\mathbf{E}\|w_t - w^\star\|^2 \leq \frac{2}{\sigma}\left(\mathbf{E}\phi(w_t) - \phi^\star\right) \leq \frac{2}{\sigma t}\Delta_t, \tag{22}$$

where in the last step we used part (a) of Theorem 3. This bound directly relate w_t to Δ_t.

Next we take a closer look at the quantity $\mathbf{E}\|\bar{w}_t - w^\star\|^2$. By convexity of $\|\cdot\|^2$, we have

$$\mathbf{E}\|\bar{w}_t - w^\star\|^2 \leq \frac{1}{t}\sum_{\tau=1}^{t}\mathbf{E}\|w_\tau - w^\star\|^2 \tag{23}$$

If $\sigma = 0$, then it is simply bounded by a constant because each $\mathbf{E}\|w_\tau - w^\star\|^2$ for $1 \leq \tau \leq t$ is bounded by a constant. When $\sigma > 0$, the optimal solution $w^\star$ is unique, and we have:

Corollary 4 *If Ψ is strongly convex with convexity parameter $\sigma > 0$ and $\beta_t = O(\ln t)$, then*

$$\mathbf{E}\|\bar{w}_t - w^\star\|^2 \leq O\left(\frac{(\ln t)^2}{t}\right).$$

Proof For the ease of presentation, we consider the case $\beta_t = 0$ for all $t \geq 1$. Substituting the bound on Δ_t in (20) into the inequality (22) gives

$$\mathbf{E}\|w_t - w^\star\|^2 \leq \frac{(6 + \ln t)\,G^2}{t\,\sigma^2}, \qquad \forall t \geq 1.$$

Then by (23),

$$\mathbf{E}\|\bar{w}_t - w^\star\|^2 \leq \frac{1}{t}\sum_{\tau=1}^{t}\left(\frac{6}{\tau} + \frac{\ln\tau}{\tau}\right)\frac{G^2}{\sigma^2} \leq \frac{1}{t}\left(6(1+\ln t) + \frac{1}{2}(\ln t)^2\right)\frac{G^2}{\sigma^2}.$$

In other words, $\mathbf{E}\|\bar{w}_t - w^\star\|^2$ converges to zero with rate $O((\ln t)^2/t)$. This can be shown for any $\beta_t = O(\ln t)$; see Section 3.2 for other choices of β_t. ■

As a further note, the conclusions in Theorem 3 still hold if the assumption (12) is weakened to

$$\mathbf{E}\,\|g_t\|_*^2 \leq G^2, \qquad \forall t \geq 1. \tag{24}$$

However, we need (12) in order to prove the high probability bounds presented next.

4.2 High Probability Bounds

For stochastic learning problems, in addition to the rates of convergence in expectation, it is often desirable to obtain confidence level bounds for approximate solutions. For this purpose, we start from part (a) of Theorem 3, which states $\mathbf{E}\phi(w_t) - \phi^\star \leq (1/t)\Delta_t$. By Markov's inequality, we have for any $\varepsilon > 0$,

$$\mathrm{Prob}\left(\phi(\bar{w}_t) - \phi^\star > \varepsilon\right) \leq \frac{\Delta_t}{\varepsilon t}. \tag{25}$$

This bound holds even with the weakened assumption (24). However, it is possible to have much tighter bounds under more restrictive assumptions. To this end, we have the following result.

Theorem 5 *Assume there exist constants D and G such that $h(w^\star) \le D^2$, and $h(w_t) \le D^2$ and $\|g_t\|_* \le G$ for all $t \ge 1$. Then for any $\delta \in (0,1)$, we have, with probability at least $1-\delta$,*

$$\phi(\bar{w}_t) - \phi^\star \le \frac{\Delta_t}{t} + \frac{8GD\sqrt{\ln(1/\delta)}}{\sqrt{t}}, \qquad \forall t \ge 1. \tag{26}$$

Theorem 5 is proved in Appendix C.

From our results in Section 3.1, with the input sequence $\beta_t = \gamma\sqrt{t}$ for all $t \ge 1$, we have $\Delta_t = O(\sqrt{t})$ regardless of $\sigma = 0$ or $\sigma > 0$. Therefore, $\phi(\bar{w}_t) - \phi^\star = O(1/\sqrt{t})$ with high probability. To simplify further discussion, let $\gamma = G/D$, hence $\Delta_t \le 2GD\sqrt{t}$ (see Section 3.1). In this case, if $\delta \le 1/e \approx 0.368$, then with probability at least $1-\delta$,

$$\phi(\bar{w}_t) - \phi^\star \le \frac{10GD\sqrt{\ln(1/\delta)}}{\sqrt{t}}.$$

Letting $\varepsilon = 10GD\sqrt{\ln(1/\delta)}/\sqrt{t}$, then the above bound is equivalent to

$$\text{Prob}\left(\phi(\bar{w}_t) - \phi^\star > \varepsilon\right) \le \exp\left(-\frac{\varepsilon^2 t}{(10GD)^2}\right),$$

which is much tighter than the one in (25). It follows that for any chosen accuracy ε and $0 < \delta \le 1/e$, the sample size

$$t \ge \frac{(10GD)^2 \ln(1/\delta)}{\varepsilon^2}$$

guarantees that, with probability at least $1-\delta$, $\bar{w}_t$ is an ε-optimal solution of the original stochastic optimization problem (1).

When Ψ is strongly convex ($\sigma > 0$), our results in Section 3.2 show that we can obtain regret bounds $\Delta_t = O(\ln t)$ using $\beta_t = O(\ln t)$. However, the high probability bound in Theorem 5 does not improve: we still have $\phi(\bar{w}_t) - \phi^\star = O(1/\sqrt{t})$, not $O(\ln t/t)$. The reason is that the concentration inequality (Azuma, 1967) used in proving Theorem 5 cannot take advantage of the strong-convexity property. By using a refined concentration inequality due to Freedman (1975), Kakade and Tewari (2009, Theorem 2) showed that for strongly convex stochastic learning problems, with probability at least $1 - 4\delta\ln t$,

$$\phi(\bar{w}_t) - \phi^\star \le \frac{R_t(w^\star)}{t} + 4\frac{\sqrt{R_t(w^\star)}}{t}\sqrt{\frac{G^2\ln(1/\delta)}{\sigma}} + \max\left\{\frac{16G^2}{\sigma}, 6B\right\}\frac{\ln(1/\delta)}{t}.$$

In our context, the constant B is an upper bound on $f(w,z) + \Phi(w)$ for $w \in \mathcal{F}_D$. Using the regret bound $R(w^\star) \le \Delta_t$, this gives

$$\phi(\bar{w}_t) - \phi^\star \le \frac{\Delta_t}{t} + O\left(\frac{\sqrt{\Delta_t \ln(1/\delta)}}{t} + \frac{\ln(1/\delta)}{t}\right).$$

Here the constants hidden in the O-notation are determined by G, σ and D. Plugging in $\Delta_t = O(\ln t)$, we have $\phi(\bar{w}_t) - \phi^\star = O(\ln t/t)$ with high probability. The additional penalty of getting the high probability bound, compared with the rate of convergence in expectation, is only $O(\sqrt{\ln t}/t)$.

5. Related Work

As we pointed out in Section 2.1, if Ψ is the indicator function of a convex set $\mathcal{C}$, then the RDA method recovers the simple dual averaging scheme in Nesterov (2009). This special case also belongs to a more general primal-dual algorithmic framework developed by Shalev-Shwartz and Singer (2006), which can be expressed equivalently in our notation:

$$w_{t+1} = \arg\min_{w\in\mathcal{C}} \left\{ \frac{1}{\gamma\sqrt{t}} \left\langle \sum_{\tau=1}^{t} d_\tau^t, w \right\rangle + h(w) \right\},$$

where $(d_1^t, \ldots, d_t^t)$ is the set of dual variables that can be chosen at time t. The simple dual averaging scheme (9) is in fact the *passive* extreme of their framework in which the dual variables are simply chosen as the subgradients and do not change over time, that is,

$$d_\tau^t = g_\tau, \qquad \forall \tau \le t, \qquad \forall t \ge 1. \tag{27}$$

However, with the addition of a general regularization term $\Psi(w)$ as in (4), the convergence analysis and $O(\sqrt{t})$ regret bound of the RDA method do *not* follow directly as corollaries of either Nesterov (2009) or Shalev-Shwartz and Singer (2006). Our analysis in Appendix B extends the framework of Nesterov (2009).

Shalev-Shwartz and Kakade (2009) extended the primal-dual framework of Shalev-Shwartz and Singer (2006) to strongly convex functions and obtained $O(\ln t)$ regret bound. In the context of this paper, their algorithm takes the form

$$w_{t+1} = \arg\min_{w\in\mathcal{C}} \left\{ \frac{1}{\sigma t} \left\langle \sum_{\tau=1}^{t} d_\tau^t, w \right\rangle + h(w) \right\},$$

where σ is the convexity parameter of Ψ, and $h(w) = (1/\sigma)\Psi(w)$. The passive extreme of this method, with the dual variables chosen in (27), is equivalent to a special case of the RDA method with $\beta_t = 0$ for all $t \ge 1$.

Other than improving the iteration complexity, the idea of treating the regularization explicitly in each step of a subgradient-based method (instead of lumping it together with the loss function and taking their subgradients) is mainly motivated by practical considerations, such as obtaining sparse solutions. In the case of ℓ_1-regularization, this leads to soft-thresholding type of algorithms, in both batch learning (e.g., Figueiredo et al., 2007; Wright et al., 2009; Bredies and Lorenz, 2008; Beck and Teboulle, 2009) and the online setting (e.g., Langford et al., 2009; Duchi and Singer, 2009; Shalev-Shwartz and Tewari, 2009). Most of these algorithms can be viewed as extensions of classical gradient methods (including mirror-descent methods) in which the new iterate is obtained by stepping from the current iterate along a single subgradient, and then followed by a truncation. Other types of algorithms include an interior-point based stochastic approximation scheme by Carbonetto et al. (2009), and Balakrishnan and Madigan (2008), where a modified shrinkage algorithm is developed based on sequential quadratic approximations of the loss function.

The main point of this paper, is to show that dual-averaging based methods can be more effective in exploiting the regularization structure, especially in a stochastic or online setting. To demonstrate this point, we compare the RDA method with the FOBOS method studied in Duchi and Singer (2009). In an online setting, each iteration of the FOBOS method consists of the following two

steps:

$$w_{t+\frac{1}{2}} = w_t - \alpha_t g_t,$$
$$w_{t+1} = \arg\min_w \left\{ \frac{1}{2} \left\| w - w_{t+\frac{1}{2}} \right\|_2^2 + \alpha_t \Psi(w) \right\}.$$

For convergence with optimal rates, the stepsize α_t is set to be $\Theta(1/\sqrt{t})$ for general convex regularizations and $\Theta(1/t)$ if Ψ is strongly convex. This method is based on a technique known as *forward-backward splitting*, which was first proposed by Lions and Mercier (1979) and later analyzed by Chen and Rockafellar (1997) and Tseng (2000). For easy comparison with the RDA method, we rewrite the FOBOS method in an equivalent form

$$w_{t+1} = \arg\min_w \left\{ \langle g_t, w \rangle + \Psi(w) + \frac{1}{2\alpha_t} \|w - w_t\|_2^2 \right\}. \tag{28}$$

Compared with this form of the FOBOS method, the RDA method (8) uses the average subgradient $\bar{g}_t$ instead of the current subgradient g_t; it uses a global proximal function, say $h(w) = (1/2)\|w\|_2^2$, instead of its local Bregman divergence $(1/2)\|w - w_t\|_2^2$; moreover, the coefficient for the proximal function is $\beta_t/t = \Theta(1/\sqrt{t})$ instead of $1/\alpha_t = \Theta(\sqrt{t})$ for general convex regularization, and $O(\ln t/t)$ instead of $\Theta(t)$ for strongly convex regularization. Although these two methods have the same order of iteration complexity, the differences list above contribute to quite different properties of their solutions.

These differences can be better understood in the special case of ℓ_1-regularization, that is, when $\Psi(w) = \lambda\|w\|_1$. In this case, the FOBOS method is equivalent to a special case of the *Truncated Gradient* (TG) method of Langford et al. (2009). The TG method truncates the solutions obtained by the standard SGD method every K steps; more specifically,

$$w_{t+1}^{(i)} = \begin{cases} \text{trnc}\left(w_t^{(i)} - \alpha_t g_t^{(i)}, \lambda_t^{\text{TG}}, \theta \right) & \text{if } \text{mod}(t,K) = 0, \\ w_t^{(i)} - \alpha_t g_t^{(i)} & \text{otherwise,} \end{cases} \tag{29}$$

where $\lambda_t^{\text{TG}} = \alpha_t \lambda K$, $\text{mod}(t,K)$ is the remainder on division of t by K, and

$$\text{trnc}(\omega, \lambda_t^{\text{TG}}, \theta) = \begin{cases} 0 & \text{if } |\omega| \le \lambda_t^{\text{TG}}, \\ \omega - \lambda_t^{\text{TG}} \,\text{sgn}(\omega) & \text{if } \lambda_t^{\text{TG}} < |\omega| \le \theta, \\ \omega & \text{if } |\omega| > \theta. \end{cases}$$

When $K = 1$ and $\theta = +\infty$, the TG method is the same as the FOBOS method (28) with ℓ_1-regularization. Now comparing the truncation threshold λ_t^{TG} and the threshold λ used in the ℓ_1-RDA method (10): with $\alpha_t = \Theta(1/\sqrt{t})$, we have $\lambda_t^{\text{TG}} = \Theta(1/\sqrt{t})\lambda$. This $\Theta(1/\sqrt{t})$ discount factor is also common for other previous work that use soft-thresholding, including Shalev-Shwartz and Tewari (2009). It is clear that the RDA method uses a much more aggressive truncation threshold, thus is able to generate significantly more sparse solutions. This is confirmed by our computational experiments in the next section.

Most recently, Duchi et al. (2010) developed a family of subgradient methods that can adaptively modifying the proximal function (squared Mahalanobis norms) at each iteration, in order to better incorporate learned knowledge about geometry of the data. Their methods include extensions for both the mirror-descent type of algorithms like (28) and the RDA methods studied in this paper.

Algorithm 2 Enhanced ℓ_1-RDA method

Input: $\gamma > 0$, $\rho \geq 0$

Initialize: $w_1 = 0$, $\bar{g}_0 = 0$.

for $t = 1, 2, 3, \ldots$ **do**

1. Given the function f_t, compute subgradient $g_t \in \partial f_t(w_t)$.
2. Compute the dual average
$$\bar{g}_t = \frac{t-1}{t}\bar{g}_{t-1} + \frac{1}{t} g_t.$$
3. Let $\lambda_t^{\mathrm{RDA}} = \lambda + \gamma\rho/\sqrt{t}$, and compute w_{t+1} entry-wise:
$$w_{t+1}^{(i)} = \begin{cases} 0 & \text{if } \left|\bar{g}_t^{(i)}\right| \leq \lambda_t^{\mathrm{RDA}}, \\ -\dfrac{\sqrt{t}}{\gamma}\left(\bar{g}_t^{(i)} - \lambda_t^{\mathrm{RDA}} \mathrm{sgn}\left(\bar{g}_t^{(i)}\right)\right) & \text{otherwise}, \end{cases} \qquad i = 1, \ldots, n. \tag{30}$$

end for

6. Computational Experiments with ℓ_1-Regularization

In this section, we provide computational experiments of the ℓ_1-RDA method on the MNIST data set of handwritten digits (LeCun et al., 1998). Our purpose here is mainly to illustrate the basic characteristics of the ℓ_1-RDA method, rather than comprehensive performance evaluation on a wide range of data sets. First, we describe a variant of the ℓ_1-RDA method that is capable of getting enhanced sparsity in the solution.

6.1 Enhanced ℓ_1-RDA Method

The enhanced ℓ_1-RDA method shown in Algorithm 2 is a special case of Algorithm 1. It is derived by setting $\Psi(w) = \lambda\|w\|_1$, $\beta_t = \gamma\sqrt{t}$, and replacing $h(w)$ with a parameterized version

$$h_\rho(w) = \frac{1}{2}\|w\|_2^2 + \rho\|w\|_1, \tag{31}$$

where $\rho \geq 0$ is a *sparsity-enhancing* parameter. Note that $h_\rho(w)$ is strongly convex with modulus 1 for any $\rho \geq 0$. Hence the convergence rate of this algorithm is the same as if we choose $h(w) = (1/2)\|w\|_2^2$. In this case, the Equation (8) becomes

$$\begin{aligned} w_{t+1} &= \arg\min_w \left\{ \langle \bar{g}_t, w\rangle + \lambda\|w\|_1 + \frac{\gamma}{\sqrt{t}}\left(\frac{1}{2}\|w\|_2^2 + \rho\|w\|_1\right)\right\} \\ &= \arg\min_w \left\{ \langle \bar{g}_t, w\rangle + \lambda_t^{\mathrm{RDA}}\|w\|_1 + \frac{\gamma}{2\sqrt{t}}\|w\|_2^2\right\}, \end{aligned}$$

where $\lambda_t^{\mathrm{RDA}} = \lambda + \gamma\rho/\sqrt{t}$. The above minimization problem has a closed-form solution given in (30) (see Appendix A for the derivation). By letting $\rho > 0$, the effective truncation threshold λ_t^{RDA} is larger than λ, especially in the initial phase of the online process. For problems without explicit ℓ_1-regularization in the objective function, that is, when $\lambda = 0$, this still gives a diminishing truncation threshold $\gamma\rho/\sqrt{t}$.

