Journal of Machine Learning Research 13 (2012) 1705-1744 Submitted 7/11; Revised 3/12; Published 6/12

Manifold Identification in Dual Averaging for Regularized Stochastic Online Learning

Sangkyun Lee SANGKYUN.LEE@TU-DORTMUND.DE
Fakultät für Informatik, LS VIII
Technische Universität Dortmund
44221 Dortmund, Germany

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

Editor: Léon Bottou

Abstract

Iterative methods that calculate their steps from approximate subgradient directions have proved to be useful for stochastic learning problems over large and streaming data sets. When the objective consists of a loss function plus a nonsmooth regularization term, the solution often lies on a low-dimensional manifold of parameter space along which the regularizer is smooth. (When an ℓ_1 regularizer is used to induce sparsity in the solution, for example, this manifold is defined by the set of nonzero components of the parameter vector.) This paper shows that a regularized dual averaging algorithm can identify this manifold, with high probability, before reaching the solution. This observation motivates an algorithmic strategy in which, once an iterate is suspected of lying on an optimal or near-optimal manifold, we switch to a "local phase" that searches in this manifold, thus converging rapidly to a near-optimal point. Computational results are presented to verify the identification property and to illustrate the effectiveness of this approach.

Keywords: regularization, dual averaging, partly smooth manifold, manifold identification

1. Introduction

Online learning algorithms based on stochastic approximation often are effective for solving large machine learning problems. Each step of these methods evaluates an approximate subgradient of the objective at the current iterate, based on a small subset (perhaps a single item) of the data. The amount of computation and data manipulation required per iteration is therefore small. Although many iterations are needed to converge to high-accuracy solutions, the tradeoffs between optimization errors and the other errors that arise in machine learning problems suggest that solutions of moderate accuracy often suffice. However, most existing stochastic algorithms do not produce approximate solutions that have the desirable structure (such as sparsity) that motivate the use of a regularization term in the objective.

We focus on the regularized dual averaging (RDA) approach developed by Nesterov (2009) and extended by Xiao (2010) to regularized problems. We show that, under appropriate assumptions, iterates generated by this method have a structure similar to the solution (with high probability) after

a sufficiently large number of iterations. This structure is characterized by a manifold, and identification of structure corresponds to identifying the manifold on which the solution lies. We sketch an algorithmic strategy that exploits this property by switching to a "local phase" that searches on the identified manifold, which often has much lower dimension than the full space.

1.1 Problem Setting and Regularized Dual Averaging

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

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

where

$$f(w) := \mathbb{E}_\xi[F(w;\xi)] = \int_\Xi F(w;\xi)dP(\xi), \tag{2}$$

and ξ is a random vector whose probability distribution P is supported on the set $\Xi \subset \mathbb{R}^d$. The regularization function $\Psi : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is assumed to be a closed proper convex function whose effective domain (denoted by $\operatorname{dom}\Psi$) is closed, and there is an open neighborhood $\mathcal{O}$ of $\operatorname{dom}\Psi$ that is contained in the domain of $F(\cdot,\xi)$, for all $\xi \in \Xi$. We assume that the expectation in (2) is well defined and finite-valued for all $w \in \mathcal{O}$ and that for every $\xi \in \Xi$, $F(w,\xi)$ is a smooth convex function of $w \in \mathcal{O}$. (An elementary argument shows that f is therefore convex.) We use w^* to denote a minimizer of (1).

The purpose of the regularizer Ψ is to promote certain desirable types of structure in the solution of (1). A common desirable property is *sparsity* (that is, w^* has few nonzero elements), which can be promoted by setting $\Psi(\cdot) = \lambda\|\cdot\|_1$ for some parameter $\lambda > 0$.

One method for finding an (approximate) solution to (1) is to draw random variables ξ_j for all $j \in \mathcal{N}$ independently from the space Ξ, where $\mathcal{N}$ is an index set of finite cardinality, and solve a *sample average approximation* (SAA) problem

$$\min_{w\in\mathbb{R}^n} \tilde{\phi}_{\mathcal{N}}(w) := \tilde{f}_{\mathcal{N}}(w) + \Psi(w) \tag{3}$$

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

Iteration t of a stochastic online learning approach examines a cost function $F(\cdot;\xi_t) : \mathbb{R}^n \to \mathbb{R}$ for some $\xi_t \in \Xi$, drawn randomly according to the distribution P, where $\{\xi_t\}_{t\geq 1}$ forms an i.i.d. sequence of samples. The next iterate w_{t+1} is obtained from information gathered up to the time t, the aim being to generate a sequence $\{w_t\}$ such that

$$\lim_{t\to\infty} \mathbb{E}[F(w_t;\xi)] + \Psi(w_t) = f(w^*) + \Psi(w^*). \tag{4}$$

We focus on objectives that consist of a smooth loss function F in conjunction with a nonsmooth regularizer Ψ. Xiao (2010) recently developed the *regularized dual averaging* (RDA) method, in which the smooth term is approximated by an averaged gradient in the subproblem at each iteration, while the regularization term appears explicitly. Xiao's approach extends the method of Nesterov (2009) in the sense that the regularization term is not handled explicitly in Nesterov's paper. Specifically, the RDA subproblem is

$$w_{t+1} = \arg\min_{w\in\mathbb{R}^n} \left\{ \langle \bar{g}_t, w\rangle + \Psi(w) + \frac{\beta_t}{t} h(w) \right\}, \tag{5}$$

where $\bar{g}_t = \frac{1}{t}\sum_{j=1}^{t} g_j$ and $g_j \in \partial F(w_j;\xi_j)$. The *prox-function* $h(\cdot)$ is a proper strongly convex function whose minimizers are also minimizers of Ψ, and for which the starting point w_1 of the algorithm is a minimizer of h. (The function $h(\cdot) = \|\cdot - w_1\|^2$ is a case of particular interest.) The sequence $\{\beta_t\}_{t\geq 1}$ is nonnegative.

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. When 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 from the one used on the full space. One local-phase algorithm that is well suited to the ℓ_1 regularizer is the *LASSO-patternsearch* (LPS) (Shi et al., 2008; Wright, 2012), a batch optimization method for ℓ_1-regularized logistic regression, which takes gradient steps in the space of nonzero variables, enhanced by Newton-like scaling. In logistic regression, as in other applications, it can be much less expensive to compute first- and second-order information on a restricted subspace than on the full space.

A second motivation for aiming to identify the optimal manifold is that for problems of large dimension, it may be quite expensive even to *store* a single iterate w_t, whereas an iterate whose structure is similar to that of the solution w^* may be stored economically. (To take the case of ℓ_1 regularization again, we would need to store only the nonzero elements of w_t and their locations.)

In this paper, we investigate the ability of the RDA algorithm to identify the optimal manifold. We also describe an enhanced, two-phase version of this algorithm, which we call RDA$^+$. We test this approach on ℓ_1-regularized logistic regression problems, in which an LPS-based algorithm is used to explore the near-optimal manifold identified by RDA.

1.2 Optimal Manifolds and the Identification

Identification of optimal manifolds has been studied in the context of constrained convex optimization (Wright, 1993; Burke and Moré, 1994) and nonsmooth nonconvex optimization (Hare and Lewis, 2004). In constrained optimization, the optimal manifold is typically a face or edge of the feasible set that contains the solution. In nonsmooth optimization, the optimal manifold is a smooth surface passing through the optimum, parameterizable by relatively few variables, such that the restriction of the objective function to the manifold is smooth. In either case, 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. Lewis and Wright (2008) analyze a framework for composite minimization (which uses a subproblem in which f is replaced by an exact linearization around w_t) and prove identification results. Part of the motivation for the current paper is to obtain similar results as in Lewis and Wright (2008) in the stochastic gradient setting.

1.3 Alternative Stochastic Approximation Approaches

Stochastic approximation algorithms have a rich history and are currently the focus of intense research in the machine learning and optimization communities. We mention a few relevant developments here, and discuss their manifold identification properties.

Stochastic approximation methods often solve formulations in which an explicit constraint $w \in \mathcal{W}$ (for a compact convex set $\mathcal{W}$) is present. These can be incorporated into the framework (1) by

defining the regularization function Ψ as follows:

$$\Psi(w) := \delta_{\mathcal{W}}(w) + \psi(w),$$

where $\delta_{\mathcal{W}}(w)$ is the indicator function (zero on $\mathcal{W}$ and $+\infty$ elsewhere) and ψ is a convex function whose domain includes $\mathcal{W}$. (In this setting, $\mathcal{O}$ is taken to be an open neighborhood of $\mathcal{W}$.) The classical approaches to solve such problems can be traced back to Robbins and Monro (1951) and Kiefer and Wolfowitz (1952). The *stochastic gradient descent* (SGD) method generates its iterates by stepping in the direction of a subgradient estimate and then projecting onto $\mathcal{W}$, as follows:

$$w_{t+1} = \Pi_{\mathcal{W}}\left(w_t - \alpha_t(g_t + \kappa_t)\right), \ \ t = 1, 2, \ldots$$

where $g_t \in \partial F(w_t; \xi_t)$ for some sampled random variable ξ_t, $\kappa_t \in \partial\psi(w_t)$, and $\Pi_{\mathcal{W}}(\cdot)$ denotes the Euclidean projection onto the set $\mathcal{W}$. With steplength choice of the form $\alpha_t = \theta/t$ (for a well-chosen constant θ), the SGD method exhibits $O(1/t)$ convergence rate in the quantity (4) for strongly convex functions f (Chung, 1954; Sacks, 1958). For general convex functions, a step length of the form $\alpha_t = \theta'/\sqrt{t}$ (for some $\theta' > 0$) yields an $O(1/\sqrt{t})$ rate of convergence in the function value (Nemirovski and Yudin, 1978; Polyak, 1990; Polyak and Juditsky, 1992). Simplified proofs of these rates are given by Nemirovski et al. (2009). These rates are known to be optimal for subgradient schemes in "black-box" algorithmic models (Nemirovski and Yudin, 1983). Although batch optimization methods based on an approximate objective, such as (3), may provide better convergence rates, SGD has been widely used in machine learning because it scales well to large data sets and provides good generalization performance in practice (Zinkevich, 2003; Bottou, 2004; Zhang, 2004; Shalev-Shwartz et al., 2007).

As discussed by Xiao (2010), the SGD method does not exploit the problem structure that is present in the regularizer ψ, and there is no reason to believe that these algorithms have good manifold identification properties. When $\psi(\cdot) = \|\cdot\|_1$, for instance, there is no particular reason for the iterates w_t to be sparse. Though equal in expectation, g_t and $\nabla f(x_t)$ may be far apart, so that even if a careful choice of κ_t is made at each iteration, the updates may have the cumulative effect of destroying sparsity in the iterates w_t.

Variants of SGD for the general convex case often work with averaged primal iterates, of the form

$$\bar{w}_t := \frac{\sum_{j=1}^{t} \nu_j w_j}{\sum_{j=1}^{t} \nu_j}, \tag{6}$$

where the ν_j are nonnegative weights (see, for example, Nemirovski et al., 2009). Averaging does not improve identification properties. For ℓ_1 regularization, we can still expect the averaged iterates $\bar{w}_t$ to be at least as dense as the "raw" iterates w_t.

