Manifold Identification of Dual Averaging Methods for Regularized Stochastic Online Learning

Sangkyun Lee SKLEE@CS.WISC.EDU
Stephen J. Wright SWRIGHT@CS.WISC.EDU
Computer Sciences Department, University of Wisconsin, 1210 W. Dayton Street, Madison, WI 53706 USA

Abstract

Iterative methods that take steps in approximate subgradient directions have proved to be useful for stochastic learning problems over large or streaming data sets. When the objective consists of a loss function plus a nonsmooth regularization term, whose purpose is to induce structure (for example, sparsity) in the solution, the solution often lies on a low-dimensional manifold along which the regularizer is smooth. This paper shows that a regularized dual averaging algorithm can identify this manifold with high probability. This observation motivates an algorithmic strategy in which, once a near-optimal manifold is identified, we switch to an algorithm that searches only in this manifold, which typically has much lower intrinsic dimension than the full space, thus converging quickly to a near-optimal point with the desired structure. Computational results are presented to illustrate these claims.

1. Introduction

Stochastic approximation methods have recently proved to be effective in solving stochastic learning problems. Each step of these methods evaluates an approximate subgradient at the current iterate, based on a subset (perhaps a single item) of the data. This information is used, possibly in combination with subgradient information from previous iterates and a damping or line-search parameter, to obtain the next iterate.

We focus on objectives that consist of a smooth loss function in conjunction with a nonsmooth regularizer. A classic problem of this form is ℓ_1-regularized logistic regression. Xiao (2010) recently described a dual-averaging method, in which the smooth term is approximated by an averaged gradient, while the regularization term appears explicitly in each subproblem. This approach can be viewed as an extension of the method of Nesterov (2009) to the case in which a regularization term is present. Other methods have been proposed to minimize nonsmooth functions (for example, approaches based on smooth approximations (Nesterov, 2005)), but these approaches are less appealing when the regularizers have simple structure that allows them to be handled explicitly.

A characteristic of problems with nonsmooth regularizers is that the solution often lies on a manifold of low dimension. In ℓ_1-regularized problems, for instance, the number of nonzero components at the solution is often a small fraction of the dimension of the full space. Where a reliable method for identifying an optimal (or near-optimal) manifold is available, we have the possibility of invoking an algorithm that searches just in the low-dimensional space defined by this manifold — possibly a very different algorithm to one that would be used on the full space. One example of this type of approach is seen in `LPS` (Shi et al., 2008; Wright, 2010), a batch optimization method for ℓ_1-regularized logistic regression, which takes inexact Newton steps on the space of apparently nonzero variables, to supplement the partial gradient steps that are used in the full space. In logistic regression, and probably in other cases, it can be much less expensive to compute first- and second-order information on a restricted subspace than on the full space.

Identification of optimal manifolds has been studied in the context of convex constrained optimization (Burke & Moré, 1994; Wright, 1993) and nonsmooth nonconvex optimization (Hare & Lewis, 2004). In the latter setting, the optimal manifold is defined to be a smooth surface passing through the optimum, parameterizable by relatively few variables, such that the restriction of the objective to the manifold is smooth.

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

When a certain nondegeneracy condition is satisfied, this manifold may be identified *without knowing the solution*, usually as a by-product of an algorithm for solving the problem.

In this paper, we investigate the ability of a regularized dual averaging (RDA) algorithm described in Xiao (2010) to identify the optimal manifold. We also investigate this behavior computationally for the case of ℓ_1-regularized logistic regression, and suggest a technique for switching to a different method once a near-optimal manifold is identified, thus avoiding the sublinear asymptotic convergence rate that characterizes stochastic gradient methods.

The RDA algorithm averages the gradient information collected at the iterates, and the averaged gradient tend to converge to the optimal gradient as the iterates converge to a solution. Taken in conjunction with the nondegeneracy condition, this property provides the key to identification.

1.1. Notations

We use $\|\cdot\|$ (without a subscript) to denote the Euclidean norm $\|\cdot\|_2$, and $\mathrm{card}(M)$ to denote the cardinality of a finite set M. The distance function $\mathrm{dist}\,(w, C)$ for $w \in \mathbb{R}^n$ and a convex set $C \subset \mathbb{R}^n$ is defined by $\mathrm{dist}\,(w, C) := \inf_{c\in C} \|w - c\|$. The effective domain of $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is defined by $\mathrm{dom}\,\Psi := \{w \in \mathbb{R}^n \,|\, \Psi(w) < +\infty\}$; $\mathrm{ri}\, C$ denotes the relative interior of a convex set C, that is, the interior relative to the affine span of C (the smallest affine set which can be expressed as the intersection of hyperplanes containing C).

2. Regularized Stochastic Online Learning

In *regularized stochastic learning*, we consider the following problem:

$$\min_{w\in\mathbb{R}^n} \quad \phi(w) := f(w) + \Psi(w) \tag{1}$$

where $f(w) := \mathbb{E}_\xi F(w;\xi) = \int_\Xi F(w;\xi)dP(\xi)$ and ξ is a random vector with a probability distribution which is supported on the set $\Xi \in \mathbb{R}^d$. We assume that there is a open neighborhood $\mathcal{O}$ of $\mathrm{dom}\,\Psi$ which is contained in $\mathrm{dom}\, F(\cdot,\xi)$ for all $\xi \in \Xi$, where the expectation is well defined and finite valued. We assume that $F(w,\xi)$ is a smooth convex function for $w \in \mathcal{O}$ and every $\xi \in \Xi$, and $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a closed proper convex function with $\mathrm{dom}\,\Psi$ closed. We use w^* to denote the optimal solution of (1).