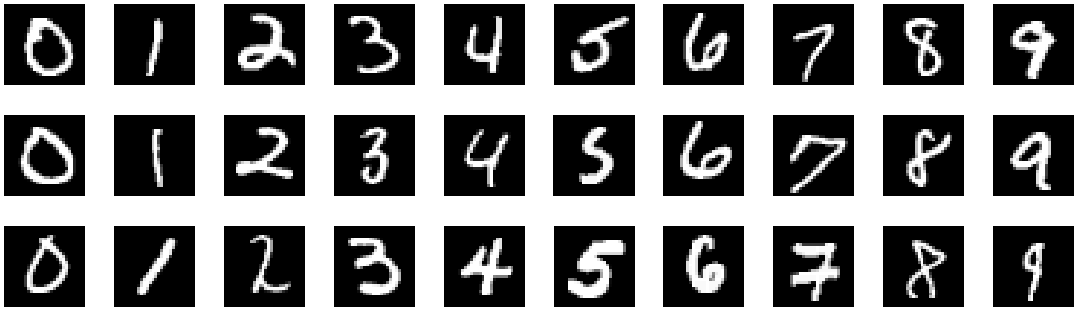

Figure 1: Sample images from the MNIST data set, with gray-scale from 0 to 255.

We can also restrict ℓ_1-regularization on part of the optimization variables only. For example, in support vector machines or logistic regression, we usually want the bias terms to be free of regularization. In this case, we can simply replace λ_t^{RDA} by 0 for the corresponding coordinates in (30).

6.2 Experiments on the MNIST Data Set

Each image in the MNIST data set is represented by a 28×28 gray-scale pixel-map, for a total of 784 features. Each of the 10 digits has roughly 6,000 training examples and 1,000 testing examples. Some of the samples are shown in Figure 1. From the perspective of using stochastic and online algorithms, the number of features and size of the data set are considered very small. Nevertheless, we choose this data set because the computational results are easy to visualize. No preprocessing of the data is employed.

We use ℓ_1-regularized logistic regression to do binary classification on each of the 45 pairs of digits. More specifically, let $z = (x, y)$ where $x \in \mathbf{R}^{784}$ represents a gray-scale image and $y \in \{+1, -1\}$ is the binary label, and let $w = (\tilde{w}, b)$ where $\tilde{w} \in \mathbf{R}^{784}$ and b is the bias. Then the loss function and regularization term in (1) are

$$f(w,z) = \log\left(1 + \exp\left(-y(\tilde{w}^T x + b)\right)\right), \qquad \Psi(w) = \lambda \|\tilde{w}\|_1.$$

Note that we do not apply regularization on the bias term b. In the experiments, we compare the (enhanced) ℓ_1-RDA method (Algorithm 2) with the SGD method

$$w_{t+1}^{(i)} = w_t^{(i)} - \alpha_t \left(g_t^{(i)} + \lambda \operatorname{sgn}(w_t^{(i)})\right), \qquad i = 1, \ldots, n,$$

and the TG method (29) with $\theta = \infty$. These three online algorithms have similar convergence rates and the same order of computational cost per iteration. We also compare them with the batch optimization approach, more specifically solving the empirical minimization problem (2) using an efficient interior-point method (IPM) of Koh et al. (2007).

Each pair of digits have about 12,000 training examples and 2,000 testing examples. We use online algorithms to go through the (randomly permuted) data only once, therefore the algorithms stop at $T = 12{,}000$. We vary the regularization parameter λ from 0.01 to 10. As a reference, the maximum λ for the batch optimization case (Koh et al., 2007) is mostly in the range of $30 - 50$ (beyond which the optimal weights are all zeros). In the ℓ_1-RDA method, we use $\gamma = 5{,}000$, and set ρ to

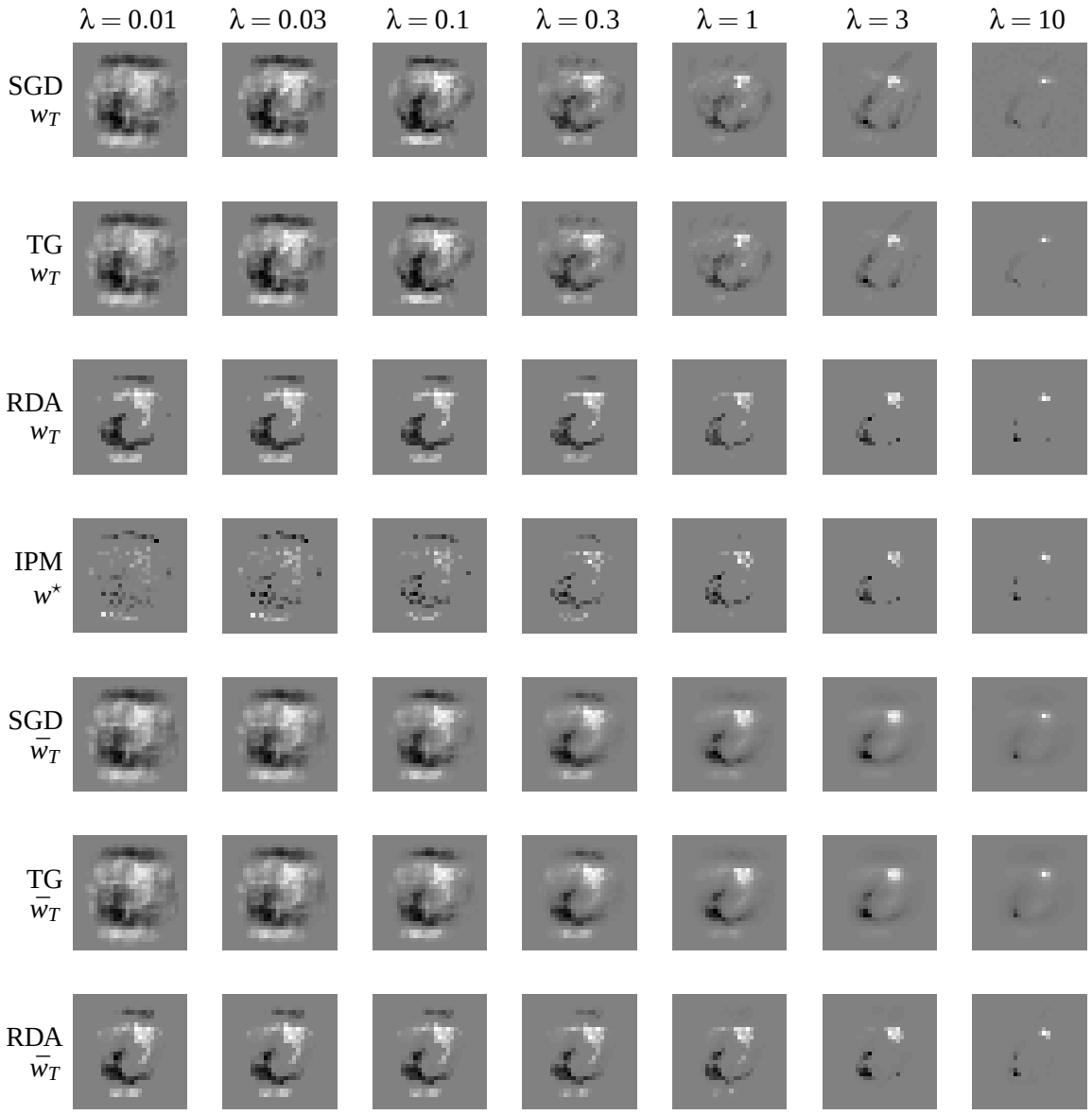

Figure 2: Sparsity patterns of w_T and $\bar{w}_T$ for classifying the digits 6 and 7 when varying the parameter λ from 0.01 to 10 in ℓ_1-regularized logistic regression. The background gray represents the value zero, bright spots represent positive values and dark spots represent negative values. Each column corresponds to a value of λ labeled at the top. The top three rows are the weights w_T (without averaging) from the last iteration of the three online algorithms; the middle row shows optimal solutions of the batch optimization problem solved by interior-point method (IPM); the bottom three rows show the averaged weights $\bar{w}_T$ in the three online algorithms. Both the TG and RDA methods were run with parameters for enhanced ℓ_1-regularization, that is, $K = 10$ for TG and $\gamma\rho = 25$ for RDA.

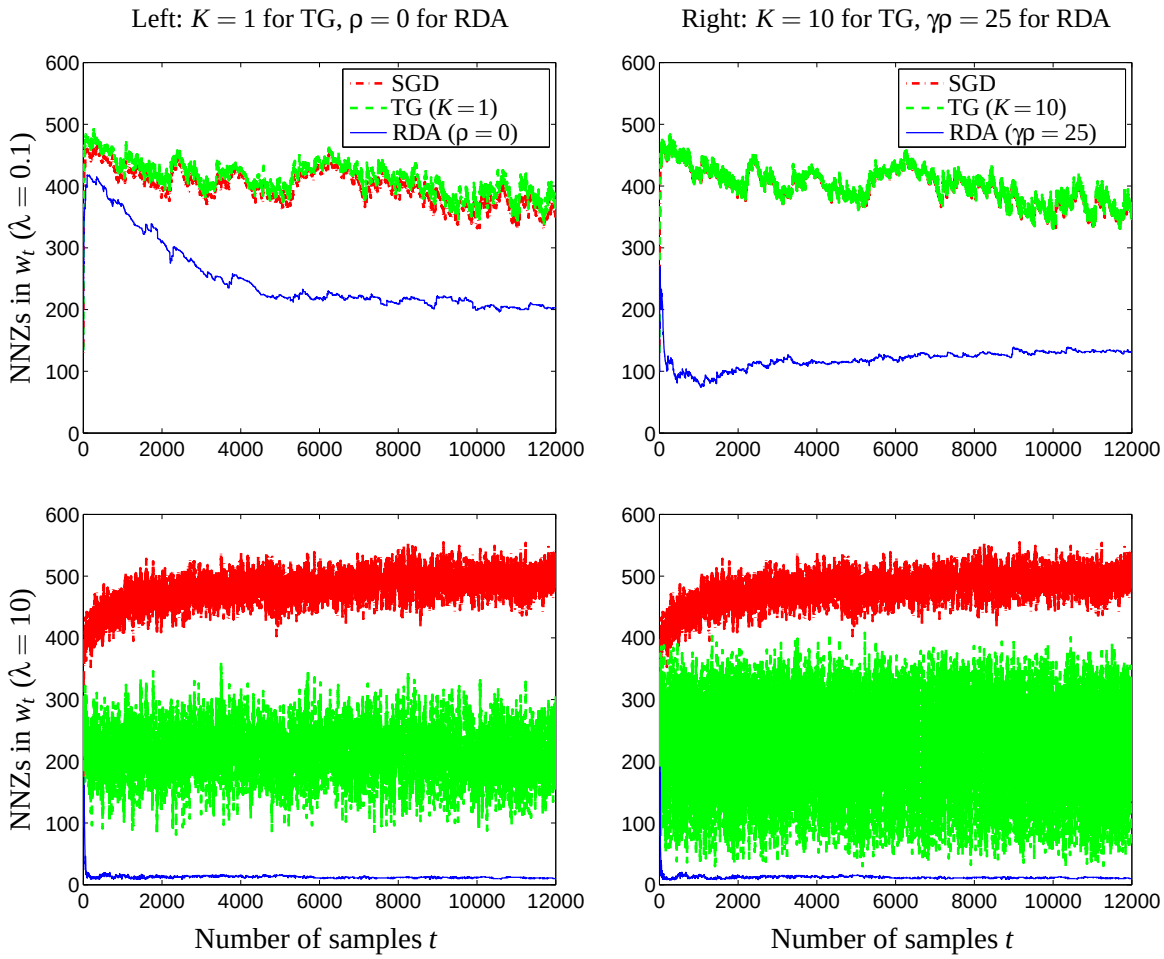

Figure 3: Number of non-zeros (NNZs) in w_t for the three online algorithms (classifying the pair 6 and 7). The left column shows SGD, TG with $K = 1$, and RDA with $\rho = 0$; the right column shows SGD, TG with $K = 10$, and RDA with $\gamma\rho = 25$. The same curves for SGD are plotted in both columns for clear comparison. The two rows correspond to $\lambda = 0.1$ and $\lambda = 10$, respectively.

be either 0 for basic regularization, or 0.005 (effectively $\gamma\rho = 25$) for enhanced regularization effect. These parameters are chosen by cross-validation. For the SGD and TG methods, we use a constant stepsize $\alpha = (1/\gamma)\sqrt{2/T}$ for comparable convergence rate; see (19) and related discussions. In the TG method, the period K is set to be either 1 for basic regularization (same as FOBOS), or 10 for periodic enhanced regularization effect.

Figure 2 shows the sparsity patterns of the solutions w_T and $\bar{w}_T$ for classifying the digits 6 and 7. The algorithmic parameters used are: $K = 10$ for the TG method, and $\gamma\rho = 25$ for the RDA method. It is clear that the RDA method gives more sparse solutions than both SGD and TG methods. The sparsity pattern obtained by the RDA method is very similar to the batch optimization results solved by IPM, especially for larger λ.

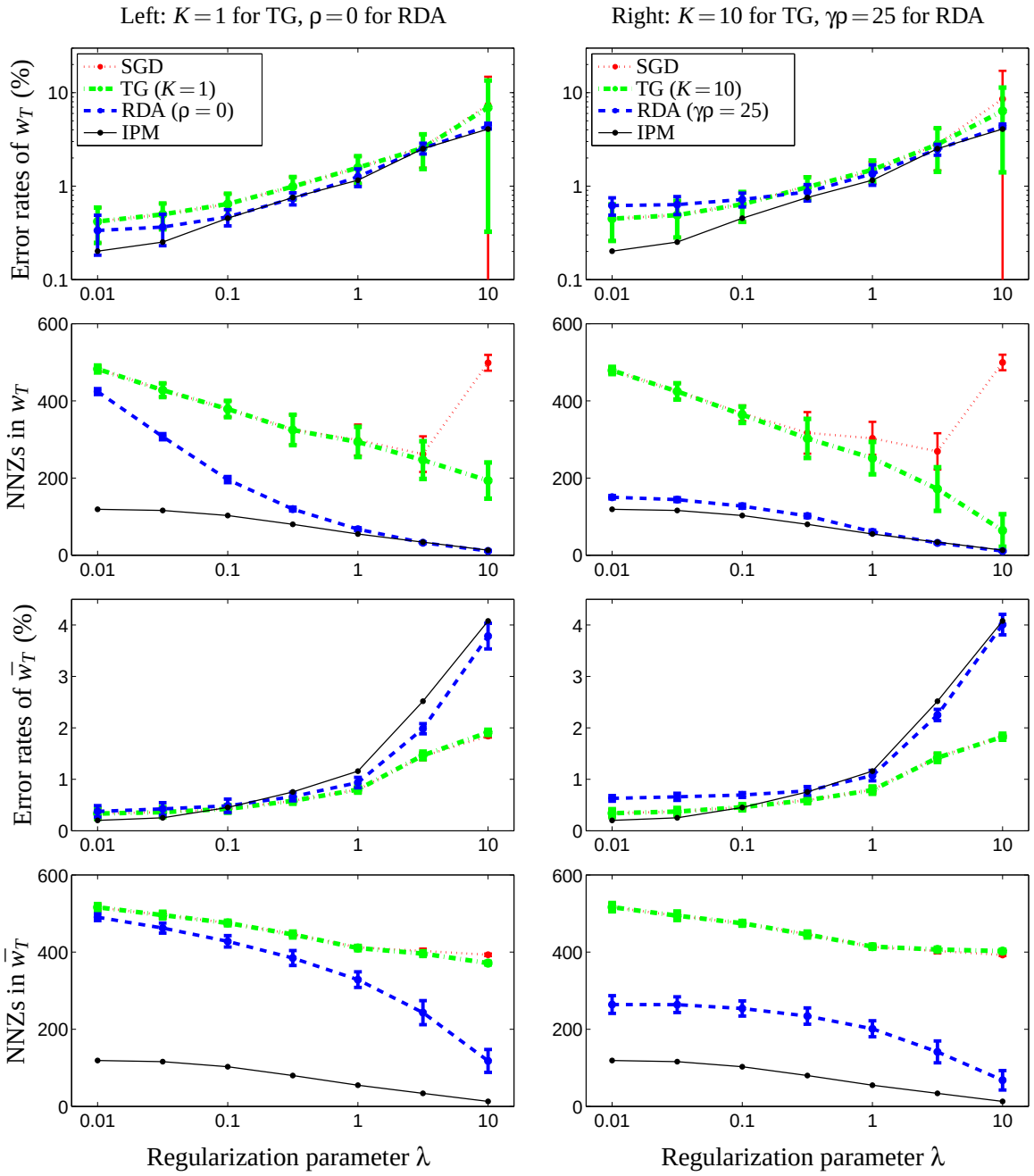

Figure 4: Tradeoffs between testing error rates and NNZs in solutions when varying λ from 0.01 to 10 (for classifying 6 and 7). The left column shows SGD, TG with $K = 1$, RDA with $\rho = 0$, and IPM. The right column shows SGD, TG with $K = 10$, RDA with $\gamma\rho = 25$, and IPM. The same curves for SGD and IPM are plotted in both columns for clear comparison. The top two rows shows the testing error rates and NNZs of the final weights w_T, and the bottom two rows are for the averaged weights $\bar{w}_T$. All horizontal axes have logarithmic scale. For vertical axes, only the two plots in the first row have logarithmic scale.