Recently, various authors have proposed modifications of SGD that account for the regularization structure. Some representative examples include *composite objective mirror descent* (COMID) (Duchi et al., 2010), *forward-backward splitting* (FOBOS) (Duchi and Singer, 2009), *truncated gradient* (TG) (Langford et al., 2009), and *sparsity-preserving stochastic gradient* (SSG) (Lin et al., 2011). The basic FOBOS subproblem is similar to that of the prox-descent algorithm of Lewis and Wright (2008) and the SpaRSA framework of Wright et al. (2009), which generate their iterates by setting $g_t = \nabla f(w_t)$ and solving

$$w_{t+1} = \arg\min_{w \in \mathbb{R}^n} \left\{ \langle g_t, w\rangle + \Psi(w) + \frac{1}{2\alpha_t}\|w - w_t\|_2^2 \right\}, \tag{7}$$

for some parameter $\alpha_t > 0$ (which plays a step-length-like role). Duchi and Singer (2009) suggest an extension to an online setting, in which g_t is replaced by an approximate gradient. SSG (Lin et al., 2011) also has the subproblem (7) at its core, but embeds it in a strategy for generating *three* sequences of iterates (rather than the single sequence $\{w_t\}$), extending an idea of Nesterov (2004). The TG method (Langford et al., 2009) solves a subproblem like (7) on some iterations; on other iterations it simply steps in the direction g_t of the latest gradient estimate for f, ignoring the regularization term. COMID (Duchi et al., 2010) is also based on a subproblem of the form (7), but with a Bregman divergence replacing the final quadratic term, thus yielding a more general framework.

Since all these methods make explicit use of the regularization term Ψ in the subproblem, they are more likely to generate iteration sequences that share the structure of the solution w^*, that is, to identify a near-optimal manifold. Such behavior is far from guaranteed, however, because g_t may be only a rough approximation to $\nabla f(w_t)$. The inaccuracy of this gradient estimate may cause the iterates to step away from the optimal manifold, even from an iterate w_t that is close to the solution w^*. (In Example 1 at the end of Section 4, we discuss a function satisfying all the assumptions of this paper, in which the subproblem (7) steps away from the optimal point and off the optimal manifold.) In contrast, the dual average $\bar{g}_t$ used by the RDA subproblem stabilizes around $\nabla f(w^*)$ as the iterates converge to w^*, allowing identification results to be derived from analysis like that of Hare and Lewis (2004) and Lewis and Wright (2008), suitably generalized to the stochastic setting.

Averaging of the primal iterates (6) does not improve the identification properties of the methods above, and will usually make them worse. Considering again ℓ_1 regularization, we observe that if component i of *any* iterate w_t is nonzero, then component i of all averaged iterates at subsequent iterations will also be nonzero (unless some fortuitous cancellation occurs). RDA itself, in the version recommended by the analysis of Xiao (2010), has the same deficiency, as the main convergence results in that paper are for averaged iterates (6). In the current paper, we facilitate the use of the raw iterates w_t by RDA by adding two assumptions. First, w^* is assumed to be a *strong* minimizer of the restriction of ϕ to the optimal manifold. Second, a nondegeneracy condition ensures that ϕ increases sharply as we move off the manifold. Together, these conditions ensure that w^* is a strong minimizer of ϕ, so convergence of $\phi(w_t)$ to $\phi(w^*)$ forces convergence of w_t to w^*.

Convergence analysis of stochastic approximation algorithms focuses largely on the *regret*, which is defined as follows, for any instantiation of the random sequence $\{w_t\}_{t\geq 1}$ with respect to any fixed decision $w \in \operatorname{dom}\Psi$:

$$R_t(w) := \sum_{j=1}^{t} [F(w_j;\xi_j) + \Psi(w_j)] - \sum_{j=1}^{t} [F(w;\xi_j) + \Psi(w)]. \tag{8}$$

As we discuss later, the RDA algorithm has $O(\sqrt{t})$ regret bounds when $\beta_t = O(\sqrt{t})$ for general convex cases, and $O(\ln t)$ bounds with $\beta_t = O(\ln t)$ for strongly convex cases. These bounds are comparable to those of the SGD method. For general convex cases, when we use $\alpha_t = O(1/\sqrt{t})$, SGD achieves an $O(\sqrt{t})$ regret bound (see, for example, Zinkevich, 2003; Nemirovski et al., 2009), which cannot be improved in general. For the strongly convex case, SGD has an $O(\ln t)$ regret bound (see, for example, Hazan et al., 2006; Bartlett et al., 2008) with $\alpha_t = O(1/t)$.

1.4 Notation and Terminology

Throughout the paper, 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 $\mathrm{dom}\,\Psi := \{w \in \mathbb{R}^n \,|\, \Psi(w) < +\infty\}$. 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).

We call a function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *strongly convex* if there exists a constant $\sigma > 0$ such that $\forall w, w' \in \mathrm{dom}\,\varphi$ and $\forall \alpha \in [0,1]$,

$$\varphi(\alpha w + (1-\alpha)w') \le \alpha\varphi(w) + (1-\alpha)\varphi(w') - \frac{\sigma}{2}\alpha(1-\alpha)\|w-w'\|^2.$$

(σ is known as the *modulus of convexity*.) Strong convexity implies that for any $w \in \mathrm{dom}\,\varphi$ and $w' \in \mathrm{ri}\,\mathrm{dom}\,\varphi$, we have

$$\varphi(w) \ge \varphi(w') + \langle s, w-w'\rangle + \frac{\sigma}{2}\|w-w'\|^2, \quad \forall s \in \partial\varphi(w'). \tag{9}$$

We say that a function φ has a *locally strong minimizer* at w^* if there exist positive constants c and $\bar{r}$ such that

$$\varphi(w) - \varphi(w^*) \ge c\|w-w^*\|^2, \text{ for all } w \in \mathcal{O} \text{ with } \|w-w^*\| \le \bar{r}.$$

w^* is a *globally strong minimizer* if this expression is true with $\bar{r} = \infty$.

The algorithm we consider in this paper makes use of an i.i.d. sequence $\{\xi_j\}_{j\ge 1}$ of random variables drawn from Ξ according to the distribution P. We denote the history of random variables up to time t by

$$\xi_{[t]} := \{\xi_1, \xi_2, \ldots, \xi_t\}.$$

The iterate w_t produced by the algorithm depends on $\xi_1, \xi_2, \ldots, \xi_{t-1}$ but not on $\xi_t, \xi_{t+1}, \cdots$; we sometimes emphasize this fact by writing $w_t = w_t(\xi_{[t-1]})$.

2. Assumptions and Basic Results

We summarize here our fundamental assumptions about the problem and its solution, together with some basic observations and results that will be used in later sections.

2.1 Unbiasedness

As in Nemirovski et al. (2009), we assume the following *unbiasedness* property:

$$\nabla f(w) = \nabla_w \mathbb{E}[F(w;\xi)] = \mathbb{E}[\nabla_w F(w;\xi)] \tag{10}$$

for any w independent of ξ. (As the differentiation of F is taken only for its first argument, we omit the subscript "w" in subsequent discussions.) Given that $w_t = w_t(\xi_{[t-1]})$, this implies

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

2.2 Uniform Lipschitz Continuity

We assume that each $F(w;\xi)$ is a smooth convex function of $w \in \mathcal{O}$ for every $\xi \in \Xi$, and in particular that $\nabla F(\cdot;\xi)$ is uniformly Lipschitz continuous with respect to its first argument, over all ξ. That is, there exists a constant $L > 0$ such that

$$\|\nabla F(w;\xi) - \nabla F(w';\xi)\| \leq L\|w - w'\|, \ \ \forall w, w' \in \mathcal{O}, \ \ \forall \xi \in \Xi. \tag{11}$$

We further assume that there exists a uniform bound G for which

$$\|\nabla F(w;\xi)\| \leq G, \ \ \forall w \in \mathcal{O}, \ \forall \xi \in \Xi. \tag{12}$$

These assumptions immediately lead to similar properties on ∇f. We prove this claim after noting a simple consequence of Jensen's inequality, whose proof is omitted.

Lemma 1 *For a vector-valued function $v : \Xi \to \mathbb{R}^n$ which is integrable with respect to P, we have*

$$\|\mathbb{E}v\|_2 \leq \mathbb{E}\|v\|_2.$$

Lemma 2 *If $\nabla F(w;\xi)$ satisfies the uniform Lipschitz continuity assumption* (11)*, then $\nabla f(w)$ is also uniformly Lipschitz continuous on $\mathcal{O}$ with the same constant L. If $\nabla F(w;\xi)$ satisfies the uniform bound* (12)*, then ∇f satisfies the same bound, that is, $\|\nabla f(x)\| \leq G$ for all $w \in \mathcal{O}$.*

Proof From unbiasedness, we have for $w, w' \in \mathcal{O}$ independent of ξ that

$$\begin{aligned}
\nabla f(w) &= \nabla \mathbb{E}[F(w;\xi)] = \mathbb{E}[\nabla F(w;\xi)] && \text{from (10)} \\
&= \mathbb{E}[\nabla F(w';\xi) + u_\xi] && \text{for } u_\xi := \nabla F(w;\xi) - \nabla F(w';\xi) \\
&= \nabla f(w') + \mathbb{E}[u_\xi] && \text{from (10) again.}
\end{aligned}$$

Since $\|u_\xi\| \leq L\|w - w'\|$, we have

$$\|\nabla f(w) - \nabla f(w')\| = \|\mathbb{E}u_\xi\| \leq \mathbb{E}\|u_\xi\| \leq L\|w - w'\|,$$

where the first inequality is due to Lemma 1. This proves the first claim.

For the second claim, we have

$$\|\nabla f(w)\| = \|\mathbb{E}[\nabla F(w;\xi)]\| \leq \mathbb{E}\|\nabla F(w;\xi)\| \leq G,$$

as required. ■

2.3 Optimality and Nondegeneracy

We specify several optimality conditions that are assumed to hold throughout the paper. The optimality of w^* for the problem (1) can be characterized as follows:

$$0 \in \nabla f(w^*) + \partial\Psi(w^*).$$

We assume that w^* is a *nondegenerate* solution—one that satisfies the stronger condition

$$0 \in \text{ri } [\nabla f(w^*) + \partial\Psi(w^*)]. \tag{13}$$

The nondegeneracy assumption is common in manifold identification analyses. It can be replaced with weaker assumptions to prove weaker results (see, for instance, Oberlin and Wright, 2006), though the analysis is somewhat more complicated.

2.4 Manifolds and Partial Smoothness

In this section we discuss properties of differential manifolds and partial smoothness by repeating some definitions from Hare and Lewis (2004).

Definition 3 (Manifold) *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 $\mathcal{C}^p$ functions ($p \geq 2$) 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 $H(z) = 0$.*

The *normal space* to $\mathcal{M}$ at $\bar{z}$, denoted by $N_{\mathcal{M}}(\bar{z})$, is the range space of $\nabla H(\bar{z})$, while the *tangent space* to $\mathcal{M}$ at $\bar{z}$ is the null space of $\nabla H(\bar{z})^T$. We assume without loss of generality that $\nabla H(\bar{z})$ has orthonormal columns.

We define *partial smoothness* as follows (Lewis, 2003, Section 2).

Definition 4 (Partial Smoothness) *A function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is ($\mathcal{C}^2$-)* partly smooth *at a point $\bar{z} \in \mathbb{R}^n$ relative to a set $\mathcal{M} \subset \mathbb{R}^n$ containing $\bar{z}$ if $\mathcal{M}$ is a manifold about $\bar{z}$ and the following properties hold:*

(i) (Smoothness) *The function φ restricted to $\mathcal{M}$, denoted by $\varphi|_{\mathcal{M}}$, is $\mathcal{C}^2$ near $\bar{z}$,*

(ii) (Regularity) *φ is subdifferentially regular at all points $z \in \mathcal{M}$ near $\bar{z}$, with $\partial\varphi(z) \neq \emptyset$,*

(iii) (Sharpness) *The affine span of $\partial\varphi(\bar{z})$ is a translate of $N_{\mathcal{M}}(\bar{z})$, and*

(iv) (Sub-continuity) *The set-valued mapping $\partial\varphi : \mathcal{M} \rightrightarrows \mathbb{R}^n$ is continuous at $\bar{z}$.*

We refer to $\mathcal{M}$ as the active manifold *at the point $\bar{z}$, and as the* optimal manifold *when $\bar{z} = w^*$, where w^* is a solution of* (1).

The regularity condition (ii) is discussed by Lewis (2003, p. 706); for our purposes it suffices to note that this condition holds for closed convex functions and continuously differentiable functions, and hence for our objective ϕ. Henceforth, we assume that Ψ is partly smooth at w^* relative to the optimal manifold, which implies partial smoothness of ϕ (by an argument like that of Lemma 2).