One method for obtaining an approximation to w^* is to draw random variables ξ_j, $j \in \mathcal{N}$ independently from the space Ξ, where $\mathcal{N}$ is an index set of finite cardinality, and solve

$$\min_{w\in\mathbb{R}^n} \quad \tilde{\phi}_{\mathcal{N}}(w) := \tilde{f}_{\mathcal{N}}(w) + \Psi(w) \tag{2}$$

where $\tilde{f}_{\mathcal{N}}(w) := \frac{1}{\mathrm{card}(\mathcal{N})}\sum_{j\in\mathcal{N}} F(w;\xi_j)$. This approach requires batch optimization, which does not scale well for $\mathcal{N}$ with large cardinality.

In *regularized stochastic online learning*, we encounter a previously unknown cost function $F(\cdot;\xi_t) : \mathbb{R}^n \to \mathbb{R}$ for $\xi_t \in \Xi$ in each time $t \geq 1$, where $\{\xi_t\}_{t\geq 1}$ forms an i.i.d. sequence of random samples. At each time t, we make a decision w_t using the information gathered up to the time t, and attempt to generate a sequence $\{w_t\}_{t\geq 1}$ such that

$$\lim_{t\to\infty} \mathbb{E}\,[F(w_t;\xi) + \Psi(w_t)] = f(w^*) + \Psi(w^*).$$

2.1. Regularized Dual Averaging Algorithm

In the regularized dual averaging (RDA) algorithm (Xiao, 2010), we define the dual average $\bar{g}_t$ to be the average of the approximate gradients $\nabla F(w_j;\xi_j)$ of $f(w_j)$ up to time t, that is,

$$\bar{g}_t := \frac{1}{t}\sum_{j=1}^{t} \nabla F(w_j;\xi_j). \tag{3}$$

(The differentiation of F is taken for its first argument.) The following subproblem is solved to obtain the iterate w_{t+1}:

$$w_{t+1} = \arg\min_{w\in\mathbb{R}^n} \left\{ \langle \bar{g}_t, w\rangle + \Psi(w) + \frac{\gamma}{\sqrt{t}}\|w - w_1\|^2 \right\}. \tag{4}$$

As the objective function in (4) is strongly convex for $\gamma > 0$, w_{t+1} is uniquely defined. Note that w_{t+1} depends on the history of random variables ξ_j up to time t; this history is denoted by $\xi_{[t]} := \{\xi_1, \xi_2, \ldots, \xi_t\}$. In particular, we have that w_{t+1} is independent of later samples $\xi_{t+1}, \xi_{t+2}, \ldots$.

2.2. Regret Bound for RDA Algorithm

We assume that there exists a bound $G > 0$ for the norms of the gradients, that is,

$$\|\nabla F(w;\xi)\| \leq G, \quad \forall w \in \mathcal{O}, \forall \xi \in \Xi.$$

Similarly to Nesterov (2009), we choose $w_1 \in \arg\min_{w\in\mathbb{R}^n} \Psi(w)$ and assume without loss of generality that $\Psi(w_1) = 0$. We also assume that the distance

between the solution w^* of (1) and w_1 is bounded by a constant $D > 0$, that is,

$$\|w^* - w_1\| \leq D.$$

We define the *regret* with respect to w^* as follows:

$$R_t := \sum_{j=1}^{t}(F(w_j;\xi_j)+\Psi(w_j))-\sum_{j=1}^{t}(F(w^*;\xi_j)+\Psi(w^*)).$$

The RDA algorithm has the following property for the regularized stochastic online learning:

Theorem 1. *For any instantiation of the sequence $\{w_j\}_{j=1}^t$ generated by the RDA algorithm, we have:*

$$R_t \leq \left(\gamma D^2 + \frac{G^2}{\gamma}\right)\sqrt{t}, \quad \forall t \geq 1.$$

Proof. See Corollary 2 of Xiao (2010). □

2.3. Manifold and Assumptions

We define some terminology regarding (differential) manifolds following Vaisman (1984).

Definition 1. *A set $\mathcal{M} \subset \mathbb{R}^n$ is a* manifold *about $\bar{z} \in \mathcal{M}$, if it can be described locally by a collection of at least twice continuously differentiable (i.e. $\mathcal{C}^p$, for $p \geq 2$) functions with linearly independent gradients. That is, there exists a map $H : \mathbb{R}^n \to \mathbb{R}^k$ that is $\mathcal{C}^p$ around $\bar{z}$ with $\nabla H(\bar{z})^T \in \mathbb{R}^{k\times n}$, surjective, such that points z near $\bar{z}$ lie in $\mathcal{M}$ if and only if $H(z) = 0$.*

Definition 2. *The* normal space *to $\mathcal{M}$ at $\bar{z}$, denoted by $N_{\mathcal{M}}(\bar{z})$, is the range space of $\nabla H(\bar{z})$.*

We assume that the following conditions hold throughout the paper:

- **Unbiasedness:** As in Nemirovski et al. (2009), we assume that $\nabla f(w) = \nabla \mathbb{E}_\xi F(w;\xi) = \mathbb{E}_\xi \nabla F(w;\xi)$ for w independent of ξ. This implies

$$\begin{aligned}\mathbb{E}[\nabla F(w_t;\xi_t)] &= \mathbb{E}\left[\mathbb{E}[\nabla F(w_t;\xi_t) \mid \xi_{[t-1]}]\right] \\ &= \mathbb{E}\left[\nabla f(w_t)\right].\end{aligned}$$

- **Lipschitz Property:** For all $t \geq 1$, $F(w;\xi_t)$ is differentiable with Lipschitz continuous derivatives in w with a uniform constant $L > 0$, i.e.,

$$\|\nabla F(w;\xi_t)-\nabla F(w';\xi_t)\| \leq L\|w-w'\|, \quad \forall w, w' \in \mathcal{O}.$$

  This assumption implies $\nabla f(w)$ is also Lipschitz continuous on $\mathcal{O}$ with the same constant L.