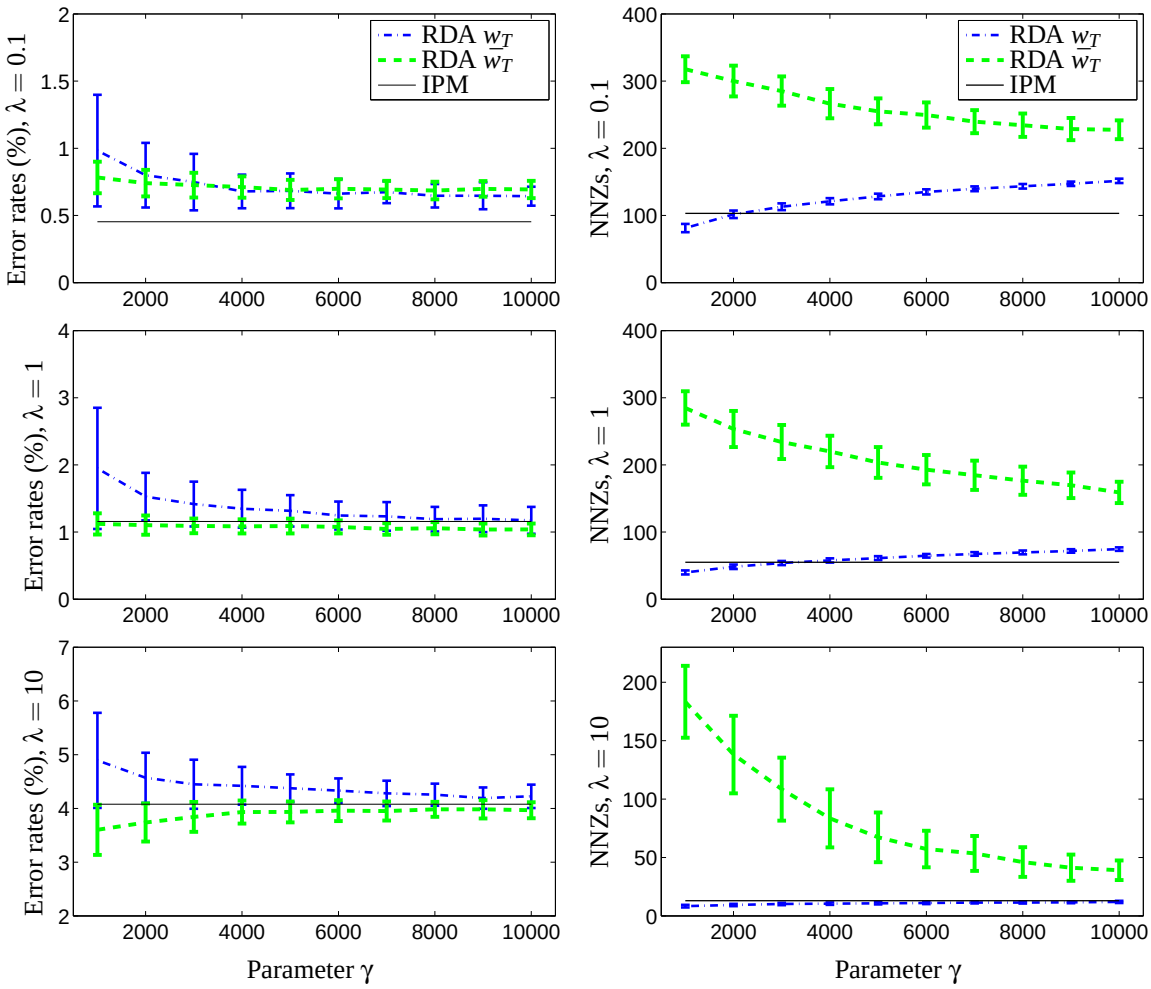

Figure 5: Testing error rates and NNZs in solutions for the RDA method when varying the parameter γ from 1,000 to 10,000, and setting ρ such that $\gamma\rho = 25$. The three rows show results for $\lambda = 0.1$, 1, and 10, respectively. The corresponding batch optimization results found by IPM are shown as a horizontal line in each plot.

To have a better understanding of the behaviors of the algorithms, we plot the number of non-zeros (NNZs) in w_t in Figure 3. Only the RDA method and TG with $K = 1$ give explicit zero weights using soft-thresholding at every step. In order to count the NNZs in all other cases, we have to set a small threshold for rounding the weights to zero. Considering that the magnitudes of the largest weights in Figure 2 are mostly on the order of 10^{-3}, we set 10^{-5} as the threshold and verified that rounding elements less than 10^{-5} to zero does not affect the testing errors. Note that we do not truncate the weights for RDA and TG with $K = 1$ further, even if some of their components are below 10^{-5}. It can be seen that the RDA method maintains a much more sparse w_t than the other online algorithms. While the TG method generates more sparse solutions than the SGD method when λ is large, the NNZs in w_t oscillates with a very big range. The oscillation becomes more severe with $K = 10$. In contrast, the RDA method demonstrates a much more smooth behavior

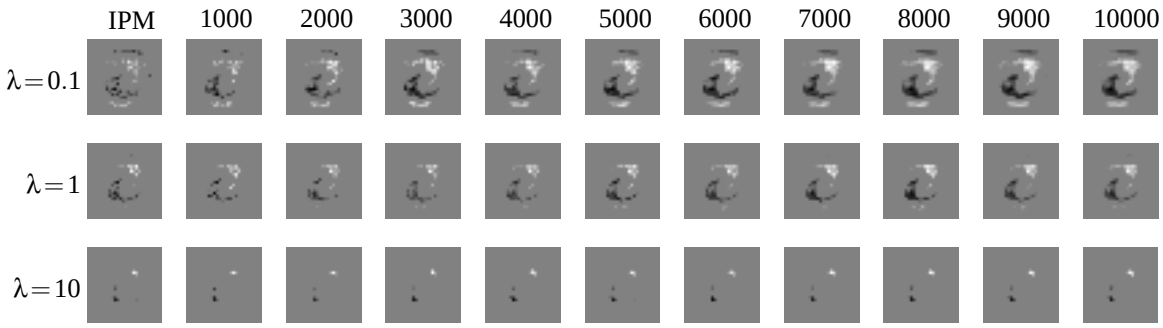

Figure 6: Sparsity patterns of w_T by varying the parameter γ in the RDA method from 1,000 to 10,000 (for classifying the pair 6 and 7). The first column shows results of batch optimization using IPM, and the other 10 columns show results of RDA method using γ labeled at the top.

of the NNZs. For the RDA method, the effect of enhanced regularization using $\gamma\rho = 25$ is more pronounced for relatively small λ.

Next we illustrate the tradeoffs between sparsity and testing error rates. Figure 4 shows that the solutions obtained by the RDA method match the batch optimization results very well. Since the performance of the online algorithms vary when the training data are given in different random sequences (permutations), we run them on 100 randomly permuted sequences of the same training set, and plot the means and standard deviations shown as error bars. For the SGD and TG methods, the testing error rates of w_T vary a lot for different random sequences. In contrast, the RDA method demonstrates very robust performance (small standard deviations) for w_T, even though the theorems only give convergence bound for the averaged weight $\bar{w}_T$. For large values of λ, the averaged weights $\bar{w}_T$ obtained by SGD and TG methods actually have much smaller error rates than those of RDA and batch optimization. This can be explained by the limitation of the SGD and TG methods in obtaining sparse solutions: these lower error rates are obtained with much more nonzero features than used by the RDA and batch optimization methods.

Figure 5 shows the results of choosing different values for the parameter γ in the RDA method. We see that smaller values of γ, which corresponds to faster learning rates, lead to more sparse w_T and higher testing error rates; larger values of γ result in less sparse w_T with lower testing error rates. But interestingly, the effects on the averaged solution $\bar{w}_T$ is almost opposite: smaller values of γ lead to less sparse $\bar{w}_T$ (in this case, we count the NNZs using the rounding threshold 10^{-5}). For large regularization parameter λ, smaller values of γ also give lower testing error rates. Figure 6 shows the sparsity patterns of w_T when varying γ from 1,000 to 10,000. We see that smaller values of γ give more sparse w_T, which are also more scattered like the batch optimization solution by IPM.

Figure 7 shows summary of classification results for all the 45 pairs of digits. For clarity, we only show results of the ℓ_1-RDA method and batch optimization using IPM. We see that the solutions obtained by the ℓ_1-RDA method demonstrate very similar tradeoffs between sparsity and testing error rates as rendered by the batch optimization solutions.

Finally, we note that one of the main reasons for regularization in machine learning is to prevent overfitting, meaning that appropriate amount of regularization may actually reduce the testing error rate. In order to investigate the possibility of overfitting, we also conducted experiments by subsam-

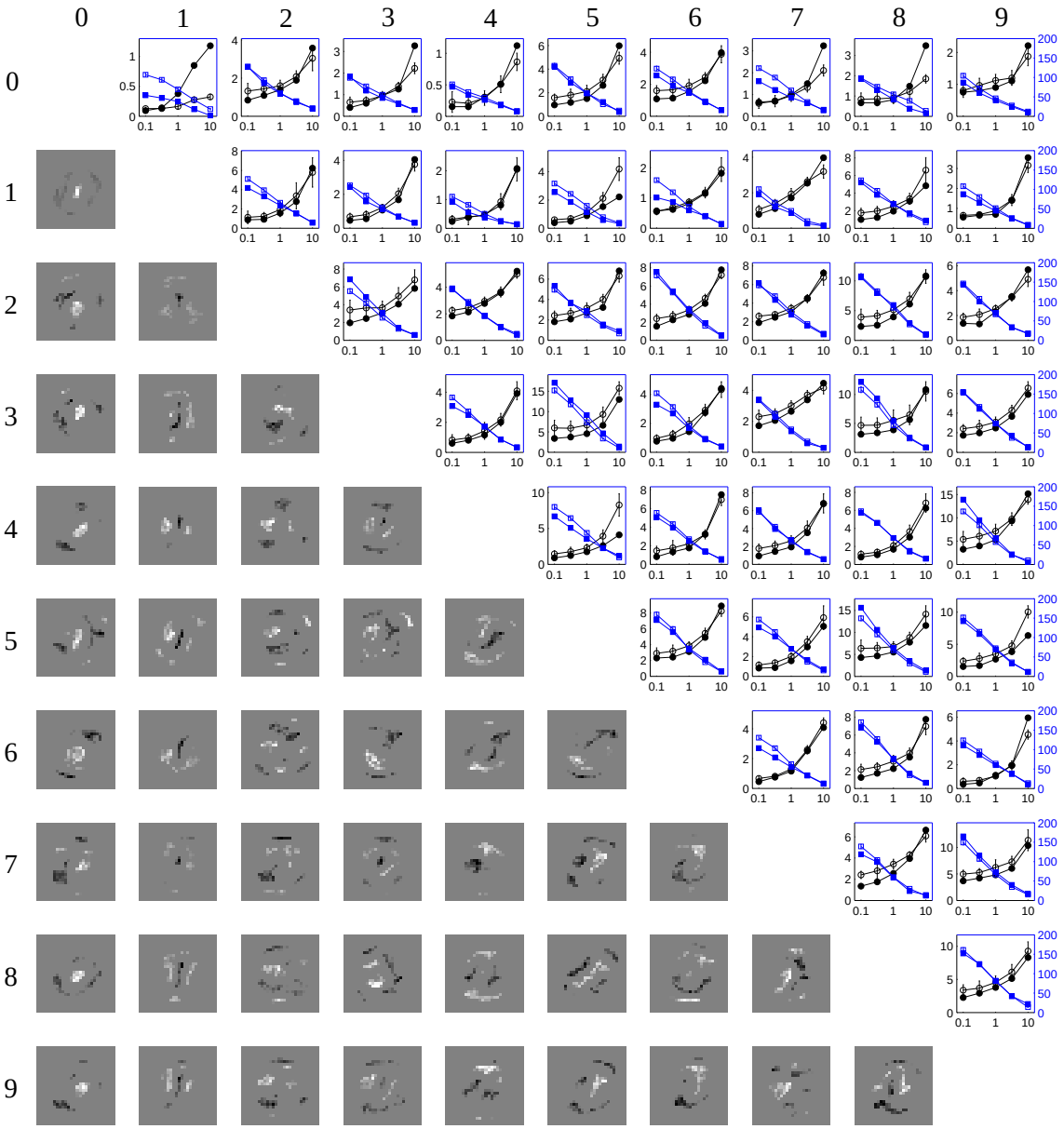

Figure 7: Binary classification for all 45 pairs of digits. The images in the lower-left triangular area show sparsity patterns of w_T with $\lambda = 1$, obtained by the ℓ_1-RDA method with $\gamma = 5000$ and $\rho = 0.005$. The plots in the upper-right triangular area show tradeoffs between sparsity and testing error rates, by varying λ from 0.1 to 10. The solid circles and solid squares show error rates and NNZs in w_T, respectively, using IPM for batch optimization. The hollow circles and hollow squares show error rates and NNZs of w_T, respectively, using the ℓ_1-RDA method. The vertical bars centered at hollow circles and squares show standard deviations by running on 100 different random permutations of the same training data. The scales of the error rates (in percentages) are marked on the left vertical axes, and the scales of the NNZs are marked on the right-most vertical axes.