We discuss the concepts outlined in the definitions above for the specific case of $\Psi(\cdot) = \|\cdot\|_1$. Given $\bar{z} \in \mathbb{R}^n$, we define the following partition of the indices $\{1, 2, \ldots, n\}$:

$$\{1, 2, \ldots, n\} = \mathcal{P} \cup \mathcal{N} \cup \mathcal{Z},$$

where $\bar{z}_i = 0$ for all $i \in \mathcal{Z}$, $\bar{z}_i > 0$ for all $i \in \mathcal{P}$, and $\bar{z}_i < 0$ for all $i \in \mathcal{N}$. A natural definition of the active manifold $\mathcal{M}$ is thus

$$\mathcal{M} = \{z \in \mathbb{R}^n \,|\, z_i = 0 \text{ for all } i \in \mathcal{Z}\}.$$

Note that $\bar{z} \in \mathcal{M}$, and that the mapping H of Definition 3 can be defined as $H(z) = [z_i]_{i \in \mathcal{Z}}$, with $k = \text{card}(\mathcal{Z})$ in that definition. The restriction of Ψ to this manifold thus has the following form for all $z \in \mathcal{M}$ near $\bar{z}$:

$$-\sum_{i \in \mathcal{N}} z_i + \sum_{i \in \mathcal{P}} z_i,$$

so that (i) of Definition 4 is satisfied. For $z \in \mathcal{M}$ near $\bar{z}$, we have

$$[\partial\phi(z)]_i = [\nabla f(z)]_i + \begin{cases} [-1,1] & \text{for } i \in \mathcal{Z}, \\ -1 & \text{for } i \in \mathcal{N}, \\ +1 & \text{for } i \in \mathcal{P}. \end{cases}$$

Clearly, condition (ii) of Definition 4 holds. The affine span of $\partial\phi(\bar{z})$ is

$$\{\nabla f(\bar{z}) + g \,|\, g_i = -1 \text{ for } i \in \mathcal{N},\ g_i = +1 \text{ for } i \in \mathcal{P}, \text{ and } g_i \in \mathbb{R} \text{ for } i \in \mathcal{Z}\},$$

whereas $N_{\mathcal{M}}(\bar{z}) = \{g \,|\, g_i = 0 \text{ for } i \in \mathcal{P} \cup \mathcal{N},\ g_i \in \mathbb{R} \text{ for } i \in \mathcal{Z}\}$. Comparison of the last two expressions shows that (iii) of Definition 4 is satisfied. Finally, the set-valued map $\partial\Psi$ is in fact constant along $\mathcal{M}$ near $\bar{z}$, so because f is smooth, Definition 4 (iv) also holds.

2.5 Strong Minimizer Properties

We assume that w^* is a *locally strong minimizer* of ϕ relative to the optimal manifold $\mathcal{M}$ with modulus $c_{\mathcal{M}}$, that is, there exists $c_{\mathcal{M}} > 0$ and $r_{\mathcal{M}} > 0$ such that $\{w \in \mathbb{R}^n \,|\, \|w - w^*\| \leq r_{\mathcal{M}}\} \subset \mathcal{O}$ and

$$\phi|_{\mathcal{M}}(w) \geq \phi|_{\mathcal{M}}(w^*) + c_{\mathcal{M}}\|w - w^*\|^2, \text{ for all } w \in \mathcal{M} \text{ with } \|w - w^*\| \leq r_{\mathcal{M}}. \tag{14}$$

Together with a nondegeneracy assumption (13), these conditions imply that w^* is a locally strong minimizer of $\phi(w)$ (without restriction to the optimal manifold), as we now prove.

Theorem 5 (Strong Minimizer for General Convex Case) *Suppose that ϕ is partly smooth at w^* relative to the optimal manifold $\mathcal{M}$, that w^* is a locally strong minimizer of $\phi|_{\mathcal{M}}$ with the modulus $c_{\mathcal{M}} > 0$ and radius $r_{\mathcal{M}} > 0$ defined in* (14)*, and that the nondegeneracy condition* (13) *holds at w^*. Then there exist $c \in (0, c_{\mathcal{M}}]$ and $\bar{r} \in (0, r_{\mathcal{M}}]$ such that*

$$\phi(w) - \phi(w^*) \geq c\|w - w^*\|^2, \text{ for all } w \text{ with } \|w - w^*\| \leq \bar{r}, \tag{15}$$

that is, w^* is a locally strong minimizer of ϕ, without restriction to the manifold $\mathcal{M}$.

Proof The proof is a simplification of the proof of Wright (2012, Theorem 2.5), which considers the more general case in which f is prox-regular rather than convex. For completeness, we present the proof in Appendix A. ■

Henceforth, we assume without loss of generality that $\bar{r} \in (0, 1]$.

The condition (15) is similar to the quadratic growth condition proposed by Anitescu (2000) in the context of nonlinear programming. It was shown by Anitescu that this fundamental condition is weaker than many other second-order conditions that are widely used in nonlinear programming.

Two corollaries follow in a straightforward fashion from the theorem above. We state these results here and give their proofs in Appendix A.

Corollary 6 *Suppose that w^* is a locally strong minimizer of* (1) *that satisfies* (15)*. For all $w \in \mathcal{O}$ with $\|w - w^*\| > \bar{r}$, we have*

$$\phi(w) - \phi(w^*) > c\bar{r}\|w - w^*\|.$$

Corollary 7 (Globally Strong Minimizer for Strongly Convex Case) *Suppose that w^* is a locally strong minimizer of* (1) *satisfying* (15)*. If ϕ is strongly convex on* $\operatorname{dom}\Psi$ *with modulus $\sigma > 0$, then w^* is a globally strong minimizer of* (1) *with modulus* $\min(c, \sigma/2)$*, that is,*

$$\phi(w) \geq \phi(w^*) + \min(c, \sigma/2)\|w - w^*\|^2, \text{ for all } w \in \mathcal{O}. \tag{16}$$

2.6 Summary of Assumptions

We summarize here the assumptions introduced in this section, for reference in the remainder of the paper.

The first assumption summarizes basic properties of the functions and the minimizer.

Assumption 1 *The unbiasedness property* (10)*, uniform Lipschitz continuity of* $\nabla F(\cdot;\xi)$ *for all* $\xi \in \Xi$ (11)*, uniform boundedness of* $\|\nabla F(\cdot,\xi)\|$ (12)*, and nondegeneracy at the optimum* w^* (13) *are satisfied.*

The second assumption provides sufficient conditions for w^* to be a locally strong minimizer.

Assumption 2 *The function* ϕ *is partly smooth at its minimizer* w^* *relative to the optimal manifold* $\mathcal{M}$ *and* w^* *is a locally strong minimizer of* $\phi|_{\mathcal{M}}$ *as defined in* (14)*.*

3. Regularized Dual Averaging Algorithm

We start this section by describing regret bounds for the regularized dual averaging (RDA) algorithm of Xiao (2010) (following Nesterov, 2009), focusing on its stochastic variant. We also describe the consequences for the analysis of the condition that the minimum is strong locally (15) or globally (16). We then analyze the properties of the averaged gradient; this analysis forms the basis of the manifold identification result in Section 4.

3.1 The RDA Algorithm

We start by specifying the RDA algorithm from Xiao (2010), noting that our assumptions on the functions F, f, and Ψ from Section 1 are stronger than the corresponding conditions in Xiao (2010), which require only subdifferentiability of $F(w;\xi_t)$ on $\operatorname{dom}\Psi$.

As introduced in (5), the prox-function $h:\mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is proper, strongly convex on $\operatorname{dom}\Psi$, and subdifferentiable on $\operatorname{ri}\operatorname{dom}\Psi$. In addition, we require h to satisfy

$$\arg\min_w h(w) \in \arg\min_w \Psi(w).$$

The *prox-center* w_1 of $\operatorname{dom}\Psi$ with respect to h, which is used as the starting point of the RDA method, is defined as follows:

$$w_1 := \arg\min_{w\in\operatorname{dom}\Psi} h(w).$$

The terms "prox-function" and "prox-center" are borrowed from Nesterov (2009). Following Xiao (2010), we assume without loss of generality that the strong convexity modulus of $h(w)$ is one, and that $\min_w h(w) = \min_w \Psi(w) = 0$. This implies $h(w_1) = 0$ and $\Psi(w_1) = 0$. The most obvious prox-function is $h(\cdot) = \frac{1}{2}\|\cdot - w_1\|^2$, where $w_1 \in \arg\min_w \Psi(w)$.

We now define a constant D that reappears throughout the analysis. For any $D > 0$, we consider a level set of the prox-function h defined as follows:

$$\mathcal{F}_D := \{w \in \operatorname{dom}\Psi \mid h(w) \le D^2\}.$$

We assume in the analysis that points of interest (specifically, w^*) lie in $\mathcal{F}_D$. Because of (9) and our assumptions on h (in particular, $h(w_1) = 0$, $0 \in \partial h(w_1)$, and h has modulus of convexity 1), we have that

$$w \in \mathcal{F}_D \;\Rightarrow\; D^2 \ge h(w) \ge \frac{1}{2}\|w - w_1\|^2 \;\Rightarrow\; \|w - w_1\| \le \sqrt{2}D. \tag{17}$$

Algorithm 1: The RDA Algorithm.

Input:

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

$$\arg\min_{w\in\mathbb{R}^n} h(w) \in \arg\min_{w\in\mathbb{R}^n} \Psi(w).$$

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

Initialize: set $w_1 = \arg\min h(w)$ and $\bar{g}_0 = 0$

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

Sample ξ_t from Ξ and compute a gradient $g_t = \nabla F(w_t; \xi_t)$;

Update the average gradient:

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

Compute the next iterate:

$$w_{t+1} = \arg\min_{w\in\mathbb{R}^n} \left\{ \langle \bar{g}_t, w\rangle + \Psi(w) + \frac{\beta_t}{t}h(w) \right\}. \tag{18}$$

end

At iteration t, the stochastic RDA algorithm samples a vector $\xi_t \in \Xi$, according to the distribution P, and evaluates an approximate gradient as follows:

$$g_t := \nabla F(w_t; \xi_t).$$

We assume that the random variables ξ_t are i.i.d. The *dual average*—an averaged approximation to the gradient of f—is defined as follows:

$$\bar{g}_t := \frac{1}{t}\sum_{j=1}^{t} g_j = \frac{1}{t}\sum_{j=1}^{t} \nabla F(w_j; \xi_j). \tag{19}$$

The RDA algorithm is specified in Algorithm 1. As the objective function in the subproblem (18) is strongly convex when $\beta_t > 0$ or when $\Psi(\cdot)$ is strongly convex (at least one of which we assume for the analysis below), w_{t+1} is uniquely defined by (18). Note that w_{t+1} depends on the history of random variables $\xi_{[t]}$ up to iteration t, and is independent of later samples $\xi_{t+1}, \xi_{t+2}, \ldots$.

We consider two choices of parameter sequences $\{\beta_t\}$ in the remainder of the paper, depending on whether the regularization function Ψ is strongly convex or not. The first choice holds for general convex regularizers Ψ:

$$\beta_t = \gamma\sqrt{t}, \quad \forall t \geq 1, \text{ for some constant } \gamma > 0. \tag{20}$$

The second choice, $\beta_t = O(1+\ln t)$, can be used when Ψ is a strongly convex function, with modulus $\sigma > 0$. Among the three choices discussed in Xiao (2010), that is, $\beta_t = \sigma$, $\beta_t = \sigma(1+\ln t)$, and $\beta_t = 0$, we focus on the zero sequence, which gives the simplest bounds:

$$\beta_t = 0, \quad \forall t \geq 1. \tag{21}$$

In this case, the subproblem (18) in Algorithm RDA has no damping term; we rely instead on the damping effect supplied by the strong convexity of Ψ to stabilize the iterates.