- **Nondegeneracy:** w^* is nondegenerate, that is,

$$0 \in \operatorname{ri} \partial\phi(w^*).$$

- **Partial Smoothness:** $\Psi(\cdot)$ is ($\mathcal{C}^2$-) *partly smooth* (Hare & Lewis, 2004) at w^* relative to a optimal manifold $\mathcal{M}$ containing w^*, that is,
  - (i) (Smoothness) The function Ψ restricted to $\mathcal{M}$, denoted by $\Psi|_{\mathcal{M}}$, is $\mathcal{C}^2$ near w^*.
  - (ii) (Regularity) Ψ is subdifferentially regular at all points $w \in \mathcal{M}$ near w^*, with $\partial\Psi(w) \neq \emptyset$.
  - (iii) (Sharpness) The affine span of $\partial\Psi(w^*)$ is a translate of $N_{\mathcal{M}}(w^*)$.
  - (iv) (Sub-continuity) The set-valued mapping $\partial\Psi : \mathcal{M} \rightrightarrows \mathbb{R}^n$ is continuous at w^*.

  Partial smoothness of $\phi(\cdot)$ follows from this condition since $f(\cdot)$ is smooth, which follows from the smoothness of $F(\cdot;\xi)$ for each $\xi \in \Xi$.

- **Strong Local Minimizer:** w^* is a *strong local minimizer* of ϕ relative to the optimal manifold $\mathcal{M}$, that is, there exists $c_{\mathcal{M}} > 0$ and $r_{\mathcal{M}} > 0$ such that

$$\begin{aligned}\phi|_{\mathcal{M}}(w) \geq \phi|_{\mathcal{M}}(w^*) + c_{\mathcal{M}}\|w - w^*\|^2,& \\ \forall w \in \mathcal{M} \cap \mathcal{O} \text{ s.t. } \|w - w^*\| \leq r_{\mathcal{M}}.&\end{aligned}$$

  Along with the other assumptions above, this implies that w^* is in fact a strong local minimizer of ϕ (Lee & Wright, 2011, Theorem 5), i.e. there exists $0 < c < c_{\mathcal{M}}$ and $0 < \bar{r} < r_{\mathcal{M}}$ such that

$$\begin{aligned}\phi(w) \geq \phi(w^*) + c\|w - w^*\|^2,& \\ \forall w \in \mathcal{O} \text{ s.t. } \|w - w^*\| \leq \bar{r}.&\end{aligned} \tag{5}$$

2.4. Manifold Identification

We now state a fundamental manifold identification result that will be used in our analysis.

Theorem 2. *For ϕ which is partly smooth at the nondegenerate minimizer w^* relative to the manifold $\mathcal{M}$, there is a threshold $\bar{\epsilon} > 0$ such that $\forall w \in \mathcal{O}$ with $\|w-w^*\| < \bar{\epsilon}$, and $\operatorname{dist}(0, \partial\phi(w)) < \bar{\epsilon}$, we have $w \in \mathcal{M}$.*

Proof. Suppose for contradiction that no such $\bar{\epsilon}$ exists. Let $\{\epsilon_j\}_{j\geq 1}$ be any sequence of positive numbers such that $\epsilon_j \downarrow 0$. Then for each $j \geq 1$ we have w_j such that $\|w_j - w^*\| < \epsilon_j$, $\operatorname{dist}(0, \partial\phi(w_j)) < \epsilon_j$ but $w_j \notin \mathcal{M}$. Considering the sequence $\{w_j\}_{j\geq 1}$, we have that $w_j \to w^*$, and $\operatorname{dist}(0, \partial\phi(w_j)) \to 0$. With convexity, these imply $\phi(w_j) \to \phi(w^*)$, since $\forall a_j \in \partial\phi(w_j)$ we have $\phi(w_j) - \phi(w^*) \leq a_j^T(w_j - w^*) \leq \|a_j\|\|w_j - w^*\|$. Convexity implies prox-regularity, so by applying Theorem 5.3 of Hare & Lewis (2004), we have that $w_j \in \mathcal{M}$ for all j sufficiently large, giving a contradiction. □

Recalling the quantity $\bar{r}$ from (5), we define the subsequence $\mathcal{S}$ by

$$\mathcal{S} := \Big\{ j \in \{1, 2, \dots\} \mid \mathbb{E}\left[I_{(\|w_j - w^*\| \le \bar{r})} \|w_j - w^*\|^2\right] \le j^{-1/4}, \text{ and } \mathbb{E}\left[I_{(\|w_j - w^*\| > \bar{r})} \|w_j - w^*\|\right] \le (1/\bar{r}) j^{-1/4} \Big\},$$

where $I_{(A)}$ is an indicator variable for the event A which satisfies $I_{(A)} = 1$ when the event A is true or $I_{(A)} = 0$ otherwise.

Then we can show the following properties for $j \in \mathcal{S}$.