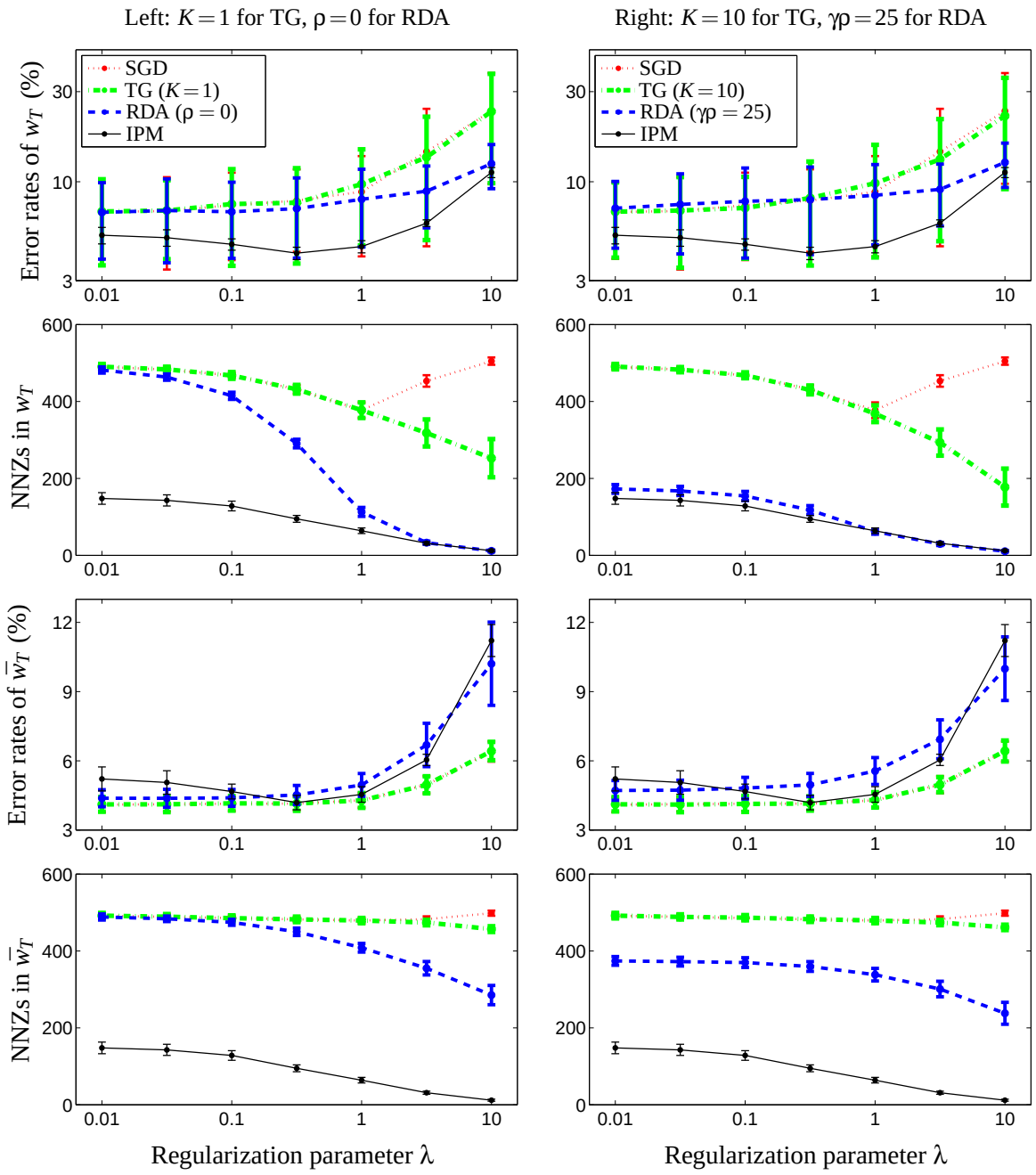

Figure 8: Tradeoffs between testing error rates and NNZs in solutions when varying λ from 0.01 to 10 (for classifying 3 and 8). In order to investigate overfitting, we used 1/10 subsampling of the training data. The error bars show standard deviations of using 10 sets of subsamples. For the three online algorithms, we averaged results on 10 random permutations for each of the 10 subsets. The left column shows SGD, TG with $K=1$, RDA with $\rho=0$, and IPM. The right column shows SGD, TG with $K=10$, RDA with $\gamma\rho=25$, and IPM. The same curves for SGD and IPM are plotted in both columns for clear comparison.

pling the training set. More specifically, we randomly partition the training sets in 10 subsets, and use each subset for training but still test on the whole testing set. The same algorithmic parameters γ and ρ are used as before. Figure 8 shows the results of classifying the more difficult pair 3 and 8. We see that overfitting does occur for batch optimization using IPM. Online algorithms, thanks for their low accuracy in solving the optimization problems, are mostly immune from overfitting.

7. Discussions and Extensions

This paper is inspired by several work in the emerging area of *structural convex optimization* (Nesterov, 2008). The key idea is that by exploiting problem structure that are beyond the conventional black-box model (where only function values and gradient information are allowed), much more efficient first-order methods can be developed for solving structural convex optimization problems. Consider the following problem with two separate parts in the objective function:

$$\underset{w}{\text{minimize}} \quad f(w) + \Psi(w) \tag{32}$$

where the function f is convex and differentiable on $\operatorname{dom}\Psi$, its gradient $\nabla f(w)$ is Lipschitz-continuous with constant L, and the function Ψ is a closed proper convex function. Since Ψ in general can be non-differentiable, the best convergence rate for gradient-type methods that are based on the black-box model is $O(1/\sqrt{t})$ (Nemirovsky and Yudin, 1983). However, if the function Ψ is *simple*, meaning that we are able to find closed-form solution for the auxiliary optimization problem

$$\underset{w}{\text{minimize}} \quad \left\{ f(u) + \langle \nabla f(u), w-u \rangle + \frac{L}{2}\|w-u\|_2^2 + \Psi(w) \right\}, \tag{33}$$

then it is possible to develop accelerated gradient methods that have the convergence rate $O(1/t^2)$ (Nesterov, 1983, 2004; Tseng, 2008; Beck and Teboulle, 2009). Accelerated first-order methods have also been developed for solving large-scale conic optimization problems (Auslender and Teboulle, 2006; Lan et al., 2009; Lu, 2009).

The story is a bit different for stochastic optimization. In this case, the convergence rate $O(1/\sqrt{t})$ cannot be improved in general for convex loss functions with a black-box model. When the loss function $f(w,z)$ have better properties such as differentiability, higher orders of smoothness, and strong convexity, it is tempting to expect that better convergence rates can be achieved. Although these better properties of $f(w,z)$ are inherited by the expected function $\varphi(w) \triangleq \mathbf{E}_z f(w,z)$, almost none of them can really help (Nesterov and Vial, 2008, Section 4). One exception is when the objective function is strongly convex. In this case, the convergence rate for stochastic optimization problems can be improved to $O(\ln t/t)$ (e.g., Nesterov and Vial, 2008), or even $O(1/t)$ (e.g., Polyak and Juditsky, 1992; Nemirovski et al., 2009). For online convex optimization problems, the regret bound can be improved to $O(\ln t)$ (Hazan et al., 2006; Bartlett et al., 2008). But these are still far short of the best complexity result for deterministic optimization with strong convexity assumptions; see, for example, Nesterov (2004, Chapter 2) and Nesterov (2007).

We discuss further the case with a stronger smoothness assumption on the stochastic objective functions. In particular, let $f(w,z)$ be differentiable with respect to w for each z, and the gradient, denoted $g(w,z)$, be Lipschitz continuous. In other words, there exists a constant L such that for any fixed z,

$$\|g(v,z) - g(w,z)\|_* \le L\|v-w\|, \qquad \forall v,w \in \operatorname{dom}\Psi. \tag{34}$$

Let $\varphi(w) = \mathbf{E}_z f(w,z)$. Then φ is differentiable and $\nabla\varphi(w) = \mathbf{E}_z g(w,z)$ (e.g., Rockafellar and Wets, 1982). By Jensen's inequality, $\nabla\varphi(w)$ is also Lipschitz continuous with the same constant L. For the regularization function Ψ, we assume there is a constant G_Ψ such that

$$|\Psi(v) - \Psi(w)| \le G_\Psi \|v - w\|, \qquad \forall v, w \in \operatorname{dom}\Psi.$$

In a black-box model, for any query point w, we are only allowed to query a stochastic gradient $g(w,z)$ and a subgradient of $\Psi(w)$. We assume the stochastic gradients have bounded variance; more specifically, let there be a constant Q such that

$$\mathbf{E}_z \|g(w,z) - \nabla\varphi(w)\|_*^2 \le Q^2, \qquad \forall w \in \operatorname{dom}\Psi. \tag{35}$$

Under these assumptions and the black-box model, the optimal convergence rate for solving the problem (1), according to the complexity theory of Nemirovsky and Yudin (1983), is

$$\mathbf{E}\,\phi(w_t) - \phi^\star \le O(1)\left(\frac{L}{t^2} + \frac{G_\Psi + Q}{\sqrt{t}}\right).$$

Lan (2010) developed an accelerated mirror-descent stochastic approximation method to achieve this rate. The stochastic nature of the algorithm dictates that the term $O(1)(Q/\sqrt{t})$ is inevitable in the convergence bound. However, by using structural optimization techniques similar to (33), it is possible to eliminate the term $O(1)(G_\Psi/\sqrt{t})$ and achieve

$$\mathbf{E}\phi(w_t) - \phi^\star \le O(1)\left(\frac{L}{t^2} + \frac{Q}{\sqrt{t}}\right). \tag{36}$$

Such a result was obtained by Hu et al. (2009). Their algorithm can be viewed as an accelerated version of the FOBOS method (28). In each iteration of their method, the regularization term $\Psi(w)$ is discounted by a factor of $\Theta(t^{-3/2})$. In terms of obtaining the desired regularization effects (see discussions in Section 5), this is even worse than the $\Theta(t^{-1/2})$ discount factor in the FOBOS method. For the case of ℓ_1-regularization, this means using an even smaller truncation threshold $\Theta(t^{-3/2})\lambda$. Next, we give an accelerated version of the RDA method, which achieves the same improved convergence rate (36), but also maintains the desired property of using the undiscounted regularization at each iteration.

7.1 Accelerated RDA Method for Stochastic Optimization

Nesterov (2005) developed an accelerated version of the dual averaging method for solving smooth convex optimization problems, where the uniform average of all past gradients is replaced by an weighted average that emphasizes more recent gradients. Several variations (Nesterov, 2007; Tseng, 2008) were also developed for minimizing composite objective functions of the form (32). They all have a convergence rate $O(L/t^2)$.

Algorithm 3 is our extension of Nesterov's method for solving stochastic optimization problems of the form (1). At the input, it needs a strongly convex function h and two positive sequences $\{\alpha_t\}_{t\ge1}$ and $\{\beta_t\}_{t\ge0}$. At each iteration $t \ge 1$, it computes three primal vectors u_t, v_t, w_t, and a dual vector $\tilde{g}_t$. Among them, u_t is the point for querying a stochastic gradient, $\tilde{g}_t$ is an weighted average of all past stochastic gradients, v_t is the solution of an auxiliary minimization problem that involves $\tilde{g}_t$ and the regularization term $\Psi(w)$, and w_t is the output vector. The computational effort

Algorithm 3 Accelerated RDA method

Input:

- a strongly convex function $h(w)$ with modulus 1 on $\mathrm{dom}\,\Psi$.
- two positive sequences $\{\alpha_t\}_{t\geq 1}$ and $\{\beta_t\}_{t\geq 0}$.

Initialize: set $w_0 = v_0 = \arg\min_w h(w)$, $A_0 = 0$, and $\tilde{g}_0 = 0$.

for $t = 1, 2, 3, \ldots$ **do**

1. Calculate the coefficients
$$A_t = A_{t-1} + \alpha_t, \qquad \theta_t = \frac{\alpha_t}{A_t}.$$
2. Compute the query point
$$u_t = (1-\theta_t) w_{t-1} + \theta_t v_{t-1}.$$
3. Query stochastic gradient $g_t = g(u_t, z_t)$, and update the weighted average $\tilde{g}_t$:
$$\tilde{g}_t = (1-\theta_t)\tilde{g}_{t-1} + \theta_t g_t.$$
4. Solve for the exploration point
$$v_t = \arg\min_w \left\{ \langle \tilde{g}_t, w\rangle + \Psi(w) + \frac{L+\beta_t}{A_t} h(w) \right\}$$
5. Compute w_t by interpolation
$$w_t = (1-\theta_t) w_{t-1} + \theta_t v_t.$$

end for

per iteration is on the same order as Algorithm 1. The additional costs are mainly the two vector interpolations (convex combinations) for computing u_t and w_t. The following theorem gives an estimate of its convergence rate.

Theorem 6 *Assume the conditions (34) and (35) hold, and the problem (1) has an optimal solution $w^\star$ with optimal value $\phi^\star$. In Algorithm 3, if the sequence $\{\alpha_t\}_{t\geq 1}$ and its accumulative sums $A_t = A_{t-1} + \alpha_t$ satisfy the condition $\alpha_t^2 \leq A_t$ for all $t \geq 1$, then*

$$\mathbf{E}\,\phi(w_t) - \phi^\star \leq \frac{L}{A_t} h(w^\star) + \frac{1}{A_t}\left(\beta_t h(w^\star) + Q^2 \sum_{\tau=1}^{t} \frac{\alpha_\tau^2}{2\beta_{\tau-1}} \right).$$

The proof of this theorem is given in Appendix D.

If we choose the two input sequences as

$$\alpha_t = 1, \qquad \forall t \geq 1,$$
$$\beta_t = \gamma\sqrt{t+1}, \qquad \forall t \geq 0,$$

then $A_t = t$, $\theta_t = 1/t$, and $\tilde{g}_t = \bar{g}_t$ is the uniform average of all past gradients. In this case, the minimization problem in Step 4 is very similar to that in Step 3 of Algorithm 1. Let D^2 be an upper

bound on $h(w^\star)$ and set $\gamma = Q/D$. Then we have

$$\mathbf{E}\phi(w_t) - \phi^\star \le \frac{LD^2}{t} + \frac{2QD}{\sqrt{t}}.$$

To achieve the optimal convergence rate stated in (36), we choose

$$\alpha_t = \frac{t}{2}, \qquad \forall t \ge 1, \tag{37}$$

$$\beta_t = \gamma\frac{(t+1)^{3/2}}{2}, \qquad \forall t \ge 0. \tag{38}$$

In this case,

$$A_t = \sum_{\tau=1}^{t}\alpha_\tau = \frac{t(t+1)}{4}, \qquad \theta_t = \frac{\alpha_t}{A_t} = \frac{2}{t+1}, \qquad \forall t \ge 1.$$

It is easy to verify that the condition $\alpha_t^2 \le A_t$ is satisfied. The following corollary is proved in Appendix D.1.

Corollary 7 *Assume the conditions (34) and (35) hold, and $h(w^\star) \le D^2$. If the two input sequences in Algorithm 3 are chosen as in (37) and (38) with $\gamma = Q/D$, then*

$$\mathbf{E}\phi(w_t) - \phi^\star \le \frac{4LD^2}{t^2} + \frac{4QD}{\sqrt{t}}.$$

We can also give high probability bound under more restrictive assumptions. Instead of requiring the deterministic condition $\|g(w,z) - \nabla\varphi(w)\|_*^2 \le Q^2$ for all z and all $w \in \operatorname{dom}\Psi$, we adopt a weaker condition used in Nemirovski et al. (2009) and Lan (2010):

$$\mathbf{E}\left[\exp\left(\frac{\|g(w,z) - \nabla\varphi(w)\|_*^2}{Q^2}\right)\right] \le \exp(1), \qquad \forall w \in \operatorname{dom}\Psi. \tag{39}$$

It is not hard to see that this implies (35) by using Jensen's inequality.

Theorem 8 *Suppose $\operatorname{dom}\Psi$ is compact, say $h(w) \le D^2$ for all $w \in \operatorname{dom}\Psi$, and let the assumptions (34) and (39) hold. If the two input sequences in Algorithm 3 are chosen as in (37) and (38) with $\gamma = Q/D$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\phi(w_t) - \phi^\star \le \frac{4LD^2}{t^2} + \frac{4QD}{\sqrt{t}} + \frac{QD}{\sqrt{t}}\left(\ln(2/\delta) + 2\sqrt{\ln(2/\delta)}\right)$$

Compared with the bound on expectation, the additional penalty in the high probability bound depends only on Q, not L. This theorem is proved in Appendix D.2.

In the special case of deterministic optimization, that is, when $Q = 0$, we have $\gamma = Q/D = 0$ and $\beta_t = 0$ for all $t \ge 0$. Then Algorithm 3 reduces to a variant of Nesterov's method given in Tseng (2008, Section 4), which has convergence rate $\phi(w_t) - \phi^\star \le 4LD^2/t^2$.

For stochastic optimization problems, the above theoretical bounds show that the algorithm can be very effective when Q is much smaller than LD. One way to make this happen is to use a mini-batch approach. More specifically, at each iteration of Algorithm 3, let g_t itself be the average of the stochastic gradients at a small batch of samples computed at u_t. We leave the empirical studies of Algorithm 3 and other accelerated schemes for future investigation.

7.2 The p-Norm RDA Methods

The p-norm RDA methods are special cases of Algorithm 1 in which the auxiliary functions h (*not* the regularization functions Ψ) are squared p-norms. They offer more flexibility than 2-norm based RDA methods in adapting to the geometry of the learning problems.

Recall that for $p \geq 1$, the p-norm of $w \in \mathbf{R}^n$ is defined as $\|w\|_p = \left(\sum_{i=1}^n |w^{(i)}|^p\right)^{1/p}$. If p and q satisfy the equality $1/p + 1/q = 1$, then the norms $\|w\|_p$ and $\|g\|_q$ are dual to each other. Moreover, the pair of functions $(1/2)\|w\|_p^2$ and $(1/2)\|g\|_q^2$ are conjugate functions of each other. As a result, their gradient mappings are a pair of inverse mappings. More formally, let $p \in (1,2]$ and $q = p/(p-1)$, and define the mapping $\vartheta : \mathcal{E} \to \mathcal{E}^*$ with

$$\vartheta_i(w) = \nabla_i \left(\frac{1}{2}\|w\|_p^2\right) = \frac{\operatorname{sgn}\left(w^{(i)}\right)\left|w^{(i)}\right|^{p-1}}{\|w\|_p^{p-2}}, \qquad i = 1, \ldots, n,$$

and the inverse mapping $\vartheta^{-1} : \mathcal{E}^* \to \mathcal{E}$ with

$$\vartheta_i^{-1}(g) = \nabla_i \left(\frac{1}{2}\|g\|_q^2\right) = \frac{\operatorname{sgn}\left(g^{(i)}\right)\left|g^{(i)}\right|^{q-1}}{\|g\|_q^{q-2}}, \qquad i = 1, \ldots, n.$$

These mappings are often called *link functions* in machine learning (e.g., Gentile, 2003).