3.2 Bounds on Regret and Expected Errors in the Iterations

Our first key result concerns bounds on the regret function defined in (8).

Theorem 8 *Suppose that the unbiasedness and uniform boundedness properties in Assumption 1 are satisfied. When $\{\beta_t\}$ is chosen according to* (20)*, we have for any $w \in \mathcal{F}_D$ that*

$$R_t(w) \leq \left(\gamma D^2 + \frac{G^2}{\gamma}\right)\sqrt{t}, \qquad t \geq 1. \tag{22}$$

Moreover, when $\Psi(w)$ is strongly convex with the modulus $\sigma > 0$, then the choice (21) *for $\{\beta_t\}$ results in the following bound for $w \in \mathcal{F}_D$:*

$$R_t(w) \leq \frac{G^2}{2\sigma}(6 + \ln t), \qquad t \geq 1. \tag{23}$$

Proof See Xiao (2010, Corollary 2) for the general convex case, and Xiao (2010, Theorem 1 and Section 3.2) for the strongly convex case. ■

The next result obtains bounds on the expected errors in the iterates generated by Algorithm 1. For the purpose of this and future results, we define the indicator function $I_{(A)}$ for the event A as follows

$$I_{(A)} = \begin{cases} 1 & \text{if event } A \text{ is true,} \\ 0 & \text{if event } A \text{ is false.} \end{cases}$$

For a random event A, $I_{(A)}$ is a random variable.

Lemma 9 (Expected Error Bounds of Iterates) *Suppose that Assumptions 1 and 2 are satisfied, and that $w^* \in \mathcal{F}_D$. Then for the iterates $w_1, w_2, \ldots, w_t$ generated by the stochastic RDA algorithm with $\{\beta_t\}$ chosen by* (20)*, we have*

$$\frac{1}{t}\sum_{j=1}^{t} \mathbb{E}\left[I_{(\|w_j - w^*\| \leq \bar{r})}\|w_j - w^*\|^2\right] \leq \frac{1}{c}\left(\gamma D^2 + \frac{G^2}{\gamma}\right) t^{-1/2}, \tag{24}$$

$$\frac{1}{t}\sum_{j=1}^{t} \mathbb{E}\left[I_{(\|w_j - w^*\| > \bar{r})}\|w_j - w^*\|\right] \leq \frac{1}{c\bar{r}}\left(\gamma D^2 + \frac{G^2}{\gamma}\right) t^{-1/2}. \tag{25}$$

When $\Psi(w)$ is strongly convex with the modulus $\sigma > 0$, then with $\{\beta_t\}$ chosen by (21) *we have*

$$\frac{1}{t}\sum_{j=1}^{t} \mathbb{E}\left[\|w_j - w^*\|^2\right] \leq \frac{G^2}{2\sigma \min(c, \sigma/2)} \frac{6 + \ln t}{t}. \tag{26}$$

Proof See Appendix B. ■

Note the differences between (24) and (25): the norms in the summation are squared in the first expression but not in the second, which includes an extra factor of $1/\bar{r}$. The next result combines these bounds into a more useful form for the results that follow.

Theorem 10 *Suppose the assumptions of Lemma 9 are satisfied. Then for the iterates $w_1, w_2, \ldots, w_t$ generated by the stochastic RDA algorithm with the choice* (20) *for $\{\beta_t\}$, we have*

$$\frac{1}{t}\sum_{j=1}^{t}\mathbb{E}\|w_j - w^*\| \leq \mu t^{-1/4}, \tag{27}$$

for all $t \geq \hat{t}$, where the constants $\hat{t}$ and μ are defined as follows:

$$\mu := \frac{2}{\sqrt{c\bar{r}}}\left(\gamma D^2 + \frac{G^2}{\gamma}\right)^{1/2}, \qquad \hat{t} := \left\lceil \frac{1}{\bar{r}^2 c^2}\left(\gamma D^2 + \frac{G^2}{\gamma}\right)^2 \right\rceil. \tag{28}$$

When $\Psi(w)$ is strongly convex with the modulus $\sigma > 0$, then with the choice (21) *for $\{\beta_t\}$ we have*

$$\frac{1}{t}\sum_{j=1}^{t}\mathbb{E}\|w_j - w^*\| \leq \mu'\left(\frac{6+\ln t}{t}\right)^{1/2} \tag{29}$$

for the constant μ' defined by

$$\mu' := \frac{G}{\sqrt{2\sigma\min(c,\sigma/2)}}. \tag{30}$$

Proof See Appendix B. ■

3.3 Stochastic Behavior of the Dual Average

We now study the convergence properties of the dual average $\bar{g}_t$, showing that the probability that $\bar{g}_t$ lies outside any given ball around $\nabla f(w^*)$ goes to zero as t increases.

Theorem 11 *Suppose that Assumptions 1 and 2 are satisfied and that $w^* \in \mathcal{F}_D$, and let $\varepsilon > 0$ be any chosen positive constant. When $\{\beta_t\}$ is chosen from* (20)*, we have*

$$\mathbb{P}(\|\bar{g}_t - \nabla f(w^*)\| \geq \varepsilon) \leq 2n\exp\left(-\frac{\varepsilon^2 t}{32n^2G^2}\right) + 2\mu\varepsilon^{-1}Lt^{-1/4}, \quad \forall t \geq \hat{t}, \tag{31}$$

where μ and $\hat{t}$ are defined in (28)*. When Ψ is strongly convex and the choice* (21) *is used for $\{\beta_t\}$, we have*

$$\mathbb{P}(\|\bar{g}_t - \nabla f(w^*)\| \geq \varepsilon) \leq 2n\exp\left(-\frac{\varepsilon^2 t}{32n^2G^2}\right) + 2\mu'\varepsilon^{-1}L\left(\frac{6+\ln t}{t}\right)^{1/2}, \quad \forall t \geq 1, \tag{32}$$

where μ' is defined in (30)*.*

Proof Using the definition (19) of $\bar{g}_t$ and the unbiasedness property (10) that $\mathbb{E}[\nabla F(w_j;\xi_j) \mid \xi_{[j-1]}] = \nabla f(w_j)$, we can write

$$t[\bar{g}_t - \nabla f(w^*)] = \sum_{j=1}^{t}\left\{\nabla F(w_j;\xi_j) - \mathbb{E}[\nabla F(w_j;\xi_j) \mid \xi_{[j-1]}]\right\} + \sum_{j=1}^{t}\left\{\nabla f(w_j) - \nabla f(w^*)\right\}.$$

Considering the norms of the vectors in both sides and using the Lipschitz property of $\nabla f(\cdot)$ in Lemma 2, we get

$$t\|\bar{g}_t - \nabla f(w^*)\| \leq \left\| \sum_{j=1}^{t} z_j \right\| + L \sum_{j=1}^{t} \|w_j - w^*\|, \tag{33}$$

where we define $z_j := \nabla F(w_j;\xi_j) - \mathbb{E}[\nabla F(w_j;\xi_j) \mid \xi_{[j-1]}]$.

For the ith component $[z_j]_i$ of the vector $z_j \in \mathbb{R}^n$, $\sum_{j=1}^{t}[z_j]_i$ forms a martingale with bounded differences since $|[z_j]_i| \leq 2G$ from (12) and $\mathbb{E}[z_j \mid \xi_{[j-1]}] = 0$. Therefore by Hoeffding-Azuma inequality (Azuma, 1967) we obtain for any $\theta > 0$,

$$\mathbb{P}\left(\left| \sum_{j=1}^{t} [z_j]_i \right| \geq \theta \right) \leq 2\exp\left(-\frac{\theta^2}{8tG^2} \right), \ \ i = 1, 2, \ldots, n.$$

Therefore, using the equivalence relation $\|v\|_2 \leq \|v\|_1$ of norms for a vector $v \in \mathbb{R}^n$, we have

$$\begin{aligned} \mathbb{P}\left(\left\| \sum_{j=1}^{t} z_j \right\|_2 \geq \theta \right) &\leq \mathbb{P}\left(\left\| \sum_{j=1}^{t} z_j \right\|_1 \geq \theta \right) = \mathbb{P}\left(\sum_{i=1}^{n} \left| \sum_{j=1}^{t} [z_j]_i \right| \geq \theta \right) \\ &\leq \mathbb{P}\left(\bigcup_{i=1}^{n} \left\{ \left| \sum_{j=1}^{t} [z_j]_i \right| \geq \frac{\theta}{n} \right\} \right) \leq 2n\exp\left(-\frac{\theta^2}{8tn^2G^2} \right). \end{aligned} \tag{34}$$

Here the second inequality uses the implication that at least one $a_i \in \mathbb{R}$ should satisfy $a_i \geq b/n$ when $\sum_{i=1}^{n} a_i \geq b$, and the last inequality is from the union bound of probabilities.

Also, from Markov's inequality and the bound (27) in Theorem 10 we get for any $\theta' > 0$ and $t \geq \hat{t}$ that

$$\mathbb{P}\left(\frac{1}{t} \sum_{j=1}^{t} \|w_j - w^*\| \geq \theta' \right) \leq \frac{\mu t^{-1/4}}{\theta'}. \tag{35}$$

Together with (33), the bounds in (34) and (35) imply that

$$\begin{aligned} \mathbb{P}\left(t\|\bar{g}_t - \nabla f(w^*)\| \geq \delta \right) &\leq \mathbb{P}\left(\left\| \sum_{j=1}^{t} z_i \right\| + L \sum_{j=1}^{t} \|w_j - w^*\| \geq \delta \right) \\ &\leq \mathbb{P}\left(\left\| \sum_{j=1}^{t} z_i \right\| \geq \frac{\delta}{2} \right) + \mathbb{P}\left(\frac{1}{t} \sum_{j=1}^{t} \|w_j - w^*\| \geq \frac{\delta}{2tL} \right) \\ &\leq 2n\exp\left(-\frac{\delta^2}{32tn^2G^2} \right) + \frac{2\mu L t^{3/4}}{\delta}. \end{aligned}$$

The first claim follows when we define $\delta := \varepsilon t$:

$$\mathbb{P}\left(\|\bar{g}_t - \nabla f(w^*)\| \geq \varepsilon \right) \leq 2n\exp\left(-\frac{\varepsilon^2 t}{32n^2G^2} \right) + \frac{2\mu L t^{-1/4}}{\varepsilon}.$$

The claim for the strongly convex case is proved similarly, using the bound (29) in Theorem 10 instead of (27) when we apply Markov inequality in (35). ■

The next result shows formally that for all t sufficiently large, the second term dominates in the bounds (31) and (32).