Lemma 1. *For any $\epsilon > 0$, we have*

$$\mathbb{P}\left(\|w_j - w^*\| > \epsilon\right) < \frac{1}{\epsilon}\left(\frac{1}{\epsilon} + \frac{1}{\bar{r}}\right) j^{-1/4}, \quad \forall j \in \mathcal{S}. \tag{6}$$

Defining

$$S_t := \mathcal{S} \cap \{1, 2, \dots, t\},$$

we have

$$\frac{1}{t}\text{card}(S_t) > 1 - \frac{2}{c}\left(\gamma D^2 + \frac{G^2}{\gamma}\right) t^{-1/4}, \tag{7}$$

that is, the density of S_t in $\{1, 2, \dots, t\}$ is $1 - O(t^{-1/4})$.

Proof. See Appendix. □

Now we state the manifold identification property of the RDA algorithm.

Theorem 3. *The iterate w_j generated by the RDA algorithm belongs to the optimal manifold $\mathcal{M}$ with probability at least $1 - (\zeta_1 + \zeta_2) j^{-1/4}$ for all sufficiently large $j \in \mathcal{S}$, where*

$$\zeta_1 := \frac{3\max(1, L)}{\bar{\epsilon}}\left(\frac{3\max(1, L)}{\bar{\epsilon}} + \frac{1}{\bar{r}}\right),$$

$$\zeta_2 := 1.2(3/\bar{\epsilon})^2 \nu,$$

$$\nu := \left[L\mu + 2\sqrt{G}(G + 4L\mu)^{1/2}\right]^2, \text{ and}$$

$$\mu := \frac{1}{\sqrt{c}}\left(\gamma D^2 + \frac{G^2}{\gamma}\right)^{1/2}\left[1 + \frac{1}{\bar{r}\sqrt{c}}\left(\gamma D^2 + \frac{G^2}{\gamma}\right)^{1/2}\right].$$

Proof. Let $\bar{\epsilon} > 0$ be the threshold defined in Theorem 2. We focus on the iterates w_j for which the following event occurs:

$$E_1 : \quad \|w_j - w^*\| \le \frac{\bar{\epsilon}}{3\max(L, 1)}, \tag{8}$$

where L is the Lipschitz constant. Note that (8) trivially implies the condition $\|w_j - w^*\| \le \bar{\epsilon}$ of Theorem 2. We have from (6) that for all $j \in \mathcal{S}$,

$$\mathbb{P}(\|w_j - w^*\| \le \bar{\epsilon}) \ge \mathbb{P}(E_1) \ge 1 - \zeta_1 j^{-1/4}. \tag{9}$$

The other condition in Theorem 2 is

$$\text{dist}\left(0, \nabla f(w_j) + \partial\Psi(w_j)\right) \le \bar{\epsilon}.$$

By adding and subtracting terms, we have

$$\begin{aligned}\nabla f(w_j) + a_j = (\nabla f(w_j) - \nabla f(w^*)) + (\nabla f(w^*) - \bar{g}_{j-1}) \\ - \frac{2\gamma}{\sqrt{j-1}}(w_j - w_1) \\ + \left(\bar{g}_{j-1} + a_j + \frac{2\gamma}{\sqrt{j-1}}(w_j - w_1)\right)\end{aligned} \tag{10}$$

for any $a_j \in \partial\Psi(w_j)$. We choose the specific a_j that satisfies the optimality of the subproblem (4), that is,

$$0 = \bar{g}_{j-1} + a_j + \frac{2\gamma}{\sqrt{j-1}}(w_j - w_1).$$

This choice eliminates the last term in (10). For the other three terms, we have the following observations.

(i) For all j satisfying (8), the Lipschitz property of ∇f implies that the following event is true:

$$E_2 : \quad \|\nabla f(w_j) - \nabla f(w^*)\| \le \bar{\epsilon}/3.$$

(ii) From Theorem 11 of Lee & Wright (2011), we have for any $\epsilon > 0$ and $t \ge 1$:

$$\mathbb{P}(\|\nabla f(w^*) - \bar{g}_t\| > \epsilon) < \epsilon^{-2}\nu t^{-1/4}.$$

By setting $\epsilon = \bar{\epsilon}/3$ and $t = j - 1$, we have

$$\mathbb{P}(\|\nabla f(w^*) - \bar{g}_{j-1}\| > \bar{\epsilon}/3) < \zeta_2 j^{-1/4}, \quad \forall j \ge 2.$$

Hence, denoting by E_3 the event $\|\nabla f(w^*) - \bar{g}_{j-1}\| \le \bar{\epsilon}/3$, we have that for $j \ge 2$

$$\mathbb{P}(\neg E_3) < \zeta_2 j^{-1/4}. \tag{11}$$

(iii) For all j satisfying (8), we have

$$\begin{aligned}&2\gamma(j-1)^{-1/2}\|w_j - w_1\| \\ &\le 2\gamma(j-1)^{-1/2}(\|w_j - w^*\| + \|w_1 - w^*\|) \\ &\le 2\gamma(j-1)^{-1/2}\left(\frac{\bar{\epsilon}}{3\max(L, 1)} + D\right).\end{aligned}$$

Therefore, the event

$$E_4 : \quad 2\gamma(j-1)^{-1/2}\|w_j - w_1\| \le \bar{\epsilon}/3$$

is true whenever $j \ge j_0$, where we define j_0 by

$$j_0 := 1 + \left\lceil \frac{36\gamma^2}{\bar{\epsilon}^2}\left(\frac{\bar{\epsilon}}{3\max(L, 1)} + D\right)^2 \right\rceil.$$