Again we focus on the ℓ_1-RDA case with $\Psi(w) = \lambda\|w\|_1$. For any $p \in (1,2]$, the function $(1/2)\|w\|_p^2$ is strongly convex with respect to $\|\cdot\|_p$ with the convexity parameter $p-1$ (e.g., Juditsky and Nemirovski, 2008). In order to have an auxiliary strongly convex function h with convexity parameter 1, we define

$$h(w) = \frac{1}{2(p-1)}\|w\|_p^2.$$

Using $\beta_t = \gamma\sqrt{t}$ for some $\gamma > 0$,, the Equation (8) in Algorithm 1 becomes

$$w_{t+1} = \arg\min_w \left\{ \langle \bar{g}_t, w \rangle + \lambda\|w\|_1 + \frac{\gamma}{\sqrt{t}} \frac{1}{2(p-1)}\|w\|_p^2 \right\}.$$

The optimality condition of the above minimization problem (Rockafellar, 1970, Section 27) states that there exists a subgradient $s \in \partial\|w_{t+1}\|_1$ such that

$$\bar{g}_t + \lambda s + \frac{\gamma}{(p-1)\sqrt{t}}\vartheta(w_{t+1}) = 0.$$

Following similar arguments as in Appendix A, we find that it has a closed-form solution

$$w_{t+1} = \vartheta^{-1}(\hat{g}_t),$$

where the elements of $\hat{g}_t$ are given as

$$\hat{g}_t^{(i)} = \begin{cases} 0 & \text{if } \left|\bar{g}_t^{(i)}\right| \leq \lambda, \\ -\dfrac{(p-1)\sqrt{t}}{\gamma}\left(\bar{g}_t^{(i)} - \lambda \operatorname{sgn}\left(\bar{g}_t^{(i)}\right)\right) & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n.$$

When $p = q = 2$, ϑ and ϑ^{-1} are identity maps and the solution is the same as (10). If p is close to 1 ($q \gg 2$), the map ϑ^{-1} penalizes small entries of the truncated vector $\hat{g}_t$ to be even smaller.

As an interesting property of the map ϑ^{-1}, we always have $\|w_{t+1}\|_p = \|\hat{g}_t\|_q$ (e.g., Gentile, 2003, Lemma 1).

In terms of regret bound or convergence rate, our results in Sections 3 and 4 apply directly. More specifically, for stochastic learning problems, let $D_p^2 = (1/2(p-1))\|w^\star\|_p^2$, and G_q be an upper bound on $\|g_t\|_q$ for all $t \geq 1$. Then by Corollary 2 and Theorem 3,

$$\mathbf{E}\,\phi(\bar{w}_t) - \phi^\star \leq \left(\gamma D_p^2 + \frac{G_q^2}{\gamma}\right)\frac{1}{\sqrt{t}}.$$

The optimal choice of γ is $\gamma^\star = G_q/D_p$, which results in

$$\mathbf{E}\,\phi(\bar{w}_t) - \phi^\star \leq \sqrt{\frac{2}{p-1}}\frac{G_q\|w^\star\|_p}{\sqrt{t}} = \sqrt{2(q-1)}\frac{G_q\|w^\star\|_p}{\sqrt{t}}.$$

In order to gain further insight, we transform the convergence bound in terms of ℓ_∞ and ℓ_1 norms. Let G_∞ be an upper bound on $\|g_t\|_\infty$, that is,

$$\left|g_t^{(i)}\right| \leq G_\infty, \qquad \forall\, i = 1,\ldots,n, \qquad \forall\, t \geq 1.$$

Then $\|g_t\|_q \leq G_\infty n^{1/q}$. If we choose $q = \ln n$ (assuming $n \geq e^2$ so that $q \geq 2$), then $\|g_t\|_q \leq G_\infty n^{1/\ln n} = G_\infty e$. Next we substitute $G_\infty e$ for G_q and use the fact $\|w^\star\|_p \leq \|w^\star\|_1$, then

$$\mathbf{E}\,\phi(\bar{w}_t) - \phi^\star \leq \sqrt{2(\ln n - 1)}\,\frac{eG_\infty\|w^\star\|_1}{\sqrt{t}} = O\left(\frac{\sqrt{\ln n}\,G_\infty\|w^\star\|_1}{\sqrt{t}}\right).$$

For 2-norm based RDA method, we have $\|g_t\|_2 \leq G_\infty\sqrt{n}$, thus

$$\mathbf{E}\,\phi(\bar{w}_t) - \phi^\star \leq \frac{\sqrt{2n}\,G_\infty\|w^\star\|_2}{\sqrt{t}}.$$

Therefore, for learning problems in which the features are dense (i.e., G_2 close to $G_\infty\sqrt{n}$) and $w^\star$ is indeed very sparse (i.e., $\|w^\star\|_2$ close to $\|w^\star\|_1$), using the p-norm RDA method, with $p = \ln n/(\ln n - 1)$, can lead to faster convergence.

The above analysis of convergence rates matches that for the p-norm based SMIDAS (Stochastic MIrror Descent Algorithm made Sparse) algorithm developed in Shalev-Shwartz and Tewari (2009). However, like other algorithms of the mirror-descent type, including TG (Langford et al., 2009) and FOBOS (Duchi and Singer, 2009), SMIDAS uses a truncation threshold $\Theta(1/\sqrt{t})\lambda$ in obtaining sparse solutions. In contrast, the p-norm based RDA method uses a much more aggressive threshold λ. This is their major difference.

The accelerated RDA method (Algorithm 3) also works in the p-norm setting.

7.3 Connection with Incremental Subgradient Methods

As an intermediate model between deterministic and stochastic optimization problems, consider the problem

$$\operatorname*{minimize}_w \quad \sum_{k=1}^m f_k(w) + \Psi(w), \tag{40}$$

which can be considered as a special case of (1) where the random variable z has a uniform distribution on a finite support; more specifically, $f_k(w) = (1/m) f(w, z_k)$ for $k = 1, \ldots, m$. The unregularized version, that is, with $\Psi(w) = 0$, has been addressed by *incremental subgradient methods* (e.g., Tseng, 1998; Nedić and Bertsekas, 2001). At each iteration of such methods, a step is taken along the negative subgradient of a single function f_k, which is chosen either in a round-robin manner or randomly with uniform distribution. The randomized version is equivalent to the SGD method. The RDA methods are well suited for solving the regularized version (40).

Randomized incremental subgradient methods with Markov jumps have also been developed for solving (40) with $\Psi(w) = 0$ (Johansson et al., 2009; Ram et al., 2009). In such methods, the functions f_k are picked randomly but not independently: they follow the transition probabilities of a Markov chain that has the uniform distribution. It would be very interesting to investigate the convergence of the RDA methods when the random examples are drawn according to a Markovian chain. This is particularly attractive for online learning problems where the assumption of i.i.d. samples does not hold.

Acknowledgments

The author is grateful to John Duchi for careful reading of a previous draft and pointing out that the regret analysis for general convex regularizations and for strongly convex regularizations can be unified. The author thanks John Platt for encouragement and suggestions for improving the computational experiments, and Paul Tseng and Zhaosong Lu for helpful discussions on minimizing composite objective functions. He also would like to thank the anonymous referees for their valuable comments.

Appendix A. Closed-form Solution for ℓ_1-RDA Method

For RDA method with ℓ_1-regularization, we set $\Psi(w) = \lambda \|w\|_1$ and use $h(w) = (1/2)\|w\|_2^2$, or use $h_\rho(w)$ in (31) for enhanced regularization effect. In such cases, the minimization problem in step 3 of Algorithm 1 can be decomposed into n independent scalar minimization problems, each of the form

$$\underset{\omega \in \mathbf{R}}{\text{minimize}} \quad \eta_t \omega + \lambda_t |\omega| + \frac{\gamma_t}{2} \omega^2,$$

where the coefficients $\lambda_t > 0$, $\gamma_t > 0$, and η_t can be arbitrary. This is an unconstrained nonsmooth optimization problem. Its optimality condition (Rockafellar, 1970, Section 27) states that $\omega^\star$ is an optimal solution if and only if there exists $\xi \in \partial |\omega^\star|$ such that

$$\eta_t + \lambda_t \xi + \gamma_t \omega^\star = 0. \tag{41}$$

The subdifferential of $|\omega|$ is

$$\partial |\omega| = \begin{cases} \{\xi \in \mathbf{R} \mid -1 \le \xi \le 1\} & \text{if } \omega = 0, \\ \{1\} & \text{if } \omega > 0, \\ \{-1\} & \text{if } \omega < 0. \end{cases}$$

We discuss the solution to (41) in three different cases:

- If $|\eta_t| \leq \lambda_t$, then $\omega^\star = 0$ and $\xi = -\eta_t/\lambda_t \in \partial|0|$ satisfy (41). We also show that there is no solution other than $\omega^\star = 0$. If $\omega > 0$, then $\xi = 1$, and

$$\eta_t + \lambda_t + \gamma_t \omega > \eta_t + \lambda_t \geq 0.$$

  Similarly, if $\omega < 0$, then $\xi = -1$, and

$$\eta_t - \lambda_t + \gamma_t \omega < \eta_t - \lambda_t \leq 0.$$

  In either cases when $\omega \neq 0$, the optimality condition (41) cannot be satisfied.

- If $\eta_t > \lambda_t > 0$, we must have $\omega^\star < 0$ and $\xi = -1$. More specifically,

$$\omega^\star = -\frac{1}{\gamma_t}(\eta_t - \lambda_t).$$

- If $\eta_t < -\lambda_t < 0$, we must have $\omega^\star > 0$ and $\xi = 1$. More specifically,

$$\omega^\star = -\frac{1}{\gamma_t}(\eta_t + \lambda_t).$$

The above discussions can be summarized as

$$\omega^\star = \begin{cases} 0 & \text{if } |\eta_t| \leq \lambda_t, \\ -\dfrac{1}{\gamma_t}\big(\eta_t - \lambda_t \operatorname{sgn}(\eta_t)\big) & \text{otherwise.} \end{cases}$$

This is the closed-form solution for each component of w_{t+1} in the ℓ_1-RDA method.

Appendix B. Regret Analysis of RDA Method

In this Appendix, we prove Theorem 1. First, let s_t denote the sum of the subgradients obtained up to time t in the RDA method, that is,

$$s_t = \sum_{\tau=1}^{t} g_\tau = t\bar{g}_t, \tag{42}$$

with the initialization $s_0 = 0$. Then the Equation (8) in Algorithm 1 is equivalent to

$$w_{t+1} = \arg\min_w \big\{ \langle s_t, w \rangle + t\Psi(w) + \beta_t h(w) \big\}. \tag{43}$$

This extends the *simple dual averaging* scheme of Nesterov (2009), where $\Psi(w)$ reduces to the indicator function of a closed convex set. Compared with the analysis in Nesterov (2009), the assumption (7), Lemma 11 and Lemma 12 (below) are new essentials that make the proof work. We also provide refined bounds on the primal and dual variables that relate to the regret with respect to an arbitrary comparison point; see part (b) and (c) of Theorem 1. It seems that the *weighted dual averaging* scheme of Nesterov (2009) cannot be extended when Ψ is a nontrivial regularization function.

B.1 Conjugate Functions and Their Properties

Let w_0 be the unique minimizer of $h(w)$. By the assumption (7), we have

$$w_0 = \arg\min_w h(w) \in \operatorname{Arg}\min_w \Psi(w).$$

Let $\{\beta_t\}_{t\geq 1}$ be the input sequence to Algorithm 1, which is nonnegative and nondecreasing. In accordance with the assumption (13), we let

$$\beta_0 = \max\{\sigma, \beta_1\} > 0, \tag{44}$$

where σ be the convexity parameter of $\Psi(w)$. For each $t \geq 0$, we define two conjugate-type functions:

$$U_t(s) = \max_{w\in\mathcal{F}_D} \big\{\langle s, w - w_0\rangle - t\Psi(w)\big\}, \tag{45}$$

$$V_t(s) = \max_{w} \big\{\langle s, w - w_0\rangle - t\Psi(w) - \beta_t h(w)\big\}, \tag{46}$$

where $\mathcal{F}_D = \big\{w \in \operatorname{dom}\Psi \,|\, h(w) \leq D^2\big\}$. The maximum in (45) is always achieved because $\mathcal{F}_D$ is a nonempty compact set (which always contains w_0). Because of (44), we have $\sigma t + \beta_t \geq \beta_0 > 0$ for all $t \geq 0$, which means the functions $t\Psi(w) + \beta_t h(w)$ are all strongly convex. Therefore, the maximum in (46) is always achieved, and the maximizer is unique. As a result, we have $\operatorname{dom} U_t = \operatorname{dom} V_t = \mathcal{E}^*$ for all $t \geq 0$. Moreover, by the assumption $\Psi(w_0) = h(w_0) = 0$, both of the functions are nonnegative.

The lemma below is similar to Lemma 2 of Nesterov (2009), but with our new definitions of U_t and V_t. We include the proof for completeness.

Lemma 9 *For any $s \in \mathcal{E}^*$ and $t \geq 0$, we have*

$$U_t(s) \leq V_t(s) + \beta_t D^2.$$

Proof Starting with the definition of $U_t(s)$ and using $\mathcal{F}_D = \{w \in \operatorname{dom}\Psi \,|\, h(w) \leq D^2\}$,

$$\begin{aligned}
U_t(s) &= \max_{w\in\mathcal{F}_D} \{\langle s, w - w_0\rangle - t\Psi(w)\} \\
&= \max_w \min_{\beta\geq 0} \big\{\langle s, w - w_0\rangle - t\Psi(w) + \beta(D^2 - h(w))\big\} \\
&\leq \min_{\beta\geq 0} \max_w \big\{\langle s, w - w_0\rangle - t\Psi(w) + \beta(D^2 - h(w))\big\} \\
&\leq \max_w \big\{\langle s, w - w_0\rangle - t\Psi(w) + \beta_t(D^2 - h(w))\big\} \\
&= V_t(s) + \beta_t D^2.
\end{aligned}$$

For the second equality and the first inequality above, we used standard duality arguments and the max-min inequality; see, for example, Boyd and Vandenberghe (2004, Section 5.4.1). ■

Let $\pi_t(s)$ denote the unique maximizer in the definition of $V_t(s)$; in other words,

$$\begin{aligned}
\pi_t(s) &= \arg\max_w \big\{\langle s, w - w_0\rangle - t\Psi(w) - \beta_t h(w)\big\} \\
&= \arg\min_w \big\{\langle -s, w\rangle + t\Psi(w) + \beta_t h(w)\big\}.
\end{aligned}$$

Comparing with the Equation (43), we have

$$w_{t+1} = \pi_t(-s_t), \qquad \forall t \geq 0.$$

Lemma 10 *The function V_t is convex and differentiable. Its gradient is given by*

$$\nabla V_t(s) = \pi_t(s) - w_0 \tag{47}$$

Moreover, the gradient is Lipschitz continuous with constant $1/(\sigma t + \beta_t)$; that is

$$\|\nabla V_t(s_1) - \nabla V_t(s_2)\| \leq \frac{1}{\sigma t + \beta_t}\|s_1 - s_2\|_*, \qquad \forall s_1, s_2 \in \mathcal{E}^*.$$

Proof Because the function $t\Psi(w) + \beta_t h(w)$ is a strongly convex with convexity parameter $\sigma t + \beta_t$, this lemma follows from classical results in convex analysis; see, for example, Hiriart-Urruty and Lemaréchal (2001, Chapter E, Theorem 4.2.1), or Nesterov (2005, Theorem 1). ■

A direct consequence of Lemma 10 is the following inequality:

$$V_t(s+g) \leq V_t(s) + \langle g, \nabla V_t(s)\rangle + \frac{1}{2(\sigma t + \beta_t)}\|g\|_*^2, \quad \forall s, g \in \mathcal{E}^*. \tag{48}$$

For a proof, see, for example, Nesterov (2004, Theorem 2.1.5).