Corollary 12 *Suppose the assumptions of Theorem 11 hold and let μ and $\hat{t}$ be defined as in* (28). *Then for $\varepsilon \in (0, 4nG/\sqrt{e}]$ and $t \geq \max(\hat{t}, \bar{t})$, with*

$$\bar{t} := \left\lceil 16n^2G^2\varepsilon^{-2} \max\left(-W\left(\frac{-\varepsilon^2}{16n^2G^2}\right), -4\ln\left(\frac{\mu L}{n\varepsilon}\right)\right)\right\rceil, \tag{36}$$

(where $W(\cdot)$ denotes the branch of the Lambert function with $W < -1$), the bound (31) *simplifies to*

$$\mathbb{P}(\|\bar{g}_t - \nabla f(w^*)\| \geq \varepsilon) \leq 4\mu\varepsilon^{-1}Lt^{-1/4}. \tag{37}$$

When Ψ is strongly convex and the choice (21) *is used for $\{\beta_t\}$, and provided that $\varepsilon \in \left(0, 4nG/\sqrt{2/e}\right]$ and $t \geq \bar{t}'$ with*

$$\bar{t}' := \left\lceil 32n^2G^2\varepsilon^{-2} \max\left(-W\left(\frac{-\varepsilon^2}{32n^2G^2}\right), -2\ln\left(\frac{\mu' L\sqrt{6}}{n\varepsilon}\right)\right)\right\rceil, \tag{38}$$

the bound (32) *simplifies to*

$$\mathbb{P}(\|\bar{g}_t - \nabla f(w^*)\| \geq \varepsilon) \leq 4\mu'\varepsilon^{-1}L\left(\frac{6+\ln t}{t}\right)^{1/2}, \qquad t \geq 1. \tag{39}$$

Proof Note first that the ratio $(\ln t)/t$ is decreasing for $t \geq e$, so for t_1 defined by

$$\ln t_1 = \frac{\varepsilon^2}{16n^2G^2}t_1, \tag{40}$$

we have for $t \geq t_1$ and for sufficiently small ε ensuring $\varepsilon^2/(16n^2G^2) \leq 1/e$ (which is guaranteed by our assumption on ε) that

$$\frac{\ln t}{t} \leq \frac{\ln t_1}{t_1} \leq \frac{\varepsilon^2}{16n^2G^2}. \tag{41}$$

Note that (40) has the form $\ln t_1 = \alpha t_1$, for $\alpha = \varepsilon^2/(16n^2G^2)$, for which the solution is $t_1 = -W(-\alpha)/\alpha$. Thus, for any $t \geq \bar{t}$, the condition (41) holds. We continue to assume that $t \geq \bar{t}$, and note that

$$-\frac{\varepsilon^2 t}{32n^2G^2} + \frac{1}{4}\ln t \overset{(41)}{\leq} -\frac{\varepsilon^2 t}{32n^2G^2} + \frac{\varepsilon^2 t}{64n^2G^2} = -\frac{\varepsilon^2 t}{64n^2G^2} \overset{(36)}{\leq} \ln\left(\frac{\mu L}{n\varepsilon}\right).$$

By rearranging this expression and taking the exponential of both sides, we obtain

$$-\frac{\varepsilon^2 t}{32n^2G^2} \leq \ln\left(\frac{\mu L}{n\varepsilon}\right) - \frac{1}{4}\ln t \;\Leftrightarrow\; \exp\left(-\frac{\varepsilon^2 t}{32n^2G^2}\right) \leq \left(\frac{\mu L}{n\varepsilon}\right) t^{-1/4},$$

which implies that the first term on the right-hand side of the bound (31) is dominated by the second. We obtain the result (37) by substituting into (31).

The strongly convex case is similar. By solving $\ln t_2 = \alpha t_2$, for $\alpha = \varepsilon^2/(32n^2G^2)$, we have for $t \geq t_2$ and sufficiently small ε that

$$\frac{\ln t}{t} \leq \frac{\ln t_2}{t_2} \leq \frac{\varepsilon^2}{32n^2G^2}. \tag{42}$$

Thus for $t \geq \bar{t}'$, we have

$$-\frac{\varepsilon^2 t}{32n^2G^2} + \frac{1}{2}\ln t \overset{(42)}{\leq} -\frac{\varepsilon^2 t}{64n^2G^2} \overset{(38)}{\leq} \ln\left(\frac{\mu' L\sqrt{6}}{\varepsilon n}\right) \leq \ln\left(\frac{\mu' L\sqrt{6+\ln t}}{\varepsilon n}\right).$$

By rearranging the outermost expressions in this bound, and taking the exponents of both sides, we obtain

$$\exp\left(-\frac{\varepsilon^2 t}{32n^2G^2}\right) \leq \frac{\mu' L}{\varepsilon n}\left(\frac{6+\ln t}{t}\right)^{1/2},$$

from which the result (39) follows. ■

The max-terms in (36) and (38) grow only slowly with the dimension n. Hence, we can base our estimate of the required size of t on the factor in front of the max terms in (36) and (38), and on the right-hand sides of (37) and (39). In particular, for the general case, we need t to be at least a modest multiple of $(4nG/\varepsilon)^2$ (for $t \geq \bar{t}$) and also $t \geq (4\mu L/\varepsilon)^4$ (for the right-hand side of (37) to be useful).

4. Manifold Identification

In this section we show that most sufficiently advanced iterates of the RDA algorithm lie on the optimal manifold. Our analysis is based upon the properties of the dual average discussed in the previous section and on basic results for manifold identification. Specifically, we make use of a result of Hare and Lewis (2004), which states that when a sequence of points approaches a limit lying on an optimal nondegenerate manifold and the subgradients at these points approach zero, then all sufficiently advanced members of the sequence lie on the manifold. We identify subsequences of the RDA sequence that lie on the optimal manifold with increasing likelihood as the iteration counter grows. We further show that these subsequences form a dense subset of the full sequence. Separate but similar results are proved for the general convex case and the strongly convex case.

4.1 Convergent Sequences

We start with two results that estimate the likelihood of w_j lying within a given radius of w^*. The first of these results is for general convex objectives.

Lemma 13 (Convergent Sequences for General Convex Case) *Suppose that Assumptions 1 and 2 hold, that $w^* \in \mathcal{F}_D$, and that $\{\beta_j\}$ is chosen according to* (20)*. Define the subsequence $\mathcal{S}$ by*

$$\mathcal{S} := \Big\{ j \in \{1,2,\ldots\} \mid \mathbb{E}\left[I_{(\|w_j - w^*\| \leq \bar{r})}\|w_j - w^*\|^2\right] \leq j^{-1/4}, \text{ and} \\ \mathbb{E}\left[I_{(\|w_j - w^*\| > \bar{r})}\|w_j - w^*\|\right] \leq (1/\bar{r})j^{-1/4}\Big\}. \tag{43}$$

For any $\varepsilon > 0$, we then have

$$\mathbb{P}(\|w_j - w^*\| \geq \varepsilon) \leq \frac{1}{\varepsilon}\left(\frac{1}{\varepsilon} + \frac{1}{\bar{r}}\right) j^{-1/4}, \ \forall j \in \mathcal{S}. \tag{44}$$

Defining

$$\mathcal{S}_t := \mathcal{S} \cap \{1,2,\ldots,t\},$$

we have

$$\frac{1}{t}\mathrm{card}(\mathcal{S}_t) \geq 1 - \frac{2}{c}\left(\gamma D^2 + \frac{G^2}{\gamma}\right) t^{-1/4}, \tag{45}$$

that is, the density of $\mathcal{S}_t$ in $\{1,2,\ldots,t\}$ is $1 - O(t^{-1/4})$.

Proof To measure the cardinality of the complement of $\mathcal{S}_t$, that is, $\mathcal{S}_t^c := \{1,2,\ldots,t\} \setminus \mathcal{S}_t$, we first define indicator variables χ_-^j and χ_+^j for $j \geq 1$ as follows:

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

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

As the set $\mathcal{S}_t^c$ contains all indices $j \in \{1,2,\ldots,t\}$ that satisfy $\chi_-^j = 1$ or $\chi_+^j = 1$, the cardinality of $\mathcal{S}_t^c$ is bounded above by $\sum_{j=1}^t (\chi_-^j + \chi_+^j)$. For $\sum_{j=1}^t \chi_-^j$, we note that

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

Using (24), 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 $\sum_{j=1}^t \chi_+^j$ with $\mathbb{E}\left[I_{(\|w_j - w^*\| > \bar{r})} \|w_j - w^*\|\right]$, $j = 1,2,\ldots,t$ and (25) 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 of the cardinality of $\mathcal{S}_t$ to $\{1,2,\ldots,t\}$ is

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

thus proving (45).

To show (44), we first observe that for any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}(\|w_j - w^*\| \geq \varepsilon) = \mathbb{P}(\|w_j - w^*\| \geq \varepsilon,\ \|w_j - w^*\| \leq \bar{r}) \\ + \mathbb{P}(\|w_j - w^*\| \geq \varepsilon,\ \|w_j - w^*\| > \bar{r}).\end{aligned} \tag{46}$$

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

$$\begin{aligned}\mathbb{P}(\|w_j - w^*\| \geq \varepsilon,\ \|w_j - w^*\| \leq \bar{r}) &= \mathbb{P}(I_{(\|w_j - w^*\| \leq \bar{r})}\|w_j - w^*\| \geq \varepsilon) \\ &\leq \varepsilon^{-2}\mathbb{E}\left[I_{(\|w_j - w^*\| \leq \bar{r})}\|w_j - w^*\|^2\right] \\ &\leq \varepsilon^{-2} j^{-1/4},\end{aligned} \tag{47}$$

where the first inequality is due to Markov and the second inequality is from the definition of $\mathcal{S}$ in (43). Similarly for the second term in (46), we have for all $j \in \mathcal{S}$ that

$$\begin{aligned}\mathbb{P}(\|w_j - w^*\| \geq \varepsilon,\ \|w_j - w^*\| > \bar{r}) &= \mathbb{P}(I_{(\|w_j - w^*\| > \bar{r})}\|w_j - w^*\| \geq \varepsilon) \\ &\leq \varepsilon^{-1}\mathbb{E}\left[I_{(\|w_j - w^*\| > \bar{r})}\|w_j - w^*\|\right] \\ &\leq \varepsilon^{-1}\bar{r}^{-1} j^{-1/4}.\end{aligned} \tag{48}$$

Applying (47) and (48) to (46) leads to the claim,

$$\mathbb{P}(\|w_j - w^*\| \geq \varepsilon) \leq \varepsilon^{-1}(\varepsilon^{-1} + \bar{r}^{-1}) j^{-1/4}, \quad \forall j \in \mathcal{S}.$$

■

This result implies that for sufficiently large j, the majority of iterates w_j converges to w^*, in probability. Similar results can be derived for the convergence of $\mathbb{E}[\phi(w_j)]$ to $\phi(w^*)$. The next theorem is the corresponding result for the strongly convex case.

Lemma 14 (Convergent Sequences for Strongly Convex Case) *Suppose that Assumptions 1 and 2 hold, that $w^* \in \mathcal{F}_D$, and that the regularizer Ψ is strongly convex with modulus $\sigma > 0$. Suppose that $\{\beta_j\}$ is defined by* (21)*. For any $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\|w_j - w^*\| \geq \varepsilon\right) \leq \frac{G^2}{\varepsilon^2\sigma^2}\left(\frac{6 + \ln j}{j}\right), \qquad j \geq 1.$$

Proof From the proof of Xiao (2010, Corollary 4), we have

$$\mathbb{E}\left[\|w_j - w^*\|^2\right] \leq \frac{G^2}{\sigma^2}\left(\frac{6 + \ln j}{j}\right), \qquad j \geq 1.$$

The claim follows from the Markov inequality. ■

4.2 Identification

In this subsection, we state the main identification results. We start with a result from Hare and Lewis (2004), stating it in a modified form that is useful for our analysis below.

Theorem 15 *Suppose that ϕ is partly smooth at the minimizer w^* relative to the optimal manifold $\mathcal{M}$ and that the nondegeneracy condition* (13) *holds. Then there exists a threshold $\bar{\varepsilon} > 0$ such that for all $w \in \mathcal{O}$ with $\|w - w^*\| < \bar{\varepsilon}$ and $\operatorname{dist}(0, \partial\phi(w)) < \bar{\varepsilon}$, we have $w \in \mathcal{M}$.*

Proof Suppose for contradiction that no such $\bar{\varepsilon}$ exists. Let $\{\varepsilon_j\}_{j\geq 1}$ be any sequence of positive numbers such that $\varepsilon_j \downarrow 0$. Then for each $j \geq 1$ we have w_j such that $\|w_j - w^*\| < \varepsilon_j$, $\operatorname{dist}(0, \partial\phi(w_j)) < \varepsilon_j$ 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 for all $a_j \in \partial\phi(w_j)$ we have $\phi(w_j) - \phi(w^*) \leq a_j^T(w_j - w^*) \leq \|a_j\|\|w_j - w^*\|$. (Because of our assumptions, we can choose a_j such that $\|a_j\| \leq \varepsilon_j$.) Convexity implies prox-regularity, so by applying Theorem 5.3 of Hare and Lewis (2004), we have that $w_j \in \mathcal{M}$ for all j sufficiently large. This contradicts our choice of w_j, so we conclude that $\bar{\varepsilon} > 0$ with the claimed properties exists. ■

The next theorem is our main result, showing that the RDA algorithm identifies the optimal manifold with increasing probability as iterations proceed. This result requires a condition (49) on h that is trivially satisfied by the usual prox-function $h(w) = \frac{1}{2}\|w - w_1\|^2$, with constant $\eta = 1$.