Therefore for $j \in \mathcal{S}$ with $j \geq j_0$, by definition of the events E_1, E_2, E_3, and E_4 above, the probability that Theorem 2 will hold is

$$\begin{aligned}&\mathbb{P}\Big(\|w_j - w^*\| \leq \bar{\epsilon} \ \wedge \ \operatorname{dist}(0, \partial\phi(w_j)) < \bar{\epsilon}\Big)\\ &\geq \mathbb{P}\Big(E_1 \wedge E_2 \wedge E_3 \wedge E_4\Big) = \mathbb{P}(E_1 \wedge E_3)\\ &\geq 1 - \mathbb{P}(\neg E_1) - \mathbb{P}(\neg E_3) \geq 1 - (\zeta_1 + \zeta_2) j^{-1/4}\end{aligned}$$

where the last inequality is due to (9) and (11). □

Lemma 1 tells us that the sequence $\mathcal{S}$ is "dense" in $\{1, 2, \dots\}$, while Theorem 3 states that for all sufficiently large $j \in \mathcal{S}$, w_j lies on the optimal manifold with probability approaching one as j increases.

3. Dual Averaging with Manifold Identification

We present a simple strategy motivated by our analysis above, in which the RDA method gives way to a local phase after a near-optimal manifold $\mathcal{M}$ is identified.

3.1. RDA$^+$ Algorithm

Algorithm 1 summarizes our algorithm called RDA$^+$. The algorithm starts with RDA steps until it identifies a near-optimal manifold, then switches to the LPS algorithm (Wright, 2010) to search a reduced space until an optimality criterion is satisfied.

For choosing a manifold as a candidate, we use a simple heuristic inspired by Theorem 3 that if the past τ consecutive iterates have been on the same manifold, we take $\mathcal{M}$ to be approximately optimal. Before commencing the local phase, however, we "safeguard" by expanding $\mathcal{M}$ to incorporate additional dimensions that may yet contain the minimizer. Our simple approach will work provided that the expanded $\mathcal{M}$ is a *superset* of the optimal manifold, since LPS is able to move to more restricted submanifolds of $\mathcal{M}$.

3.2. Specification for ℓ_1-regularization

We describe the details of Algorithm 1 for ℓ_1-regularization ($\Psi(w) = \lambda\|w\|_1$ for some $\lambda > 0$, yielding a starting point of $w_1 = 0$). The ℓ_1-norm encourages sparsity in the solution w^*. The manifold embracing $w^* \in \mathbb{R}^n$ corresponds to the set of points in $\mathbb{R}^n$ that have the same sign and nonzero patterns as w^*.

Computation of w_{j+1}: For this regularizer, we have a closed-form solution for the subproblem (4):

$$[w_{j+1}]_i = \frac{\sqrt{j}}{2\gamma}\operatorname{soft}(-[\bar{g}_j]_i, \lambda), \quad i = 1, 2, \dots, n,$$

Algorithm 1 RDA$^+$ Algorithm.

Input: $\gamma > 0$, $\tau \in \mathbb{N}$.
Initialize: set $w_1 \in \arg\min_w \Psi(w)$ and $\bar{g}_0 = 0$.
Dual Averaging:
for $j = 1, 2, \dots$ **do**
  Compute a gradient $g_j \leftarrow \nabla F(w_j; \xi_j)$.
  Update the average gradient:
    $\bar{g}_j \leftarrow \frac{j-1}{j}\bar{g}_{j-1} + \frac{1}{j} g_j$.
  Compute the next iterate:
    $w_{j+1} \leftarrow$ the solution of the subproblem (4).
  if there is $\mathcal{M}$ such that $w_{j+2-i} \in \mathcal{M}$ for $i = 1, 2, \dots, \tau$ **then**
    Local Phase:
    Expand $\mathcal{M}$ and use LPS to search for solution on manifold $\mathcal{M}$, starting at w_{j+1};
  end if
end for

where $\operatorname{soft}(u, a) := \operatorname{sgn}(u)\max\{|u| - a, 0\}$ is the well-known soft-threshold function, where $\operatorname{sgn}(u)$ equals 1 if $u > 0$, -1 if $u < 0$, and 0 if $u = 0$.

Acceleration: To generate the approximate solution in the local phase of Algorithm 1, we use an empirical estimate $\tilde{\phi}_{\mathcal{N}}$ in (2) as a surrogate objective function and then solve

$$\min_{w \in \mathcal{M}} \ \tilde{\phi}_{\mathcal{N}}|_{\mathcal{M}}(w),$$

where $\mathcal{N}$ is drawn from available samples. LPS calculates first- and second-order information for $\tilde{\phi}_{\mathcal{N}}$ on the subset of components defined by $\mathcal{M}$. Since the intrinsic dimension of $\mathcal{M}$ is usually much smaller than the dimension n of the full space, these restricted gradients and Hessians are much cheaper to compute than their full-space counterparts.

Checking Optimality: From the optimality condition for (2), we define the optimality measure $\delta(w_j)$:

$$\delta(w_j) := \frac{1}{\sqrt{n}} \inf_{a_j \in \partial\Psi(w_j)} \|\nabla \tilde{f}_{\mathcal{N}}(w_j) + a_j\|. \tag{12}$$

Since $\delta(w^*) \approx 0$ for sufficiently large sample set $\mathcal{N}$ because of the law of large numbers, we can stop the algorithm when $\delta(w_j)$ drops below a certain threshold.

Safeguarding: For a more robust implementation, we augment $\mathcal{M}$ before starting the local phase, adding components i for which $[w_{j+1}]_i = 0$ but $[\bar{g}_j]_i$ is close to one of the endpoints of its allowable range; that is,

$$[w_{j+1}]_i = 0 \text{ and } |[\bar{g}_j]_i| > \rho\lambda \tag{13}$$

for some fixed $\rho \in (0, 1]$. This is motivated from Theorem 11 of Lee & Wright (2011), which indicates that $\bar{g}_j$ approaches $\nabla f(w^*)$ in probability as j increases.