Lemma 11 *For each $t \geq 1$, we have*

$$V_t(-s_t) + \Psi(w_{t+1}) \leq V_{t-1}(-s_t) + (\beta_{t-1} - \beta_t)h(w_{t+1}).$$

Proof We start with the definition of $V_{t-1}(-s_t)$:

$$\begin{aligned} V_{t-1}(-s_t) &= \max_w \{\langle -s_t, w - w_0\rangle - (t-1)\Psi(w) - \beta_{t-1}h(w)\} \\ &\geq \langle -s_t, w_{t+1} - w_0\rangle - (t-1)\Psi(w_{t+1}) - \beta_{t-1}h(w_{t+1}) \\ &= \left\{\langle -s_t, w_{t+1} - w_0\rangle - t\Psi(w_{t+1}) - \beta_t h(w_{t+1})\right\} + \Psi(w_{t+1}) + (\beta_t - \beta_{t-1})h(w_{t+1}). \end{aligned}$$

Comparing with (43) and (46), we recognize that the expression in the last braces above is precisely $V_t(-s_t)$. Making the substitution and rearranging terms give the desired result. ■

Since by assumption $h(w_{t+1}) \geq 0$ and the sequence $\{\beta_t\}_{t\geq 1}$ is nondecreasing, we have

$$V_t(-s_t) + \Psi(w_{t+1}) \leq V_{t-1}(-s_t), \quad \forall t \geq 2. \tag{49}$$

For $t = 1$, Lemma (11) gives

$$V_1(-s_1) + \Psi(w_2) \leq V_0(-s_1) + (\beta_0 - \beta_1)h(w_2). \tag{50}$$

Since it may happen that $\beta_0 > \beta_1$, we need the following upper bound on $h(w_2)$.

Lemma 12 *Assume* $\max\{\sigma, \beta_1\} > 0$, *and let* $h(w) = (1/\sigma)\Psi(w)$ *if* $\sigma > 0$. *Then*

$$h(w_2) \leq \frac{2\|g_1\|_*^2}{(\beta_1+\sigma)^2}. \tag{51}$$

Proof For $t = 1$, we have $w_1 = w_0$, $\Psi(w_1) = \Psi(w_0) = 0$, $h(w_1) = h(w_0) = 0$, and $\bar{g}_1 = g_1$. Since w_2 is the minimizer in (43) for $t = 1$, we have

$$\langle g_1, w_2\rangle + \Psi(w_2) + \beta_1 h(w_2) \leq \langle g_1, w_1\rangle + \Psi(w_1) + \beta_1 h(w_1) = \langle g_1, w_1\rangle.$$

Therefore,

$$\Psi(w_2) + \beta_1 h(w_2) \leq \langle g_1, w_1 - w_2\rangle \leq \|g_1\|_*\|w_2 - w_1\|.$$

On the other hand, by strong convexity of $\Psi(w)$ and $h(w)$, we have

$$\Psi(w_2) + \beta_1 h(w_2) \geq \frac{\sigma+\beta_1}{2}\|w_2 - w_1\|^2.$$

Combining the last two inequalities together, we have

$$\Psi(w_2) + \beta_1 h(w_2) \leq \frac{2\|g_1\|_*^2}{\sigma+\beta_1}.$$

By assumption, if $\sigma = 0$, we must have $\beta_1 > 0$. In this case, since $\Psi(w_2) \geq 0$, we have

$$\beta_1 h(w_2) \leq \Psi(w_2) + \beta_1 h(w_2) \leq \frac{2\|g_1\|_*^2}{\beta_1} \quad \Longrightarrow \quad h(w_2) \leq \frac{2\|g_1\|_*^2}{\beta_1^2} = \frac{2\|g_1\|_*^2}{(\sigma+\beta_1)^2}.$$

If $\sigma > 0$, we have $\Psi(w) = \sigma h(w)$ by assumption, and therefore

$$\Psi(w_2) + \beta_1 h(w_2) = (\sigma+\beta_1)h(w_2) \leq \frac{2\|g_1\|_*^2}{\sigma+\beta_1},$$

which also results in (51). ■

B.2 Bounding the Regret

To measure the quality of the solutions $w_1, \ldots, w_t$, we define the following *gap* sequence:

$$\delta_t = \max_{w\in\mathcal{F}_D}\left\{\sum_{\tau=1}^{t}\big(\langle g_\tau, w_\tau - w\rangle + \Psi(w_\tau)\big) - t\Psi(w)\right\}, \quad t = 1, 2, 3, \ldots. \tag{52}$$

The gap δ_t is an upper bound on the regret $R_t(w)$ for all $w \in \mathcal{F}_D$. To see this, we use the assumption $w \in \mathcal{F}_D$ and convexity of $f_t(w)$ in the following:

$$\begin{aligned}
\delta_t &\geq \sum_{\tau=1}^{t}\big(\langle g_\tau, w_\tau - w\rangle + \Psi(w_\tau)\big) - t\Psi(w) \\
&\geq \sum_{\tau=1}^{t}\big(f_\tau(w_\tau) - f_\tau(w) + \Psi(w_\tau)\big) - t\Psi(w) \\
&= \sum_{\tau=1}^{t}\big(f_\tau(w_\tau) + \Psi(w_\tau)\big) - \sum_{\tau=1}^{t}\big(f_\tau(w) + \Psi(w)\big) = R_t(w).
\end{aligned} \tag{53}$$

We can also derive an upper bound on δ_t. For this purpose, we add and subtract the sum $\sum_{\tau=1}^t \langle g_\tau, w_0\rangle$ in the definition (52), which leads to

$$\delta_t = \sum_{\tau=1}^t \big(\langle g_\tau, w_\tau - w_0\rangle + \Psi(w_\tau)\big) + \max_{w\in\mathcal{F}_D} \big\{\langle s_t, w_0 - w\rangle - t\Psi(w)\big\}. \tag{54}$$

We observe that the maximization term in (54) is in fact $U_t(-s_t)$. Therefore, by applying Lemma 9, we have

$$\delta_t \le \sum_{\tau=1}^t \big(\langle g_\tau, w_\tau - w_0\rangle + \Psi(w_\tau)\big) + V_t(-s_t) + \beta_t D^2. \tag{55}$$

Next, we show that Δ_t defined in (14) is an upper bound for the right-hand side of the inequality (55). For any $\tau \ge 2$, we have

$$\begin{aligned} V_\tau(-s_\tau) + \Psi(w_{\tau+1}) &\le V_{\tau-1}(-s_\tau) \\ &= V_{\tau-1}(-s_{\tau-1} - g_\tau) \\ &\le V_{\tau-1}(-s_{\tau-1}) + \langle -g_\tau, \nabla V_{\tau-1}(-s_{\tau-1})\rangle + \frac{\|g_\tau\|_*^2}{2(\sigma(\tau-1)+\beta_{\tau-1})} \\ &= V_{\tau-1}(-s_{\tau-1}) + \langle -g_\tau, w_\tau - w_0\rangle + \frac{\|g_\tau\|_*^2}{2(\sigma(\tau-1)+\beta_{\tau-1})}, \end{aligned}$$

where the four steps above used (49), (42), (48), and (47), respectively. Therefore,

$$\langle g_\tau, w_\tau - w_0\rangle + \Psi(w_{\tau+1}) \le V_{\tau-1}(-s_{\tau-1}) - V_\tau(-s_\tau) + \frac{\|g_\tau\|_*^2}{2(\sigma(\tau-1)+\beta_{\tau-1})}, \quad \forall \tau \ge 2.$$

For $\tau = 1$, we have a similar inequality

$$\langle g_1, w_1 - w_0\rangle + \Psi(w_2) \le V_0(-s_0) - V_1(-s_1) + \frac{\|g_1\|_*^2}{2\beta_0} + (\beta_0 - \beta_1)h(w_2),$$

where the additional term $(\beta_0 - \beta_1)h(w_2)$ comes from using (50). Summing the above inequalities for $\tau = 1, \ldots, t$, and noting that $V_0(-s_0) = V_0(0) = 0$, we arrive at

$$\sum_{\tau=1}^t \big(\langle g_\tau, w_\tau - w_0\rangle + \Psi(w_{\tau+1})\big) + V_t(-s_t) \le (\beta_0 - \beta_1)h(w_2) + \frac{1}{2}\sum_{\tau=1}^t \frac{\|g_\tau\|_*^2}{\sigma(\tau-1)+\beta_{\tau-1}}.$$

Using $w_1 = w_0 \in \operatorname{Arg\,min}_w \Psi(w)$, we have $\Psi(w_{t+1}) \ge \Psi(w_0) = \Psi(w_1)$. Therefore, adding the nonpositive quantity $\Psi(w_1) - \Psi(w_{t+1})$ to the left-hand side of the above inequality yields

$$\sum_{\tau=1}^t \big(\langle g_\tau, w_\tau - w_0\rangle + \Psi(w_\tau)\big) + V_t(-s_t) \le (\beta_0 - \beta_1)h(w_2) + \frac{1}{2}\sum_{\tau=1}^t \frac{\|g_\tau\|_*^2}{\sigma(\tau-1)+\beta_{\tau-1}}. \tag{56}$$

Combining the inequalities (53), (55) and (56), and using Lemma 12,

$$R_t(w) \le \delta_t \le \beta_t D^2 + \frac{1}{2}\sum_{\tau=1}^t \frac{\|g_\tau\|_*^2}{\sigma(\tau-1)+\beta_{\tau-1}} + \frac{2(\beta_0-\beta_1)\|g_1\|_*^2}{(\beta_1+\sigma)^2}.$$

Finally using the assumption (12) and the definition of Δ_t in (14), we conclude

$$R_t(w) \le \delta_t \le \Delta_t.$$

This proves the regret bound (15).

B.3 Bounding the Primal Variable

We start with the optimality condition for the minimization problem in (43): there exist subgradients $b_{t+1} \in \partial\Psi(w_{t+1})$ and $d_{t+1} \in \partial h(w_{t+1})$ such that

$$\langle s_t + t b_{t+1} + \beta_t d_{t+1},\ w - w_{t+1} \rangle \geq 0, \quad \forall w \in \operatorname{dom}\Psi. \tag{57}$$

By the strong convexity of h and Ψ, we have for any $w \in \operatorname{dom}\Psi$,

$$\Psi(w) \geq \Psi(w_{t+1}) + \langle b_{t+1}, w - w_{t+1} \rangle + \frac{\sigma}{2}\|w_{t+1} - w\|^2, \tag{58}$$

$$h(w) \geq h(w_{t+1}) + \langle d_{t+1}, w - w_{t+1} \rangle + \frac{1}{2}\|w_{t+1} - w\|^2. \tag{59}$$

We multiply both sides of the inequality (58) by t, multiply both sides of the inequality (59) by β_t, and then add them together. This gives

$$\begin{aligned}\frac{1}{2}(\sigma t + \beta_t)\|w_{t+1} - w\|^2 \leq{}& \beta_t h(w) - \beta_t h(w_{t+1}) - \langle t b_{t+1} + \beta_t d_{t+1}, w - w_{t+1} \rangle \\ &+ t\Psi(w) - t\Psi(w_{t+1}).\end{aligned}$$

Using the optimality condition (57), we have

$$\begin{aligned}\frac{1}{2}(\sigma t + \beta_t)\|w_{t+1} - w\|^2 &\leq \beta_t h(w) - \beta_t h(w_{t+1}) + \langle s_t, w - w_{t+1} \rangle + t\Psi(w) - t\Psi(w_{t+1}) \\ &= \beta_t h(w) + \left\{ \langle -s_t, w_{t+1} - w_0 \rangle - t\Psi(w_{t+1}) - \beta_t h(w_{t+1}) \right\} \\ &\quad + t\Psi(w) + \langle s_t, w - w_0 \rangle.\end{aligned}$$

Using (43), we recognize that the collection in the braces is precisely $V_t(-s_t)$. Therefore

$$\frac{1}{2}(\sigma t + \beta_t)\|w_{t+1} - w\|^2 \leq \beta_t h(w) + V_t(-s_t) + t\Psi(w) + \langle s_t, w - w_0 \rangle. \tag{60}$$

Now we expand the last term $\langle s_t, w - w_0 \rangle$ using the definition of s_t:

$$\langle s_t, w - w_0 \rangle = \sum_{\tau=1}^{t} \langle g_\tau, w - w_0 \rangle = \sum_{\tau=1}^{t} \langle g_\tau, w_\tau - w_0 \rangle + \sum_{\tau=1}^{t} \langle g_\tau, w - w_\tau \rangle.$$

By further adding and subtracting $\sum_{\tau=1}^{t} \Psi(w_\tau)$, the right-hand side of (60) becomes

$$\beta_t h(w) + \left\{ V_t(-s_t) + \sum_{\tau=1}^{t} \left(\langle g_\tau, w_\tau - w_0 \rangle + \Psi(w_\tau) \right) \right\} + \sum_{\tau=1}^{t} \langle g_\tau, w - w_\tau \rangle + t\Psi(w) - \sum_{\tau=1}^{t} \Psi(w_\tau).$$

We recognize that the expression in the braces above is exactly the left-hand side in (56). Furthermore, by convexity of f_τ for $\tau \geq 1$,

$$\begin{aligned}\sum_{\tau=1}^{t} \langle g_\tau, w - w_\tau \rangle + t\Psi(w) - \sum_{\tau=1}^{t} \Psi(w_\tau) &\leq \sum_{\tau=1}^{t} \left(f_\tau(w) - f_\tau(w_\tau) \right) + t\Psi(w) - \sum_{\tau=1}^{t} \Psi(w_\tau) \\ &= \sum_{\tau=1}^{t} \left(f_\tau(w) + \Psi(w) \right) - \sum_{\tau=1}^{t} \left(f_\tau(w_\tau) + \Psi(w_\tau) \right) \\ &= -R_t(w).\end{aligned}$$

Putting everything together, and using (56), we have

$$\frac{1}{2}(\sigma t+\beta_t)\|w_{t+1}-w\|^2 \le \beta_t h(w)+(\beta_0-\beta_1)h(w_2)+\frac{1}{2}\sum_{\tau=1}^{t}\frac{\|g_\tau\|_*^2}{\sigma(\tau-1)+\beta_{\tau-1}}-R_t(w).$$

Finally, using $w\in\mathcal{F}_D$, Lemma 12 and the assumption (12), we conclude

$$\frac{1}{2}(\sigma t+\beta_t)\|w_{t+1}-w\|^2 \le \Delta_t - R_t(w),$$

which is the same as (16).

B.4 Bounding the Dual Average

First notice that (54) still holds if we replace w_0 with an arbitrary, fixed $w\in\mathcal{F}_D$, that is,

$$\delta_t=\sum_{\tau=1}^{t}\big(\langle g_\tau,w_\tau-w\rangle+\Psi(w_\tau)\big)+\max_{u\in\mathcal{F}_D}\big\{\langle s_t,w-u\rangle-t\Psi(u)\big\}.$$

By convexity of f_τ for $\tau\ge 1$, we have

$$\begin{aligned}
\delta_t &\ge \sum_{\tau=1}^{t}\big(f_\tau(w_\tau)-f_\tau(w)+\Psi(w_\tau)\big)+\max_{u\in\mathcal{F}_D}\big\{\langle s_t,w-u\rangle-t\Psi(u)\big\}\\
&=\sum_{\tau=1}^{t}\big(f_\tau(w_\tau)+\Psi(w_\tau)-f_\tau(w)-\Psi(w)\big)+\max_{u\in\mathcal{F}_D}\big\{\langle s_t,w-u\rangle-t\big(\Psi(u)-\Psi(w)\big)\big\}\\
&=R_t(w)+\max_{u\in\mathcal{F}_D}\big\{\langle s_t,w-u\rangle-t\big(\Psi(u)-\Psi(w)\big)\big\}. \qquad (61)
\end{aligned}$$

Let $d(u)$ denote a subgradient of Ψ at u with minimum norm, that is,

$$d(u)=\operatorname*{arg\,min}_{g\in\partial\Psi(u)}\|g\|_*. \qquad (62)$$

Since Ψ has convexity parameter σ, we have

$$\Psi(w)-\Psi(u)\ge\langle d(u),w-u\rangle+\frac{\sigma}{2}\|w-u\|^2.$$

Therefore,

$$\begin{aligned}
\delta_t &\ge R_t(w)+\max_{u\in\mathcal{F}_D}\left\{\langle s_t,w-u\rangle+t\langle d(u),w-u\rangle+\frac{\sigma t}{2}\|w-u\|^2\right\}\\
&\ge R_t(w)+\max_{u\in\mathcal{B}(w,r)}\left\{\langle s_t,w-u\rangle+t\langle d(u),w-u\rangle+\frac{\sigma t}{2}\|w-u\|^2\right\},
\end{aligned}$$

where in the last inequality, we used the assumption $\mathcal{B}(w,r)\subset\mathcal{F}_D$ for some $r>0$. Let $u^\star$ be the maximizer of $\langle s_t,w-u\rangle$ within the set $B(w,r)$, that is,

$$u^\star=\operatorname*{arg\,max}_{u\in B(w,r)}\big\{\langle s_t,w-u\rangle\big\}.$$

Then $\|w-u^\star\| = r$ and

$$\langle s_t, w-u^\star\rangle = \|w-u^\star\|\|s_t\|_* = r\|s_t\|_*.$$

So we can continue with the inequality:

$$\begin{aligned}
\delta_t &\geq R_t(w) + \langle s_t, w-u^\star\rangle + t\langle d(u^\star), w-u^\star\rangle + \frac{\sigma t}{2}\|w-u^\star\|^2 \\
&= R_t(w) + r\|s_t\|_* + t\langle d(u^\star), w-u^\star\rangle + \frac{1}{2}\sigma t r^2 \\
&\geq R_t(w) + r\|s_t\|_* - t\|d(u^\star)\|_*\|w-u^\star\| + \frac{1}{2}\sigma t r^2 \\
&\geq R_t(w) + r\|s_t\|_* - rt\Gamma_D + \frac{1}{2}\sigma t r^2
\end{aligned}$$

where in the last inequality, we used $\|d(u^\star)\|_* \leq \Gamma_D$, which is due to (62) and (11). Therefore

$$\|s_t\|_* \leq t\Gamma_D - \frac{1}{2}\sigma t r + \frac{1}{r}\big(\delta_t - R_t(w)\big).$$

Finally, we have (17) by noting $\delta_t \leq \Delta_t$ and $s_t = t\bar{g}$.