Theorem 16 (Identification for General Convex Case) *Suppose that Assumptions 1 and 2 hold, that $w^* \in \mathcal{F}_D$, that*

$$\sup_{b_j \in \partial h(w_j)} \|b_j\| \leq \eta \|w_j - w_1\|, \quad j = 1, 2, \ldots \tag{49}$$

for some $\eta > 0$, and that $\{\beta_j\}$ is defined as in (20)*. Given the set of indices $\mathcal{S}$ defined in* (43)*, we have*

$$\mathbb{P}(w_j \in \mathcal{M}) \geq 1 - (\zeta_1 + \zeta_2) j^{-1/4}$$

for all $j \in \mathcal{S}$ sufficiently large, where

$$\zeta_1 := \frac{3\max(1, L)}{\bar{\varepsilon}} \left(\frac{3\max(1, L)}{\bar{\varepsilon}} + \frac{1}{\bar{r}} \right), \text{ and } \zeta_2 := \frac{15\mu L}{\bar{\varepsilon}}.$$

Here $\bar{\varepsilon} > 0$ has the value defined in Theorem 15, L is the Lipschitz constant of (11)*, $\bar{r}$ is the radius of strong local minimization from* (15)*, and μ is defined in* (28)*.*

Proof We focus on the iterate w_j and the random events associated with it. First we denote the following event as E_1:

$$E_1: \quad \|w_j - w^*\| \leq \frac{\bar{\varepsilon}}{3\max(L, 1)}.$$

Note that E_1 depends on the history $\xi_{[j-1]}$ of random variables prior to iteration j. If E_1 is true, it trivially implies the condition $\|w_j - w^*\| \leq \bar{\varepsilon}$ of Theorem 15. From Lemma 13, with $\varepsilon = \frac{\bar{\varepsilon}}{3\max(L,1)}$, we have that

$$\mathbb{P}(\|w_j - w^*\| \leq \bar{\varepsilon}) \geq \mathbb{P}(E_1) \geq 1 - \zeta_1 j^{-1/4}, \quad \text{for all } j \in \mathcal{S}. \tag{50}$$

We now examine the other condition in Theorem 15, namely

$$\operatorname{dist}\big(0, \nabla f(w_j) + \partial\Psi(w_j)\big) \leq \bar{\varepsilon}.$$

By adding and subtracting terms, we obtain

$$\begin{aligned}\nabla f(w_j) + a_j = (\nabla f(w_j) - \nabla f(w^*)) + (\nabla f(w^*) - \bar{g}_{j-1}) - \frac{\beta_{j-1}}{j-1} b_j \\ + \left(\bar{g}_{j-1} + a_j + \frac{\beta_{j-1}}{j-1} b_j\right). \end{aligned} \tag{51}$$

for any $a_j \in \partial\Psi(w_j)$ and $b_j \in \partial h(w_j)$. We choose the specific a_j and b_j that satisfy the optimality of the subproblem (18), that is,

$$0 = \bar{g}_{j-1} + a_j + \frac{\beta_{j-1}}{j-1} b_j.$$

This choice eliminates the last term in (51). For the other three terms, we have the following observations.

(i) For those w_j satisfying E_1, the Lipschitz property of ∇f (Lemma 2) implies that

$$\|\nabla f(w_j) - \nabla f(w^*)\| \leq L\|w_j - w^*\| \leq \frac{\bar{\varepsilon} L}{3\max(L,1)} \leq \frac{\bar{\varepsilon}}{3}.$$

Hence, E_1 implies the following event:

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

(ii) From Corollary 12, we have by setting $\varepsilon = \bar{\varepsilon}/3$ and $t = j-1$ that

$$\mathbb{P}(\|\nabla f(w^*) - \bar{g}_{j-1}\| \geq \bar{\varepsilon}/3) \leq 4\mu L \left(\frac{3}{\bar{\varepsilon}}\right)(j-1)^{-1/4} < \zeta_2 j^{-1/4},$$

for $j - 1 \geq \max(\hat{t}, \bar{t})$, where $\hat{t}$ is defined in (28) and $\bar{t}$, which depends on $\bar{\varepsilon}$, is defined in (36). Hence, denoting by E_3 the event

$$E_3: \quad \|\nabla f(w^*) - \bar{g}_{j-1}\| \leq \bar{\varepsilon}/3,$$

we have that

$$\mathbb{P}(\neg E_3) < \zeta_2 j^{-1/4}, \qquad j \geq \max(\hat{t}, \bar{t}) + 1. \tag{52}$$

(iii) Since $\beta_{j-1} = \gamma(j-1)^{1/2}$, we have for w_j satisfying E_1 that

$$\begin{aligned}\frac{\beta_{j-1}}{j-1}\|b_j\| &= \gamma(j-1)^{-1/2}\|b_j\| \\ &\leq \gamma\eta(j-1)^{-1/2}\|w_j - w_1\| && \text{from (49)} \\ &\leq \gamma\eta(j-1)^{-1/2}(\|w_j - w^*\| + \|w_1 - w^*\|) \\ &\leq \gamma\eta(j-1)^{-1/2}\left(\frac{\bar{\varepsilon}}{3\max(L,1)} + \sqrt{2}D\right) && \text{from } w^* \in \mathcal{F}_D \text{ and (17)}.\end{aligned}$$

Therefore, E_1 implies the event

$$E_4: \quad \frac{\beta_{j-1}}{j-1}\|b_j\| \leq \frac{\bar{\varepsilon}}{3}, \qquad j \geq t_0+1,$$

where we define t_0 by

$$t_0 := \left\lceil \frac{9\gamma^2\eta^2}{\bar{\varepsilon}^2}\left(\frac{\bar{\varepsilon}}{3\max(L,1)}+\sqrt{2}D\right)^2 \right\rceil.$$

Therefore for $j \in \mathcal{S}$ with $j \geq \max\{\hat{t},\bar{t},t_0\}+1$, by definition of the events E_1, E_2, E_3, and E_4 above, the probability that the conditions of Theorem 15 hold is bounded as follows:

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

where the last inequality is due to (50) and (52). Our claim follows. ■

Theorem 17 (Identification for Strongly Convex Case) *Suppose that Assumptions 1 and 2 hold, that Ψ is strongly convex with modulus $\sigma > 0$, that $w^* \in \mathcal{F}_D$, that $h(\cdot)$ satisfies* (49)*, and that $\beta_j = 0$ for all $j \geq 1$, as defined in* (21)*. Then we have*

$$\mathbb{P}(w_j \in \mathcal{M}) \geq 1 - (\zeta_1' + \zeta_2')\left(\frac{6+\ln j}{j}\right)^{1/2}.$$

for all j sufficiently large, where

$$\zeta_1' := \frac{G^2}{\sigma^2}\left(\frac{3\max(1,L)}{\bar{\varepsilon}}\right)^2, \text{ and } \zeta_2' := \frac{17\mu' L}{\bar{\varepsilon}}.$$

Here $\bar{\varepsilon} > 0$ has the value defined in Theorem 15, L is the Lipschitz constant of (11)*, G is the uniform bound on gradient norms in* (12)*, and μ' is defined in* (30)*.*

Proof This proof is almost identical to that of Theorem 16; here we briefly mention the required changes for the strongly convex case. Consider $\bar{\varepsilon} > 0$ and the event E_1 defined in the proof of Theorem 16. From Lemma 14 with $\varepsilon = \frac{\bar{\varepsilon}}{3\max(L,1)}$, we have

$$\mathbb{P}(\|w_j - w^*\| \leq \bar{\varepsilon}) \geq \mathbb{P}(E_1) \geq 1 - \zeta_1'(6+\ln j)/j, \qquad j \geq 1.$$

Instead of (ii) and (iii) in the proof of Theorem 16, we use the following:

(ii') From Corollary 12, we have by setting $\varepsilon = \bar{\varepsilon}/3$ and $t = j-1$ that

$$\mathbb{P}(\|\nabla f(w^*) - \bar{g}_{j-1}\| \geq \bar{\varepsilon}/3) \leq 4\mu'\frac{3}{\bar{\varepsilon}}L\left(\frac{6+\ln(j-1)}{j-1}\right)^{1/2} < \frac{17\mu' L}{\bar{\varepsilon}}\left(\frac{6+\ln j}{j}\right)^{1/2},$$

for all $j > \max(\bar{t}'+1, 2)$, where $\bar{t}'$ is defined in (38). Hence, denoting by E_3 the event $\|\nabla f(w^*) - \bar{g}_{j-1}\| \leq \bar{\varepsilon}/3$, we have that

$$\mathbb{P}(\neg E_3) < \zeta_2' \left(\frac{6+\ln j}{j}\right)^{1/2},$$

for all j sufficiently large.

(iii') With $\beta_{j-1} = 0$ and the given conditions, the event E_4 holds for all $j \geq 2$.

Using the modified probability bounds for E_1 and E_3, we have

$$\begin{aligned} P\Big(\|w_j - w^*\| \leq \bar{\varepsilon} \wedge \operatorname{dist}(0, \partial\phi(w_j)) \leq \bar{\varepsilon}\Big) &\geq P\Big(E_1 \wedge E_3\Big) \\ &\geq 1 - \zeta_1' \left(\frac{6+\ln j}{j}\right) - \zeta_2' \left(\frac{6+\ln j}{j}\right)^{1/2} \\ &\geq 1 - (\zeta_1' + \zeta_2') \left(\frac{6+\ln j}{j}\right)^{1/2}, \end{aligned}$$

for all $j \geq \max(\bar{t}'+1, 9)$, using the fact that $(6+\ln j)/j \leq 1$ for $j \geq 9$. Our claim follows. ■

Lemma 13 tells us that the sequence $\mathcal{S}$ is "dense" in $\{1, 2, \dots\}$, while Theorem 16 states that for all sufficiently large $j \in \mathcal{S}$, w_j lies on the optimal manifold with probability approaching one as j increases. When the regularizer Ψ is strongly convex, Theorem 17 tells that similar results hold earlier in the sequence $\{w_j\}$.

We conclude this section with a simple example to show that algorithms based on subproblem (7) that were discussed in Section 1.3 do not have reliable identification properties. The major reason is that each iteration uses a "raw" sampled gradient g_t of f, rather than the averaged (and thus smoothed) approximate gradient $\bar{g}_t$ of RDA. Thus, as we see now, even when the current iterate w_t is optimal, the subproblem may step away from this point, and away from the optimal manifold.

Example 1 Consider the following definitions for the problem (1):

- $n = 1$ (a scalar problem)
- $\Xi = [-2, 2]$ with ξ uniformly distributed on this interval.
- $F(w;\xi) = \xi w$. Thus $\nabla F(w;\xi) = \xi$ and $f(w) \equiv 0$.
- $\Psi(w) = |w|$.

With these definitions, the solution is $w^* = 0$ and the optimal manifold is zero-dimensional: $\mathcal{M} = \{0\}$. Thus Assumption 2 is trivially satisfied. Regarding Assumption 1, the nondegeneracy condition is satisfied, since $\partial\Psi(0) = [-1, 1]$ while $\nabla f(0) = 0$. It is easy to verify too that F satisfies the unbiasedness, uniform Lipschitz, and uniform boundedness properties of Assumption 1.