4. Experiments

We use the MNIST data set which consists of gray-scale images of digits represented by 28×28 pixels, and focus on differentiation between the digits 6 and 7, yielding 12183 training and 1986 test examples. The digits are classified via logistic regression with ℓ_1-regularization ($\Psi(w) = \lambda\|w\|_1$ for some $\lambda > 0$). The empirical estimate $\tilde{\phi}_{\mathcal{N}}$ is taken for the full training set. For the training example selected by ξ_t at time $t \geq 1$, we use its feature vector $x_t \in \mathbb{R}^{n-1}$ and label $y_t \in \{-1, 1\}$ to define the corresponding loss function for $\tilde{w} \in \mathbb{R}^{n-1}$, $b \in \mathbb{R}$ and $w = (\tilde{w}, b)$:

$$F(w; \xi_t) = \log\Big(1 + \exp\big(-y_t(\tilde{w}^T x_t + b)\big)\Big).$$

We compare RDA$^+$ to several other algorithms: SGD, TG, RDA, and LPS. The *stochastic gradient descent* (SGD) method (for instance, Nemirovski et al., 2009) for ℓ_1-regularization consists of the iterations

$$[w_{t+1}]_i = [w_t]_i - \alpha_t\big([g_t]_i + \lambda \mathrm{sgn}\,([w_t]_i)\big),\ i = 1, \ldots, n.$$

The *truncated gradient* (TG) method (Langford et al., 2009) truncates the iterates in every K steps. That is,

$$[w_{t+1}]_i = \begin{cases} \mathrm{trnc}\,\big([w_t]_i - \alpha_t[g_t]_i,\ \lambda_t^{\mathrm{TG}}\big) & \text{if } \mathrm{mod}(t, K) = 0, \\ [w_t]_i - \alpha_t[g_t]_i & \text{otherwise,} \end{cases}$$

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

$$\mathrm{trnc}(\omega,\ \lambda_t^{\mathrm{TG}}) = \begin{cases} 0 & \text{if } |\omega| \leq \lambda_t^{\mathrm{TG}}, \\ \omega - \lambda_t^{\mathrm{TG}}\mathrm{sgn}(\omega) & \text{otherwise.} \end{cases}$$

We use $K = 10$ for enhanced regularization effect. For the stepsize α_t in SGD and TG, we adopt a variable stepsize scheme (Zinkevich, 2003; Nemirovski et al., 2009), choosing $\alpha_t = (D/G)\sqrt{2/t}$. This gives SGD a regret bound of $R_t \leq 2\sqrt{2}GD\sqrt{t}$ for $\lambda \ll G$ (as can be shown by a slight modification of Theorem 1 in Zinkevich (2003)), which is comparable to the simplified bound of RDA for $\gamma = G/D$, i.e. $R_t \leq 2GD\sqrt{t}$. We use $\gamma = 5000$ (determined by cross validation) to run RDA$^+$ and RDA, and set $\alpha_t = \gamma^{-1}\sqrt{2/t}$ for SGD and TG. For LPS and the local phase of RDA$^+$, we set the Newton threshold to 200 and use no sampling in the gradient and Hessian computation. We set $w_1 = 0$ for all algorithms.

Progress in Time: We first run RDA$^+$ with random permutations of the training samples, stopping when $\tau = 100$ consecutive iterates have the same sparsity pattern, after seeing all samples at least once. (All repeated runs required at most 19327 iterations to stop.)

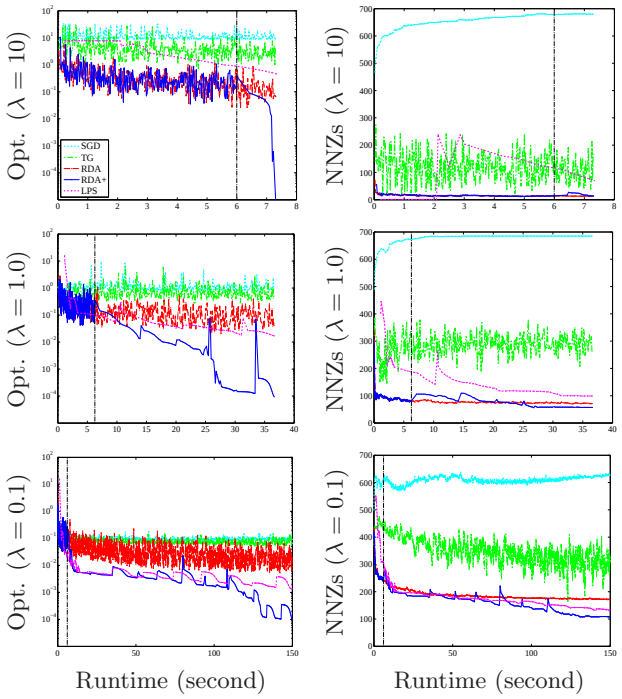

Figure 1. Convergence of iterates, measured in terms of the optimality measure (left) and the number of nonzero components in the iterates (right). SGD, TG, RDA and LPS are run up to the time taken for RDA$^+$ to achieve 10^{-4} optimality value. The black dot-dashed lines indicate the event of phase switching in RDA$^+$.

In the safeguarding test (13), we use $\rho = 0.85$. Then we run the local phase of RDA$^+$ until we obtain a solution for which the optimality measure value in (12) is less than 10^{-4}. We record the total runtime of RDA$^+$, and then run all the other algorithms up to the runtime of RDA$^+$. (They may stop earlier if they achieve the desired optimality.)