Appendix C. Proof of High Probability Bounds

In this Appendix, we prove Theorem 5. First let $\varphi(w) = \mathbf{E}_z f(w,z)$, then by definition $\phi(w) = \varphi(w) + \Psi(w)$. Let $\hat{g}_t$ be the conditional expectation of g_t given w_t, that is,

$$\hat{g}_t = \mathbf{E}[g_t \,|\, w_t] = \mathbf{E}\big[g_t \,|\, \mathbf{z}[t-1]\big].$$

Since $g_t \in \partial f(w_t, z_t)$, we have $\hat{g}_t \in \partial\varphi(w_t)$ (e.g., Rockafellar and Wets, 1982). By the definition of δ_t in (52), for any $w^\star \in \mathcal{F}_D$,

$$\begin{aligned}
\delta_t &\geq \sum_{\tau=1}^{t}\Big(\langle g_\tau, w_\tau - w^\star\rangle + \Psi(w_\tau)\Big) - t\Psi(w^\star) \\
&= \sum_{\tau=1}^{t}\Big(\langle \hat{g}_\tau, w_\tau - w^\star\rangle + \Psi(w_\tau)\Big) - t\Psi(w^\star) + \sum_{\tau=1}^{t}\langle g_\tau - \hat{g}_\tau, w_\tau - w^\star\rangle \\
&\geq \sum_{\tau=1}^{t}\Big(\varphi(w_\tau) - \varphi(w^\star) + \Psi(w_\tau)\Big) - t\Psi(w^\star) + \sum_{\tau=1}^{t}\langle g_\tau - \hat{g}_\tau, w_\tau - w^\star\rangle \\
&= \sum_{\tau=1}^{t}\big(\phi(w_\tau) - \phi(w^\star)\big) + \sum_{\tau=1}^{t}\langle g_\tau - \hat{g}_\tau, w_\tau - w^\star\rangle. \qquad (63)
\end{aligned}$$

where in the second inequality above we used convexity of φ. Now define the random variables

$$\xi_\tau = \langle g_\tau - \hat{g}_\tau, w^\star - w_\tau\rangle, \qquad \forall\, \tau \geq 1.$$

Combining (63) and the result $\delta_t \leq \Delta_t$ leads to

$$\sum_{\tau=1}^{t}\big(\phi(w_\tau) - \phi(w^\star)\big) \leq \Delta_t + \sum_{\tau=1}^{t}\xi_\tau. \qquad (64)$$

Since w_t is a deterministic function of $\mathbf{z}[t-1]$ and $\hat{g}_t = \mathbf{E}[g_t \,|\, w_t]$, we have

$$\mathbf{E}\big[\xi_\tau \,\big|\, \mathbf{z}[\tau-1]\big] = 0.$$

Therefore the sum $\sum_{\tau=1}^t \xi_\tau$ is a martingale. By the assumptions $h(w_\tau) \le D^2$ and $\|g_\tau\|_* \le L$ for all w_τ, we have

$$\|w - w_\tau\| \le \|w - w_0\| + \|w_\tau - w_0\| \le (2h(w))^{1/2} + (2h(w_\tau))^{1/2} \le 2\sqrt{2}D,$$

and $\|g_\tau - G_\tau\|_* \le \|g_\tau\|_* + \|G_\tau\|_* \le 2L$. Therefore,

$$|\xi_\tau| \le \|g_\tau - G_\tau\|_* \|w - w_\tau\| \le 4\sqrt{2}LD$$

So the sequence of random variables $\{\xi_\tau\}_{\tau=1}^t$ form a bounded martingale difference. Now by Hoeffding-Azuma inequality (Azuma, 1967), we have

$$\text{Prob}\left(\sum_{\tau=1}^t \xi_\tau \ge \Theta\right) \le \exp\left(\frac{-\Theta^2}{2t(4\sqrt{2}LD)^2}\right) = \exp\left(-\frac{\Theta^2}{64L^2D^2t}\right), \qquad \forall \Theta > 0.$$

Let $\Omega = \Theta/(8LD\sqrt{t})$, we have

$$\text{Prob}\left(\sum_{\tau=1}^t \xi_\tau \ge 8LD\sqrt{t}\,\Omega\right) \le \exp(-\Omega^2).$$

Now combining with (64) yields

$$\text{Prob}\left(\phi(\bar{w}_t) - \phi^\star \ge \frac{\Delta_t}{t} + \frac{8LD\Omega}{\sqrt{t}}\right) \le \exp(-\Omega^2).$$

Setting $\delta = \exp(-\Omega^2)$ gives the desired result (26).

Appendix D. Convergence Analysis of Accelerated RDA Method

In this appendix, we prove Theorem 6. We will need the following lemma.

Lemma 13 *Let ψ be a closed proper convex function, and h be strongly convex on* $\operatorname{dom}\psi$ *with convexity parameter σ_h. If*

$$v = \arg\min_w \big\{\psi(w) + h(w)\big\}, \tag{65}$$

then

$$\psi(w) + h(w) \ge \psi(v) + h(v) + \frac{\sigma_h}{2}\|w - v\|^2, \qquad \forall w \in \operatorname{dom}\psi.$$

Proof By the optimality condition for (65), there exist $b \in \partial\psi(v)$ and $d \in \partial h(v)$ such that

$$\langle b + d, w - v\rangle \ge 0, \qquad \forall w \in \operatorname{dom}\psi.$$

Since ψ is convex and h is strongly convex, we have

$$\begin{aligned}\psi(w) &\ge \psi(v) + \langle b, w - v\rangle,\\ h(w) &\ge h(v) + \langle d, w - v\rangle + \frac{\sigma_h}{2}\|w - v\|^2.\end{aligned}$$

The lemma is proved by combining the three inequalities above. ■

In Lemma 13, we do not assume differentiability of either ψ or h. Similar results assuming differentiability have appeared in, for example, Chen and Teboulle (1993) and Tseng (2008), where the term $(\sigma_h/2)\|w-v\|^2$ was replaced by the Bregman divergence induced by h.

Our proof combines several techniques appeared separately in Nesterov (2005), Tseng (2008), Lan (2010), and Nemirovski et al. (2009). First, let $\varphi(w) = \mathbf{E}_z f(w,z)$. For every $t \geq 1$, define the following two functions:

$$\ell_t(w) = \varphi(u_t) + \langle \nabla\varphi(u_t), w-u_t\rangle + \Psi(w),$$
$$\hat{\ell}_t(w) = \varphi(u_t) + \langle g_t, w-u_t\rangle + \Psi(w).$$

Note that $\ell_t(w)$ is a lower bound of $\phi(w)$ for all $t \geq 1$. Let $q_t = g_t - \nabla\varphi(u_t)$, then

$$\hat{\ell}_t(w) = \ell_t(w) + \langle q_t, w-u_t\rangle.$$

For each $t \geq 1$, we also define the function

$$\psi_t(w) = \sum_{\tau=1}^{t} \alpha_\tau \hat{\ell}_\tau(w).$$

For convenience, let $\psi_0(w) = 0$. Then step 4 in Algorithm 3 is equivalent to

$$v_t = \arg\min_w \big\{\psi_t(w) + (L+\beta_t)h(w)\big\}. \tag{66}$$

Since $\nabla\varphi$ is Lipschitz continuous with a constant L (see discussions following (34)), we have

$$\varphi(w_t) \leq \varphi(u_t) + \langle \nabla\varphi(u_t), w_t-u_t\rangle + \frac{L}{2}\|w_t-u_t\|^2.$$

Adding $\Psi(w_t)$ to both sides of the above inequality yields

$$\begin{aligned}
\phi(w_t) &\leq \ell_t(w_t) + \frac{L}{2}\|w_t-u_t\|^2 \\
&= \ell_t\big((1-\theta_t)w_{t-1} + \theta_t v_t\big) + \frac{L}{2}\big\|(1-\theta_t)w_{t-1} + \theta_t v_t - u_t\big\|^2 \\
&\leq (1-\theta_t)\ell_t(w_{t-1}) + \theta_t\ell_t(v_t) + \frac{L}{2}\big\|\theta_t v_t - \theta_t v_{t-1}\big\|^2 \\
&= (1-\theta_t)\ell_t(w_{t-1}) + \theta_t\hat{\ell}_t(v_t) - \theta_t\langle q_t, v_t-u_t\rangle + \theta_t^2\frac{L}{2}\big\|v_t-v_{t-1}\big\|^2 \\
&= (1-\theta_t)\ell_t(w_{t-1}) + \frac{1}{A_t}\left(\alpha_t\hat{\ell}_t(v_t) + \frac{\alpha_t^2}{A_t}\frac{L}{2}\big\|v_t-v_{t-1}\big\|^2\right) - \theta_t\langle q_t, v_t-u_t\rangle \\
&\leq (1-\theta_t)\phi(w_{t-1}) + \frac{1}{A_t}\left(\alpha_t\hat{\ell}_t(v_t) + \frac{L}{2}\big\|v_t-v_{t-1}\big\|^2\right) - \theta_t\langle q_t, v_t-u_t\rangle.
\end{aligned}$$

In the second inequality above, we used convexity of ℓ_t and $u_t = (1-\theta_t)w_{t-1} + \theta_t v_{t-1}$, and in the last inequality above, we used $\ell_t(w) \leq \phi(w)$ and the assumption $\alpha_t^2 \leq A_t$. Multiplying both sides of

the above inequality by A_t and noticing $A_t(1-\theta_t) = A_t - \alpha_t = A_{t-1}$, we have

$$\begin{aligned}
A_t\phi(w_t) &\le A_{t-1}\phi(w_{t-1}) + \alpha_t\hat{\ell}_t(v_t) + \frac{L}{2}\left\|v_t - v_{t-1}\right\|^2 - \alpha_t\langle q_t, v_t - u_t\rangle \\
&= A_{t-1}\phi(w_{t-1}) + \alpha_t\hat{\ell}_t(v_t) + \frac{L+\beta_{t-1}}{2}\left\|v_t - v_{t-1}\right\|^2 - \frac{\beta_{t-1}}{2}\left\|v_t - v_{t-1}\right\|^2 \\
&\quad - \alpha_t\langle q_t, v_t - v_{t-1}\rangle - \alpha_t\langle q_t, v_{t-1} - u_t\rangle \\
&\le A_{t-1}\phi(w_{t-1}) + \alpha_t\hat{\ell}_t(v_t) + \frac{L+\beta_{t-1}}{2}\left\|v_t - v_{t-1}\right\|^2 - \frac{\beta_{t-1}}{2}\left\|v_t - v_{t-1}\right\|^2 \\
&\quad + \alpha_t\|q_t\|_*\|v_t - v_{t-1}\| - \alpha_t\langle q_t, v_{t-1} - u_t\rangle.
\end{aligned}$$

Now using the inequality

$$bc - \frac{a}{2}c^2 \le \frac{b^2}{2a}, \qquad \forall a > 0,$$

with $a = \beta_{t-1}$, $b = \alpha_t\|q_t\|_*$, and $c = \|v_t - v_{t-1}\|$, we have

$$A_t\phi(w_t) \le A_{t-1}\phi(w_{t-1}) + \alpha_t\hat{\ell}_t(v_t) + \frac{L+\beta_{t-1}}{2}\left\|v_t - v_{t-1}\right\|^2 + \frac{\alpha_t^2\|q_t\|_*^2}{2\beta_{t-1}} - \alpha_t\langle q_t, v_{t-1} - u_t\rangle.$$

By (66), v_{t-1} is the minimizer of $\psi_{t-1}(v) + (L+\beta_{t-1})h(v)$. Then by Lemma 13, we have

$$\psi_{t-1}(v_t) + (L+\beta_{t-1})h(v_t) \ge \psi_{t-1}(v_{t-1}) + (L+\beta_{t-1})h(v_{t-1}) + \frac{L+\beta_{t-1}}{2}\left\|v_t - v_{t-1}\right\|^2,$$

therefore,

$$\begin{aligned}
A_t\phi(w_t) - \psi_t(v_t) - (L+\beta_{t-1})h(v_t) \le{}& A_{t-1}\phi(w_{t-1}) - \psi_{t-1}(v_{t-1}) - (L+\beta_{t-1})h(v_{t-1}) \\
&+ \frac{\alpha_t^2\|q_t\|_*^2}{2\beta_{t-1}} - \alpha_t\langle q_t, v_{t-1} - u_t\rangle.
\end{aligned}$$

Since $\beta_t \ge \beta_{t-1}$ and $h(v_t) \ge 0$, we can replace the β_{t-1} on the left-hand side with β_t:

$$\begin{aligned}
A_t\phi(w_t) - \psi_t(v_t) - (L+\beta_t)h(v_t) \le{}& A_{t-1}\phi(w_{t-1}) - \psi_{t-1}(v_{t-1}) - (L+\beta_{t-1})h(v_{t-1}) \\
&+ \frac{\alpha_t^2\|q_t\|_*^2}{2\beta_{t-1}} - \alpha_t\langle q_t, v_{t-1} - u_t\rangle.
\end{aligned}$$

Summing the above inequality from $\tau = 1$ to t results in

$$\begin{aligned}
A_t\phi(w_t) \le{}& \psi_t(v_t) + (L+\beta_t)h(v_t) + A_0\phi(w_0) - \psi_0(v_0) - (L+\beta_0)h(v_0) \\
&+ \sum_{\tau=1}^{t}\frac{\alpha_\tau^2\|q_\tau\|_*^2}{2\beta_{\tau-1}} + \sum_{\tau=1}^{t}\alpha_\tau\langle q_\tau, u_\tau - v_{\tau-1}\rangle.
\end{aligned}$$

Using $A_0 = 0$, $\psi_0(v_0) = 0$, $h(v_0) = 0$, and (66), we have

$$\begin{aligned}
A_t\phi(w_t) &\le \psi_t(w^\star) + (L+\beta_t)h(w^\star) + \sum_{\tau=1}^{t} \frac{\alpha_\tau^2 \|q_\tau\|_*^2}{2\beta_{\tau-1}} + \sum_{\tau=1}^{t} \alpha_\tau \langle q_\tau, u_\tau - v_{\tau-1}\rangle \\
&= \sum_{\tau=1}^{t} \alpha_\tau \hat{\ell}_\tau(w^\star) + (L+\beta_t)h(w^\star) + \sum_{\tau=1}^{t} \frac{\alpha_\tau^2 \|q_\tau\|_*^2}{2\beta_{\tau-1}} + \sum_{\tau=1}^{t} \alpha_\tau \langle q_\tau, u_\tau - v_{\tau-1}\rangle \\
&= \sum_{\tau=1}^{t} \alpha_\tau \big(\ell_\tau(w^\star) + \langle q_\tau, w^\star - u_\tau\rangle\big) + (L+\beta_t)h(w^\star) + \sum_{\tau=1}^{t} \frac{\alpha_\tau^2 \|q_\tau\|_*^2}{2\beta_{\tau-1}} \\
&\quad + \sum_{\tau=1}^{t} \alpha_\tau \langle q_\tau, u_\tau - v_{\tau-1}\rangle \\
&= \sum_{\tau=1}^{t} \alpha_\tau \ell_\tau(w^\star) + (L+\beta_t)h(w^\star) + \sum_{\tau=1}^{t} \frac{\alpha_\tau^2 \|q_\tau\|_*^2}{2\beta_{\tau-1}} + \sum_{\tau=1}^{t} \alpha_\tau \langle q_\tau, w^\star - v_{\tau-1}\rangle.
\end{aligned}$$

Next, by $\ell_\tau(w^\star) \le \phi(w^\star)$ for all $\tau \ge 1$ and $A_t = \sum_{\tau=1}^{t} \alpha_\tau$, we have

$$A_t\phi(w_t) \le A_t\phi(w^\star) + (L+\beta_t)h(w^\star) + \sum_{\tau=1}^{t} \frac{\alpha_\tau^2 \|q_\tau\|_*^2}{2\beta_{\tau-1}} + \sum_{\tau=1}^{t} \alpha_\tau \langle q_\tau, w^\star - v_{\tau-1}\rangle. \tag{67}$$

Since $\mathbf{E}\big[q_\tau | \mathbf{z}[\tau-1]\big] = 0$ and q_τ is independent of $v_{\tau-1}$, we have $\mathbf{E}\big[\langle q_\tau, w^\star - v_{\tau-1}\rangle | \mathbf{z}[\tau-1]\big] = 0$. Together with the assumption $\mathbf{E}\|q_\tau\|_*^2 \le Q^2$ for all $\tau \ge 1$, we conclude

$$\mathbf{E}\,\phi(w_t) - \phi(w^\star) \le \frac{L+\beta_t}{A_t} h(w^\star) + \frac{1}{A_t}\left(\frac{Q^2}{2} \sum_{\tau=1}^{t} \frac{\alpha_\tau^2}{\beta_{\tau-1}}\right).$$

By rearranging terms on the right-hand side, this finishes the proof for Theorem 6.