Setting $w_t = w^* = 0$, the subproblem (7) is

$$w_{t+1} = \arg\min_w \xi_t w + |w| + \frac{1}{2\alpha_t} w^2,$$

where ξ_t is selected uniformly at random from $[-2,2]$. For $\xi_t \in [-1,1]$, we have $w_{t+1} = 0$, so the next iterate remains at the optimal point. However, if $\xi_t \in [-2,-1)$, we have $w_{t+1} = -\alpha_t(\xi_t + 1) > 0$, while if $\xi_t \in (1,2]$, we have $w_{t+1} = -\alpha_t(\xi_t - 1) < 0$. In both cases, the next iterate steps away from the solution $w^* = 0$ and from the optimal manifold. Because ξ_t is uniformly distributed in $[-2,2]$, this event happens with probability $1/2$ in this example. (The probability of this behavior can be made arbitrarily close to 1 by suitable extension of the interval Ξ.)

5. Enhancing the Regularized Dual Averaging Algorithm

Here we present a simple optimization strategy motivated by our analysis above, in which the RDA method gives way to a local phase after a near-optimal manifold is identified.

Algorithm 2 summarizes our algorithm RDA$^+$. This algorithm starts with RDA steps until it identifies a near-optimal manifold, then searches this manifold using a different optimization strategy, possibly better suited to lower-dimensional spaces and possibly with better local convergence properties. If an explicit parametrization of $\mathcal{M}$ is available, the "local phase" could take the form of a Newton-like method applied to the composition of the objective with this parametrization. In the important special case of $\Psi(\cdot) = \lambda\|\cdot\|_1$, $\mathcal{M}$ can be represented simply by its nonzero variables, and LPS (Shi et al., 2008; Wright, 2012) can be applied on the space of just these variables.

To decide when to make the switch to the local phase, we use a simple heuristic inspired by Theorem 16 and 17 that if the past τ consecutive iterates have been on the same manifold $\mathcal{M}$, we take $\mathcal{M}$ to be approximately optimal. However, we "safeguard" by expanding $\mathcal{M}$ to incorporate additional dimensions that may yet contain the minimizer. This simple approach will work provided that the $\mathcal{M}$ so constructed is a *superset* of the optimal manifold, since our implementation of the local phase is able to move to more restricted submanifolds of $\mathcal{M}$ but does not expand its search to include dimensions not originally included in the manifold. When sufficient progress is not attained in the local phase, we can resume the paused dual-averaging phase.

We describe the details of Algorithm 2 for ℓ_1 regularization, where $\Psi(w) = \lambda\|w\|_1$ for some $\lambda > 0$. (Thus, the starting point will be $w_1 = 0$.) The optimal manifold corresponds (near w^*) to the points in $\mathbb{R}^n$ that have the same nonzero patterns as w^*. We use the simple quadratic prox-function $h(w) = \frac{1}{2}\|w - w_1\|^2$. Since Ψ is not strongly convex, we use the sequence $\{\beta_t\}$ defined in (20).

We now describe various specifics of the implementation of Algorithm 2 for this case.

5.1 Computation of w_{j+1}

The closed-form solution for the subproblem (18) with $t = j$ is

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

where $\text{soft}(u,a) := \text{sgn}(u)\max\{|u| - a, 0\}$ is the well-known soft-threshold function.

5.2 Surrogate Objective

To apply the batch optimization method LPS in the local phase of Algorithm 2, we use an empirical estimate $\tilde{\phi}_{\mathcal{N}}$ in (3) as a surrogate objective function (where ξ_j, $j \in \mathcal{N}$ is a sample of random variables from the space Ξ according to the distribution P), and then solve

$$\min_{w \in \mathcal{M}} \;\; \tilde{\phi}_{\mathcal{N}}(w) = \tilde{f}_{\mathcal{N}}(w) + \lambda\|w\|_1. \tag{53}$$

Algorithm 2: RDA$^+$ Algorithm.

Input:

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

$$\arg\min_{w\in\mathbb{R}^n} h(w) \in \arg\min_{w\in\mathbb{R}^n} \Psi(w),$$
$$\sup_{b\in\partial h(w)} \|b\| \le \eta\|w-w_1\|, \ \forall w \in \operatorname{dom}\Psi, \quad \text{where } w_1 \in \arg\min_w \Psi(w).$$

- a nonnegative and nondecreasing sequence $\{\beta_j\}$, $j \ge 1$.

- a positive integer τ.

Initialize: Set $\bar{g}_0 = 0$;
for $j = 1,2,\ldots$ **do** (Dual Averaging)
  Choose a random vector $\xi_j \in \Xi$;
  Compute a gradient $g_j = \nabla F(w_j;\xi_j)$;
  Update the average gradient:

$$\bar{g}_j = \frac{j-1}{j}\bar{g}_{j-1} + \frac{1}{j}g_j.$$

  Compute the next iterate by solving the subproblem (18), which is

$$w_{j+1} = \arg\min_{w\in\mathbb{R}^n} \left\{ \langle \bar{g}_j, w\rangle + \Psi(w) + \frac{\beta_j}{j}h(w) \right\}.$$

  if *there is* $\mathcal{M}$ *such that* $w_{j+2-i} \in \mathcal{M}$ *for* $i = 1,2,\ldots,\tau$ **then** (Local Phase)
    Safeguard $\mathcal{M}$ by adding dimensionality as appropriate to encompass w^*;
    Use a technique for low-dimensional optimization to search for a solution on manifold $\mathcal{M}$, starting at w_{j+1};
  end
end

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.

5.3 Local Phase: LPS

We give further information on the LPS approach, following Wright (2012) but specializing the description to the problem (53). Many details are omitted; we refer the reader to Wright (2012) for a complete description and analysis of the approach, and to Shi et al. (2008) for an earlier version together with an application to ℓ_1-regularized logistic regression.

Algorithm 3: LPS Approach in Local Phase

Input: $\nu_{\max} > \nu_{\min} > 0$, $K > 1$, $s > 1$, w_{j+1};
Initialize: $z_0 \leftarrow w_{j+1}$;
for $k = 0, 1, 2, \ldots$ **do**
  Choose $\mathcal{M}_k \subset \mathcal{M}$ such that each nonzero component in $\mathcal{M}$ appears in at least one of the manifolds $\mathcal{M}_k, \mathcal{M}_{k-1}, \ldots, \mathcal{M}_{k-K+1}$, for $k \geq K$;
  Choose $\nu_k \in [\nu_{\min}, \nu_{\max}]$;
  Solve (54) for d_k;
  while $\tilde{\phi}_{\mathcal{N}}(z_k + d_k) > \tilde{\phi}_{\mathcal{N}}(z_k) - |d_k|^3$ **do**
    Set $\nu_k \leftarrow s\nu_k$;
    Solve (54) for d_k;
  end
RN Find $\tilde{d}_k$ with $\tilde{\phi}_{\mathcal{N}}(z_k + \tilde{d}_k) \leq \tilde{\phi}_{\mathcal{N}}(z_k + d_k)$ and $\tilde{\phi}_{\mathcal{N}}(z_k + \tilde{d}_k) \leq \tilde{\phi}_{\mathcal{N}}(z_k) - .01|\tilde{d}_k|^3$;
  Set $z_{k+1} \leftarrow z_k + \tilde{d}_k$;
end

The scheme is outlined as Algorithm 3. We use $\mathcal{M}_k$ to denote a subset of the restricted manifold $\mathcal{M}$, again defined by the indices of the components in which we consider a move at iteration k. We require that each nonzero component in $\mathcal{M}$ be considered for a possible step at least once every K iterations, where K is a defined parameter. The basic step at each iteration, given an LPS iterate z_k, is obtained by solving the following subproblem:

$$\min_{d \in \mathcal{M}_k} \nabla \tilde{f}_{\mathcal{N}}(z_k)^T d + \frac{\nu_k}{2} d^T d + \lambda \|z_k + d\|_1, \tag{54}$$

where ν_k is manipulated by the algorithm to produce a decrease in the objective at each iteration. Formulation of this subproblem requires evaluation only of those elements of the gradient $\nabla \tilde{f}_{\mathcal{N}}$ corresponding to the nonzero set $\mathcal{M}_k$. Since it is separable in the components of d, it can be solved in closed form in a number of operations proportional to the number of nonzero components in $\mathcal{M}_k$.

The enhancement in the line marked as RN of Algorithm 3 can be carried out by a reduced Newton-type method, applied to the current set of nonzero components of $z_k + d_k$. Here, we obtain an estimate of the restriction of the Hessian $\nabla^2 \tilde{f}_{\mathcal{N}}$ to the nonzero set, possibly by taking a random sub-batch of $\mathcal{N}$.

5.4 Checking Optimality

From the optimality condition for (3), we define the optimality measure $\delta(w_j)$ as follows,

$$\delta(w_j) := \frac{1}{\sqrt{n}} \inf_{a_j \in \partial \|w_j\|_1} \|\nabla \tilde{f}_{\mathcal{N}}(w_j) + \lambda a_j\|. \tag{55}$$

Since $\delta(w^*) \approx 0$ for a sufficiently large sample set $\mathcal{N}$ because of the law of large numbers, it makes sense to terminate when $\delta(w_j)$ drops below a certain threshold.

5.5 Safeguarding

At the start of the local phase, we augment $\mathcal{M}$ by 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,

$$[w_{j+1}]_i = 0 \text{ and } |[\bar{g}_j]_i| > \rho\lambda \tag{56}$$

for some ρ between 0 and 1 (but closer to 1). This conservative strategy allows for the possibility that $|[\bar{g}_j]_i|$ will exceed λ on a later iteration, causing $[w_{j+1}]_i$ to move away from zero.

6. Computational Experiments

We report here on computational experiments involving binary classification tasks via ℓ_1-regularized logistic regression. Given a set of m training examples, we select one at time t—indexed by ξ_t—and use its feature vector $x_{\xi_t} \in \mathbb{R}^{n-1}$ and label $y_{\xi_t} \in \{-1,1\}$ to define the corresponding loss function for $\tilde{w} \in \mathbb{R}^{n-1}$, $b \in \mathbb{R}$ and $w = (\tilde{w}, b)$:

$$F(w;\xi_t) = \ln\left(1 + \exp\left(-y_{\xi_t}(\tilde{w}^T x_{\xi_t} + b)\right)\right).$$

We choose $\Psi(w) = \lambda\|\tilde{w}\|_1$ as the regularizer for some $\lambda > 0$, and set $w_1 = 0$, as required in Algorithm 2.

For the second phase of Algorithm 2, we set $\mathcal{N} = \{1,2,\ldots,m\}$ to obtain the empirical estimate $\tilde{\phi}_{\mathcal{N}}$ from the full training set.

6.1 Manifold Identification

To investigate the identification behavior of the RDA algorithm in practical circumstances, we use five data sets from the UCI Machine Learning Repository, which have the sizes / dimensions shown in Table 1. We apply the original LPS to acquire the reference solution $w^*_{\mathcal{N}}$ of (3), with the tight optimality threshold of 10^{-6}. We then tabulate how many iterations of RDA are required before it generates a point in the optimal manifold $\mathcal{M}$ containing $w^*_{\mathcal{N}}$. We also check when the iterates of RDA reach a modest superset of the optimal manifold—a "2×" superset composed of the points in $\mathbb{R}^n$ having the same sign pattern for the active components in $\mathcal{M}$, and up to twice as many nonzeros as the points in $\mathcal{M}$. For each data set we use three values of λ equally spaced in the log-scale range of $[0.3, 0.9]\lambda_{\max}$, where $\lambda_{\max}$, computed accordingly to Koh et al. (2007), is the value beyond which the solution $w^*_{\mathcal{N}}$ has all zero components, except for the intercept term.

Table 1 shows performance of the RDA algorithm, over 100 repeated runs for each data set (using random permutations of training data for each run and for each sweep through the data), as measured by the number of iterations required for the algorithm to identify the optimal manifold and its 2× superset. Since the empirical distributions of the iteration counts are skewed, we show the median (rather than the mean) and the standard deviation. The table also shows the values of the optimality measure δ defined in (55) for the iterate at the moment we identify the optimal manifold. These results demonstrate that a huge number of iterations may be required to identify the optimal manifold, whereas identifying the superset is often much easier. In cases in which only a few components of $w^*_{\mathcal{N}}$ are nonzero, just a small fraction of the training examples usually suffice to identify the 2× superset. We note too that even when optimal identification is achieved, the iterate is still far from being optimal, by the criterion (55), suggesting the need for a local phase to achieve tighter optimality.