We compare the convergence speed of the algorithms, in terms of the optimality measure and the number of nonzero components. Figure 1 presents the results for three different values of λ. The optimality plots (on the left) show that RDA$^+$ achieves the target optimality much faster than other algorithms, including LPS. The RDA algorithm behaves better than SGD and TG, but it still hardly achieves the target value.

The plots on the right in Figure 1 show the number of nonzeros in the iterates. RDA tends to produce much sparser iterates with less fluctuation than SGD and TG, but it fails to reduce the number of nonzeros to the smallest number identified by RDA$^+$ in the given time, apparently for $\lambda = 1.0$ and $\lambda = 0.1$.

We mark the events of switching to the local phase for

RDA$^+$ with black dot-dashed lines. In the local phase, RDA$^+$ behaves very similarly to LPS, sharing the typical behavior of nonmonotonic decrease in optimality. However, the local phase often converges faster than LPS, because it can operate on the reduced space chosen by the initial phase of RDA$^+$. The number of nonzeros often increases in the event of switching, since the safeguarding can add more elements. This behavior can be diminished by using more conservative (larger) ρ values.

Quality of Solutions: In Figure 2, we compare the quality of the solutions in terms of optimality, the number of nonzeros, and test error rate. We run the algorithms with the same setting used in the previous experiments, except for LPS; now we run LPS without any time limit to use it as the baseline of comparison. (The runtime of LPS was about four times longer than that of RDA$^+$ on average.) The experiments are repeated for 100 different random seeds, for each of the seven λ values in the range of $[0.01, 10]$.

We can observe that only the solutions from RDA$^+$ achieve the desired optimality and the smallest number of nonzeros, with almost identical quality to the solutions from LPS. The solutions (both with and without averaging) from SGD, TG, and RDA are suboptimal, leaving much scope for zeroing out many more components of the iterates. RDA achieves a similar number of nonzeros to RDA$^+$ for large λ values, but more nonzeros on smaller values of λ. In terms of the test error rate, RDA$^+$ produces slightly better solutions than SGD, TG, and RDA overall. Although the improvement is marginal, we note that high accuracy is difficult to achieve solely with the stochastic online learning algorithms in limited time. The averaged iterates of SGD and TG show smaller test error for $\lambda \geq 1$ than others, but they need a large number of nonzero components, despite the strong regularization imposed.

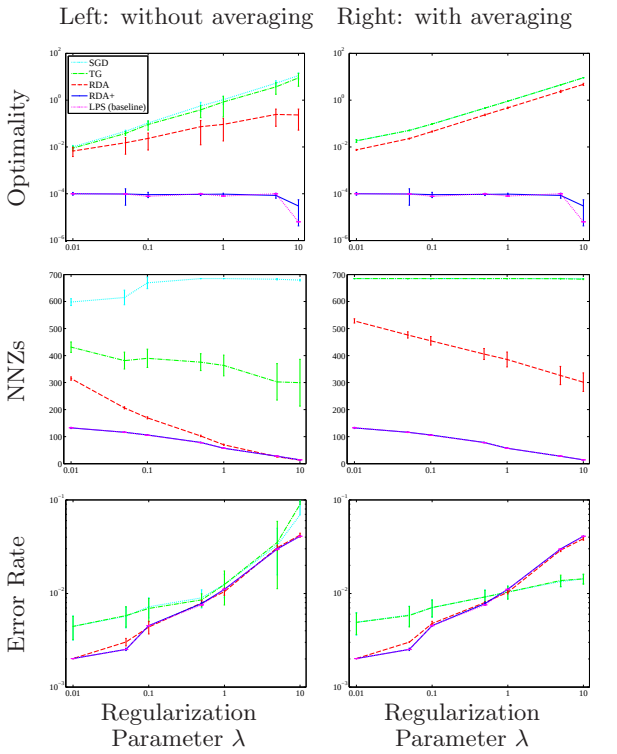

Figure 2. Quality of the solutions, in terms of the optimality, the number of nonzero components, and the test error rate (measured for 100 different random permutations). SGD, TG, and RDA are run up to the time taken for RDA$^+$ to achieve 10^{-4} optimality solutions, whereas LPS is run without such limit. The plots for RDA$^+$ and LPS on the left are duplicated to the right for comparison.

5. Conclusion

We have shown that the RDA algorithm is effective for producing solutions with a smaller set of active elements than other subgradient methods, and also identifies the optimal manifold with probability approaching one as iterations proceed. This observation enables us to apply alternative optimization techniques with faster convergence rate on the near-optimal manifold, enabling more rapid convergence to near-optimal points than plain stochastic gradient approaches.

Acknowledgements

Research supported in part by National Science Foundation Grant DMS-0914524.

References

Burke, J. V. and Moré, J. J. Exposing constraints. *SIAM Journal on Optimization*, 4(3):573–595, 1994.

Hare, W. L. and Lewis, A. S. Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis*, 11(2):251–266, 2004.

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

Lee, S. and Wright, S. J. Manifold identification of dual averaging algorithm for regularized stochastic online learning. Technical report, University of Wisconsin-Madison, April 2011.

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

Nesterov, Y. Smooth minimization of nonsmooth func-

tions. *Mathematical Programming, Series A*, 103: 127–152, 2005.

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

Shi, W., Wahba, G., Wright, S. J., Lee, K., Klein, R., and Klein, B. LASSO-Patternsearch algorithm with application to opthalmology data. *Statistics and its Interface*, 1:137–153, 2008.