D.1 Proof of Corollary 7

Using the two input sequences given in (37) and (38), we have

$$\sum_{\tau=1}^{t} \frac{\alpha_\tau^2}{2\beta_{\tau-1}} = \frac{1}{4\gamma} \sum_{\tau=1}^{t} \tau^{1/2} \le \frac{1}{4\gamma} \int_0^{t+1} \tau^{1/2} d\tau = \frac{(t+1)^{3/2}}{6\gamma}. \tag{68}$$

Plugging them into the conclusion of Theorem 6 gives

$$\mathbf{E}\phi(w_t) - \phi^\star \le \frac{4L}{t(t+1)} h(w^\star) + \frac{(t+1)^{1/2}}{t}\left(2\gamma h(w^\star) + \frac{2Q^2}{3\gamma}\right).$$

Next we use the assumption $h(w^\star) \le D^2$ and let $\gamma = Q/D$. Then

$$\mathbf{E}\phi(w_t) - \phi^\star \le \frac{4LD^2}{t(t+1)} + \frac{(t+1)^{1/2}}{t} \frac{8QD}{3} \le \frac{4LD}{t^2} + \frac{4QD}{\sqrt{t}}.$$

D.2 Proof of Theorem 8

We start with the inequality (67). We will first show high probability bounds for the two summations on the right-hand side of (67) that involve the stochastic quantities q_τ, and then combine them to prove Theorem 8. We need the following result on large-deviation bound for martingales, which can be viewed as an extension to the Hoeffding-Azuma inequality.

Lemma 14 *(Lan et al., 2008, Lemma 6) Let $z_1, z_2, \ldots$ be a sequence of i.i.d. random variables and let $\mathbf{z}[t]$ denote the collection $[z_1, \ldots, z_t]$. If $\xi_t = \xi_t(\mathbf{z}[t])$ are deterministic Borel functions of $\mathbf{z}[t]$ such that the conditional expectations $\mathbf{E}\big[\xi_t \,|\, \mathbf{z}[t-1]\big] = 0$ almost surely and*

$$\mathbf{E}\big[\exp(\xi_t^2/\nu_t^2) \,|\, \mathbf{z}[t-1]\big] \le \exp(1) \tag{69}$$

almost surely, where $\nu_t > 0$ are deterministic. Then for all $t \ge 1$,

$$\text{Prob}\left(\sum_{\tau=1}^{t} \xi_\tau > \Omega \sqrt{\sum_{\tau=1}^{t} \nu_\tau^2}\right) \le \exp\left(-\frac{\Omega^2}{3}\right), \qquad \forall\, \Omega \ge 0.$$

Lemma 15 *Let $\xi_t = \alpha_t \langle q_t, w^\star - v_{t-1}\rangle$. Then for all $t \ge 1$ and any $\Omega > 0$,*

$$\text{Prob}\left(\sum_{\tau=1}^{t} \xi_\tau > \Omega QD \sqrt{\frac{2}{3}(t+1)^3}\right) \le \exp(-\Omega^2/3).$$

Proof Since $\mathbf{E}\big[q_t \,|\, \mathbf{z}[t-1]\big] = 0$ and q_t is independent of $w^\star$ and v_{t-1}, we have

$$\mathbf{E}\big[\xi_t \big| \mathbf{z}[t-1]\big] = \mathbf{E}\big[\alpha_t \langle q_t, w^\star - v_{t-1}\rangle \big| \mathbf{z}[t-1]\big] = 0.$$

Therefore, $\sum_{\tau=1}^{t} \xi_\tau$ is a martingale. By the assumption $(1/2)\|w\|^2 \le h(w) \le D^2$ for all $w \in \text{dom}\,\Psi$, we have $\|w\| \le \sqrt{2}D$ for all $w \in \text{dom}\,\Psi$, and therefore

$$|\xi_t| \le \alpha_t \|q_t\|_* \|w^\star - v_{t-1}\| \le \alpha_t \|q_t\|_* \big(\|w^\star\| + \|v_{t-1}\|\big) \le \alpha_t \|q_t\|_* 2\sqrt{2}D.$$

Using the assumption $\mathbf{E}\big[\exp\big(\|q_t\|_*^2/Q^2\big)\big] \le \exp(1)$, we have

$$\mathbf{E}\left[\exp\left(\frac{\xi_t^2}{(8\alpha_t^2 Q^2 D^2)^2}\right)\Big|\mathbf{z}[t-1]\right] \le \mathbf{E}\left[\exp\left(\frac{(\alpha_t \|q_t\|_* 2\sqrt{2}D)^2}{8\alpha_t^2 Q^2 D^2}\right)\Big|\mathbf{z}[t-1]\right] \le \exp(1).$$

Therefore the condition (69) holds with $\nu_t^2 = 8\alpha_t^2 Q^2 D^2$. We bound $\sum_{\tau=1}^{t} \nu_\tau^2$ as follows:

$$\sum_{\tau=1}^{t} \nu_\tau^2 \le 8Q^2D^2 \sum_{\tau=1}^{t} \alpha_\tau^2 = 2Q^2D^2 \sum_{\tau=1}^{t} \tau^2 \le 2Q^2D^2 \int_0^{t+1} \tau^2 d\tau = \frac{2Q^2D^2}{3}(t+1)^3.$$

Then applying Lemma 14 gives the desired result. ■

Lemma 16 *For all $t \ge 1$ and any $\Lambda > 0$,*

$$\text{Prob}\left(\sum_{\tau=1}^{t} \frac{\alpha_\tau^2}{2\beta_{\tau-1}} \|q_\tau\|_*^2 > (1+\Lambda)\frac{Q^2}{6\gamma}(t+1)^{3/2}\right) \le \exp(-\Lambda).$$

Proof For any given $t \geq 1$, let

$$\Theta_t = \sum_{\tau=1}^{t} \frac{\alpha_\tau^2}{2\beta_{\tau-1}},$$

and

$$\eta_\tau = \frac{\alpha_\tau^2}{2\beta_{\tau-1}} \frac{1}{\Theta_t}, \qquad \tau = 1, \ldots, t.$$

Therefore $\sum_{\tau=1}^{t} \eta_\tau = 1$. By convexity of the function $\exp(\cdot)$,

$$\exp\left(\sum_{\tau=1}^{t} \eta_\tau \frac{\|q_\tau\|_*^2}{Q^2} \right) \leq \sum_{\tau=1}^{t} \eta_\tau \exp\left(\frac{\|q_\tau\|_*^2}{Q^2} \right).$$

Taking expectation and using the assumption (39),

$$\mathbf{E} \exp\left(\sum_{\tau=1}^{t} \eta_\tau \frac{\|q_\tau\|_*^2}{Q^2} \right) \leq \sum_{\tau=1}^{t} \eta_\tau \mathbf{E} \exp\left(\frac{\|q_\tau\|_*^2}{Q^2} \right) \leq \sum_{\tau=1}^{t} \eta_\tau \exp(1) = \exp(1).$$

By Markov's inequality,

$$\mathrm{Prob}\left(\exp\left(\sum_{\tau=1}^{t} \eta_\tau \frac{\|q_\tau\|_*^2}{Q^2} \right) > \exp(1+\Lambda) \right) \leq \frac{\exp(1)}{\exp(1+\Lambda)} = \exp(-\Lambda),$$

which is the same as

$$\mathrm{Prob}\left(\sum_{\tau=1}^{t} \frac{\alpha_\tau^2}{2\beta_{\tau-1}} \|q_\tau\|_*^2 > (1+\Lambda)\Theta_t Q^2 \right) \leq \exp(-\Lambda).$$

Then using the upper bound on Θ_t derived in (68) gives the desired result. ■

Combining Lemma 15, Lemma 16, and the inequality (67), we have

$$\mathrm{Prob}\left(A_t(\phi(w_t) - \phi^\star) > (L+\beta_t)h(w^\star) + (1+\Lambda)\frac{Q^2}{6\gamma}(t+1)^{3/2} + \Omega Q D \sqrt{\frac{2}{3}}(t+1)^{3/2} \right)$$
$$\leq \exp(-\Lambda) + \exp\left(-\frac{\Omega^2}{3} \right).$$

Plugging in $A_t = t(t+1)/4$, $\beta_t = (\gamma/2)(t+1)^{3/2}$, and letting $\gamma = Q/D$, $\Omega = \sqrt{3\Lambda}$, we get

$$\mathrm{Prob}\left(\phi(w_t) - \phi^\star > \frac{4LD^2}{t(t+1)} + \left(\frac{8QD}{3} + \frac{2\Lambda QD}{3} + \sqrt{2\Lambda} QD \right) \frac{(t+1)^{1/2}}{t} \right) \leq 2\exp(-\Lambda).$$

Then using the fact $\sqrt{(t+1)/t} \leq \sqrt{2} \leq 3/2$ for all $t \geq 1$ results in

$$\mathrm{Prob}\left(\phi(w_t) - \phi^\star > \frac{4LD^2}{t^2} + \frac{4QD}{\sqrt{t}} + \frac{(\Lambda + 2\sqrt{\Lambda})QD}{\sqrt{t}} \right) \leq 2\exp(-\Lambda).$$

Finally, let $\delta = 2\exp(-\Lambda)$, hence $\Lambda = \ln(2/\delta)$. Then with probability at least $1-\delta$,

$$\phi(w_t) - \phi^\star \leq \frac{4LD^2}{t^2} + \frac{4QD}{\sqrt{t}} + \frac{QD}{\sqrt{t}} \left(\ln(2/\delta) + 2\sqrt{\ln(2/\delta)} \right).$$

This finishes the proof of Theorem 8.

References

G. Andrew and J. Gao. Scalable training of l_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 33–40, Corvallis, OR, USA, 2007.

A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16:697–725, 2006.

K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19:357–367, 1967.

S. Balakrishnan and D. Madigan. Algorithms for sparse linear classifiers in the massive data setting. *Journal of Machine Learning Research*, 9:313–337, 2008.

P. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 65–72. MIT Press, Cambridge, MA, 2008.

A. Beck and M. Teboulle. A fast iterative shrinkage-threshold algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 161–168. MIT Press, Cambridge, MA, 2008.

L. Bottou and Y. LeCun. Large scale online learning. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 217–224. MIT Press, Cambridge, MA, 2004.

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

D. M. Bradley and J. A. Bagnell. Differentiable sparse coding. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 113–120. MIT Press, Cambridge, MA, USA, 2009.

K. Bredies and D. A. Lorenz. Iterated hard shrinkage for minimization problems with sparsity constraints. *SIAM Journal on Scientific Computing*, 30(2):657–683, 2008.

P. Carbonetto, M. Schmidt, and N. De Freitas. An interior-point stochastic approximation method and an l_1-regularized delta rule. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 233–240. MIT Press, 2009.

N. Cesa-Bianchi and G. Lugosi. *Predictioin, Learning, and Games*. Cambridge University Press, 2006.

G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, August 1993.

G. H.-G. Chen and R. T. Rockafellar. Convergence rates in forward-backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2873–2898, 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 (ICML)*, pages 272–279, 2008.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. To appear in *Journal of Machine Learning Research*, 2010.

M. C. Ferris and T. S. Munson. Interior-point methods for massive support vector machines. *SIAM Journal on Optimization*, 13(3):783–804, 2003.

M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal on Selected Topics in Signal Processing*, 1(4):586–597, 2007.

D. A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

C. Gentile. The robustness of the *p*-norm algorithms. *Machine Learning*, 53:265–299, 2003.

R. Goebel and R. T. Rockafellar. Local strong convexity and local Lipschitz continuity of the gradient of convex functions. *Journal of Convex Analysis*, 15(2):263–270, 2008.

E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. In *Proceedings of 19th Annual Conference on Computational Learning Theory (COLT)*, pages 499–513, Pittsburgh, PA, USA, 2006.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2001.

C. Hu, J. T. Kwok, and W. Pan. Accelerated gradient methods for stochastic optimization and online learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 781–789. 2009.

B. Johansson, M. Rabi, and M. Johansson. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3):1157–1170, 2009.

A. Juditsky and A. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. Manuscript submitted to *The Annals of Probability*, 2008. arXiv:0809.0813v1.

A. Juditsky, A. Nazin, A. Tsybakov, and N. Vayatis. Recursive aggregation of estimators by mirror descent algorithm with averaging. *Problems of Information Transmission*, 41(4):368–384, 2005.

S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 801–808. MIT Press, Cambridge, MA, USA, 2009.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annuals of Mathematical Statistics*, 23:462–466, 1952.

J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

K. Koh, S.-J. Kim, and S. Boyd. An interior-point method for large-scale ℓ_1-regularized logistic regression. *Journal of Machine Learning Research*, 8:1519–1555, 2007.

G. Lan. An optimal method for stochastic composite optimization. To appear in *Mathematical Programming*, 2010.

G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of robust stochastic approximation methods. Submitted to *Mathematical Programming*, 2008.

G. Lan, Z. Lu, and R. D. C. Monteiro. Primal-dual first-order methods with $O(1/\varepsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, February 2009. Published online, DOI 10.1007/s10107-008-0261-6.

J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777–801, 2009.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. Dataset available at `http://yann.lecun.com/exdb/mnist`.

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

Z. Lu. Primal-dual first-order methods for a class of cone programming with applications to the Dantzig selector. Submitted manuscript, 2009.

A. Nedić and D. P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

A. Nemirovsky and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. J. Wiley & Sons, New York, 1983.

Yu. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Doklady*, 27(2):372–376, 1983. Translated from Russian by A. Rosa.

Yu. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, Boston, 2004.

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

Yu. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 2007/76, Catholic University of Louvain, Center for Operations Research and Econometrics, 2007.

Yu. Nesterov. How to advance in structural convex optimization. *OPTIMA: Mathematical Programming Society Newsletter*, 78:2–5, November 2008.

Yu. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009. Appeared early as CORE discussion paper 2005/67, Catholic University of Louvain, Center for Operations Research and Econometrics.

Yu. Nesterov and J.-Ph. Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6):1559–1568, 2008.

B. T. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 1992.

S. Sundhar Ram, A. Nedić, and V. V. Veeravalli. Incremental stochastic subgradient algorithms for convex optimization. *SIAM Journal on Optimization*, 20(2):691–717, 2009.

H. Robbins and S. Monro. A stochastic approximation method. *The Annuals of Mathematical Statistics*, 22:400–407, 1951.

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

R. T. Rockafellar and R. J. B. Wets. On the interchange of subdifferentiation and conditional expectation for convex functionals. *Stochastics An International Journal of Probability and Stochastic Processes*, 7(3):173–182, 1982.

S. Shalev-Shwartz and S. M. Kakade. Mind the duality gap: Logarithmic regret algorithms for online optimization. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1457–1464. MIT Press, 2009.

S. Shalev-Shwartz and Y. Singer. Convex repeated games and Fenchel duality. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems*, volume 19, pages 1265–1272. MIT Press, 2006.

S. Shalev-Shwartz and A. Tewari. Stochastic methods for ℓ_1 regularized loss minimization. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 929–936, Montreal, Canada, 2009.

S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 807–814, 2007.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

P. Tseng. An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.

P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Manuscript submitted to *SIAM Journal on Optimization*, 2008.

P. Tseng and D. P. Bertsekas. On the convergence of exponential multiplier method for convex programming. *Mathematical Programming*, 60:1–19, 1993.

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

T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 116–123, Banff, Alberta, Canada, 2004.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 928–936, Washington DC, 2003.