Vaisman, I. *A First Course in Differential Geometry.* Monographs and Textbooks in Pure and Applied Mathematics. Marcel Dekker, 1984.

Wright, S. J. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31:1063–1079, 1993.

Wright, S. J. Accelerated block-coordinate relaxation for regularized optimization. Technical report, University of Wisconsin-Madison, August 29 2010.

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, October 2010.

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

Appendix: Proof of Lemma 1

To measure $\text{card}(\mathcal{S}_t^c)$ for $\mathcal{S}_t^c := \{1, 2, \ldots, t\} \setminus \mathcal{S}_t$, we define indicator variables for $j \geq 1$:

$$\chi_-^j := \begin{cases} 1 & \text{if } \mathbb{E}\left[I_{(\|w_j - w^*\| \leq \bar{r})}\|w_j - w^*\|^2\right] > j^{-1/4}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\chi_+^j := \begin{cases} 1 & \text{if } \mathbb{E}\left[I_{(\|w_j - w^*\| > \bar{r})}\|w_j - w^*\|\right] > (1/\bar{r})j^{-1/4}, \\ 0 & \text{otherwise.} \end{cases}$$

Since $\mathcal{S}_t^c$ contains all $j \in \{1, 2, \ldots, t\}$ that satisfy $\chi_-^j = 1$ or $\chi_+^j = 1$, $\text{card}(\mathcal{S}_t^c) \leq \sum_{j=1}^t (\chi_-^j + \chi_+^j)$. For $\sum_{j=1}^t \chi_-^j$, we note that

$$\begin{aligned} &\sum_{j=1}^t \mathbb{E}\left[I_{(\|w_j - w^*\| \leq \bar{r})}\|w_j - w^*\|^2\right] \\ &\quad \geq \sum_{j=1}^t \chi_-^j \mathbb{E}\left[I_{(\|w_j - w^*\| \leq \bar{r})}\|w_j - w^*\|^2\right] \\ &\quad > \sum_{j=1}^t \chi_-^j j^{-1/4} \geq t^{-1/4} \sum_{j=1}^t \chi_-^j. \end{aligned} \tag{14}$$

From Theorem 9 of Lee & Wright (2011), we have

$$\begin{aligned} &\frac{1}{t}\sum_{j=1}^t \mathbb{E}[I_{(\|w_j - w^*\| \leq \bar{r})}\|w_j - w^*\|^2] \\ &\qquad \leq \frac{1}{c}\left(\gamma D^2 + \frac{G^2}{\gamma}\right) t^{-1/2}. \end{aligned}$$

Applying this bound to (14), we deduce that

$$\frac{1}{t}\sum_{j=1}^t \chi_-^j \leq \frac{1}{c}\left(\gamma D^2 + \frac{G^2}{\gamma}\right) t^{-1/4}.$$

Similar arguments for the term $\sum_{j=1}^t \chi_+^j$ with $\mathbb{E}[I_{(\|w_j - w^*\| > \bar{r})}\|w_j - w^*\|]$, $j = 1, 2, \ldots, t$ lead to

$$\frac{1}{t}\sum_{j=1}^t \chi_+^j \leq \frac{1}{c}\left(\gamma D^2 + \frac{G^2}{\gamma}\right) t^{-1/4}.$$

Therefore, the fraction $\text{card}(\mathcal{S}_t)/\text{card}(\{1, 2, \ldots, t\})$ is

$$\begin{aligned} \frac{1}{t}\text{card}(\mathcal{S}_t) &= 1 - \frac{1}{t}\text{card}(\mathcal{S}_t^c) \\ &\geq 1 - \frac{1}{t}\sum_{j=1}^t (\chi_-^j + \chi_+^j) \\ &> 1 - \frac{2}{c}\left(\gamma D^2 + \frac{G^2}{\gamma}\right) t^{-1/4}, \end{aligned}$$

thus proving (7). For (6), we have for any $\epsilon > 0$ that

$$\begin{aligned} &\mathbb{P}(\|w_j - w^*\| > \epsilon) \\ &\quad = \mathbb{P}(\|w_j - w^*\| > \epsilon,\ \|w_j - w^*\| \leq \bar{r}) \\ &\quad + \mathbb{P}(\|w_j - w^*\| > \epsilon,\ \|w_j - w^*\| > \bar{r}) \end{aligned} \tag{15}$$

Focusing on the first term, we have for all $j \in \mathcal{S}$ that

$$\begin{aligned} &\mathbb{P}(\|w_j - w^*\| > \epsilon,\ \|w_j - w^*\| \leq \bar{r}) \\ &\quad = \mathbb{P}(I_{(\|w_j - w^*\| \leq \bar{r})}\|w_j - w^*\| > \epsilon) \\ &\quad < \epsilon^{-2}\mathbb{E}\left[I_{(\|w_j - w^*\| \leq \bar{r})}\|w_j - w^*\|^2\right] \\ &\quad \leq \epsilon^{-2} j^{-1/4} \end{aligned} \tag{16}$$

due to the Markov inequality and the definition of $\mathcal{S}$. Similarly for the second term in (15), we have for all $j \in \mathcal{S}$

$$\begin{aligned} &\mathbb{P}(\|w_j - w^*\| > \epsilon,\ \|w_j - w^*\| > \bar{r}) \\ &\quad < \epsilon^{-1}\mathbb{E}\left[I_{(\|w_j - w^*\| > \bar{r})}\|w_j - w^*\|\right] \\ &\quad \leq \epsilon^{-1}\bar{r}^{-1} j^{-1/4} \end{aligned} \tag{17}$$

Applying (16) and (17) to (15) results in (6). □