Data set	λ	$2\times$ Superset		Optimal $\mathcal{M}$		Optimality δ	NNZs $w^*_{\mathcal{N}}$
Glass ($m=214, n=10$)	0.29	14	(25)	20	(27)	0.068	1
	0.17	13	(10)	116	(428)	0.063	2
	0.10	13	(11)	28392	(6907)	0.016	3
Iono ($m=351, n=35$)	0.22	38	(84)	122	(95)	0.015	2
	0.13	44	(28)	30812	(15575)	0.008	3
	0.07	86	(41)	404	(150)	0.019	5
Arrhythmia ($m=452, n=280$)	0.15	192	(110)	304	(141)	0.001	2
	0.09	272	(88)	2036	(1076)	0.002	8
	0.05	447	(195)	27750	(4590)	0.001	13
Spambase ($m=4601, n=58$)	0.17	137	(219)	357	(325)	0.006	1
	0.10	722	(2495)	4340	(3097)	0.004	8
	0.06	812	(1247)	4680	(2209)	0.004	17
Pageblock ($m=5473, n=11$)	0.11	26	(326)	58	(395)	0.063	1
	0.07	182	(941)	524	(1233)	0.038	3
	0.04	103	(913)	461	(1232)	0.040	4

Table 1: Manifold identification properties of the RDA algorithm over 100 runs for each data set. The median number of iterations required to identify the optimal manifold $\mathcal{M}$, and the number required to identify a $2\times$ superset, are presented, along with the standard deviations over the 100 tests (in parentheses). δ represents the optimality measure at the moment of identifying $\mathcal{M}$, while the last column shows the number of nonzeros (excluding intercepts) in the reference solution obtained by LPS.

6.2 Performance on the MNIST Data Set

We now focus on the effects of the local phase on the performance of RDA$^+$. For this purpose, we use the MNIST data set which consists of gray-scale images of digits represented by 28×28 pixels. We choose the binary classification problem of distinguishing between the digits 6 and 7, for which the data set has 12183 training and 1986 test examples. Although the "6 vs 7" task is relatively easy, we choose this setting so that we can compare our results to those reported by Xiao (2010) for the original RDA algorithm.

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

$$[w_{t+1}]_i = [w_t]_i - \alpha_t \left([g_t]_i + \lambda \mathrm{sgn}\left([w_t]_i\right) \right), \; i = 1, 2, \ldots, n,$$

where g_t is a sampled approximation to the gradient of f at w_t, obtained from a single training example. The TG method (Langford et al., 2009) truncates the iterates obtained by the standard SGD at every Kth step (where K is a user-defined constant). That is,

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

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

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

We follow Xiao (2010) in using $\theta = \infty$ and $K = 10$.

For the stepsize α_t in SGD and TG, we adopt a variable stepsize scheme (Zinkevich, 2003; Nemirovski et al., 2009), choosing α_t to be a multiple of $1/\sqrt{t}$ so that the methods can achieve regret bounds of $O(\sqrt{t})$ similar to that of RDA.

In our implementations of RDA$^+$ and RDA, we set $\gamma = 5000$ in (20). (This value is determined by cross validation with RDA, using a single scan through the data set.) For LPS and the local phase of RDA$^+$, we compute a reduced Newton step on the current active manifold only when the number of nonzeros falls below 200. We also use the full set of training examples to compute the reduced gradient and reduced Hessian of the surrogate function $f_{\mathcal{N}}$.

For SGD, TG, and RDA, we keep track not only of the primal iteration sequence $\{w_t\}$, but also the primal averages $\bar{w}_T := \frac{1}{T}\sum_{t=1}^{T} w_t$, where T for each algorithm denotes the iteration number where the algorithm is stopped. We include these in the comparison because the convergence of the stochastic subgradient algorithms is often described in terms of the primal averages. Note that RDA$^+$ and LPS do not make use of primal averages.

We first run the RDA$^+$ algorithm 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, which is less than two complete sweeps through the data set.) In the safeguarding test (56), we use $\rho = 0.85$. We run the local phase of RDA$^+$ until the optimality measure in (55) falls below 10^{-4}. We record the total runtime of the RDA$^+$ algorithm, then run other algorithms SGD, TG, RDA, and LPS up to the runtime of RDA$^+$, stopping them earlier if they achieve the desired optimality before that point.

6.2.1 PROGRESS IN TIME

We compare the convergence of the algorithms in terms of the optimality measure and the number of nonzero components. Figure 1 presents the plots for the iterates without averaging, for three different values of λ: 10, 1, and 0.1. The optimality plots (on the left) show that RDA$^+$ achieves the target optimality much faster than other algorithms, including LPS. RDA behaves better than SGD and TG, but still does not come close to the target value of optimality. There is only a modest decrease in the optimality measure for SGD, TG, and RDA over the time frame of this experiment.

The plots on the right of Figure 1 show the number of nonzeros (excluding intercepts hereafter) 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, for the values $\lambda = 1.0$ and $\lambda = 0.1$.

In its local phase, RDA$^+$ behaves very similarly to LPS, sharing the typical behavior of non-monotonic decrease in the optimality measure (55). However, the local phase often converges faster than LPS, because it starts with the reduced space chosen by the dual-averaging phase of RDA$^+$. The number of nonzeros often increases at the point of switching between phases, as the safeguarding strategy adds more elements to the nonzero set.

Figure 2 shows similar plots but for the primal averaged iterates $\bar{w}_t$. We duplicate the plots of RDA$^+$ and LPS (which do not use iterate averaging) from Figure 1 for easy comparison. The number of nonzero components is clearly higher for averaged iterates.

6.2.2 QUALITY OF SOLUTIONS

In Figure 3, we compare the quality of the solutions in terms of optimality, the number of nonzeros, and test error rate. We present the results for the iterates without averaging in the three plots on the left, and those for the primal-averaged iterates (for algorithms SGD, TG, and RDA) in the plots on the right. (The plots of RDA$^+$ and LPS on the left are duplicated in the right-hand plots for easy comparison.) We run the algorithms with the same setting used in the previous experiments, except for LPS, which we run to optimality (10^{-4}, without time limit) to provide a baseline for comparison. (The runtime of LPS was about four times longer than that of RDA$^+$ on average.) The experiments are repeated for 100 runs of the data (using random permutations of training data for each run and for each sweep through the data), for each of seven λ values in the interval $[.01, 10]$. (The value of $\lambda_{\max}$ for this data set is 45.8.)

In Figure 3, 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 for 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 require a large number of nonzero components, despite the strong regularization imposed.

In Figure 4, we show typical solutions obtained from the various algorithms for different values of λ. The first three rows present the solutions acquired without averaging, and the last three rows present those obtained with primal averaging. The solutions from RDA$^+$ reveal almost identical sparsity pattern to those from the baseline algorithm LPS, achieving smallest nonzero patterns. RDA produces solutions of similar sparsity to RDA$^+$ for large λ values, but much denser solutions for smaller λ values. Again, we see that primal-averaged solutions are denser than those without averaging.

7. Conclusion

We have shown that under assumptions of nondegeneracy and strong local minimization at the optimum, the (non-averaged) iterates generated by the RDA algorithm identify the optimal manifold with probability approaching one as iterations proceed. This observation enables us to develop a new algorithmic framework that enjoys the low computational footprint of stochastic gradient methods as well as the rapid local convergence of Newton-type optimization techniques, which can be used to search for solutions on near-optimal manifolds that often have much lower intrinsic dimension than the full space.

We believe that our analysis can be extended in several directions. First, the use of ℓ_p norms in definition of the prox-function h can lead to faster convergence of RDA, but the interaction with the strong minimizer assumption and the manifold identification results that we use here is

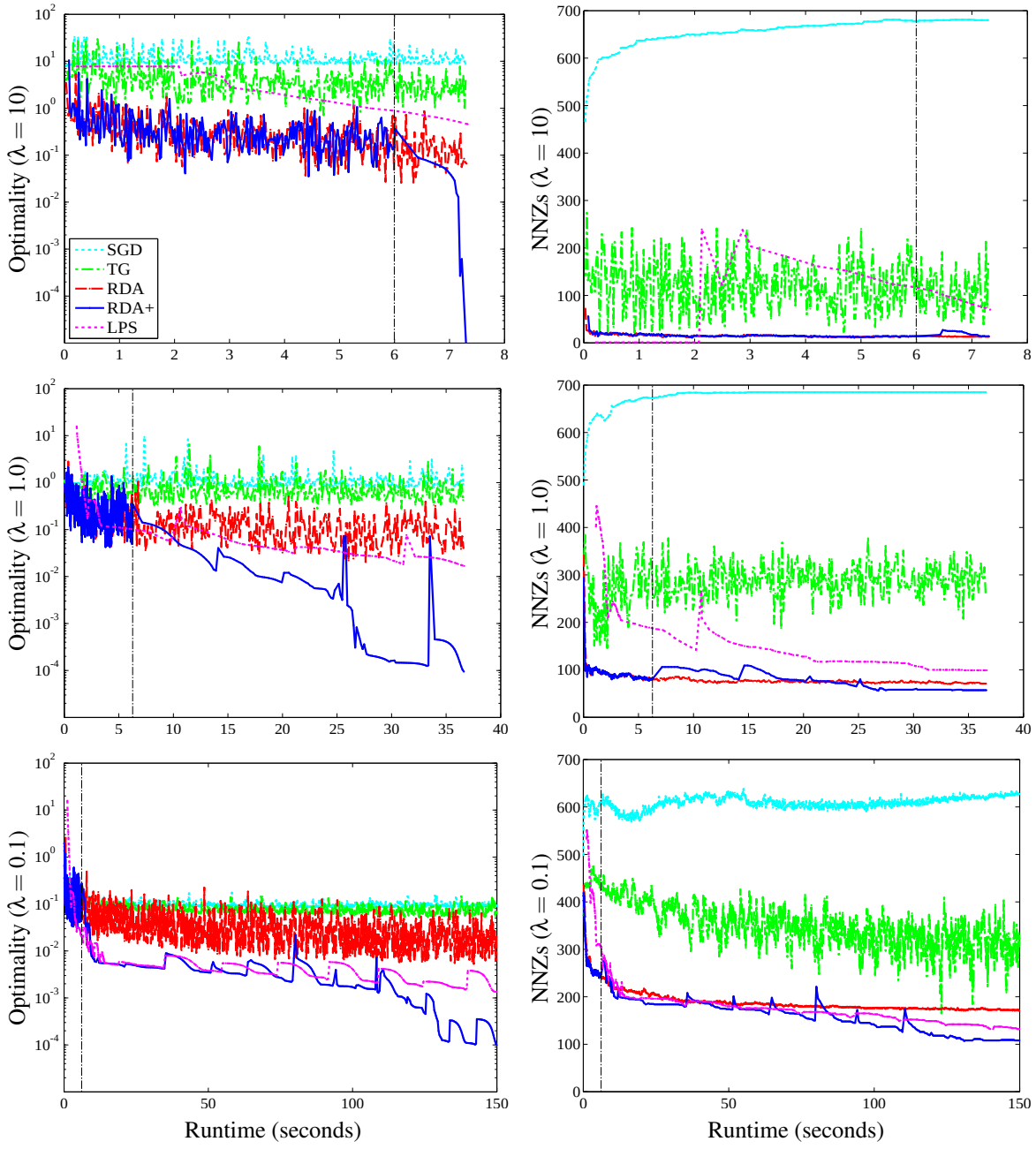

Figure 1: Convergence of iterates for various algorithms applied to an ℓ_1-regularized logistic regression function constructed from the digits 6 and 7 in the MNIST data set. Convergence is measured in terms of the optimality measure (left) and the number of nonzero components (excluding intercepts) 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 vertical lines indicate the event of phase switching in RDA$^+$. The vertical axes on the left are in logarithmic scale.

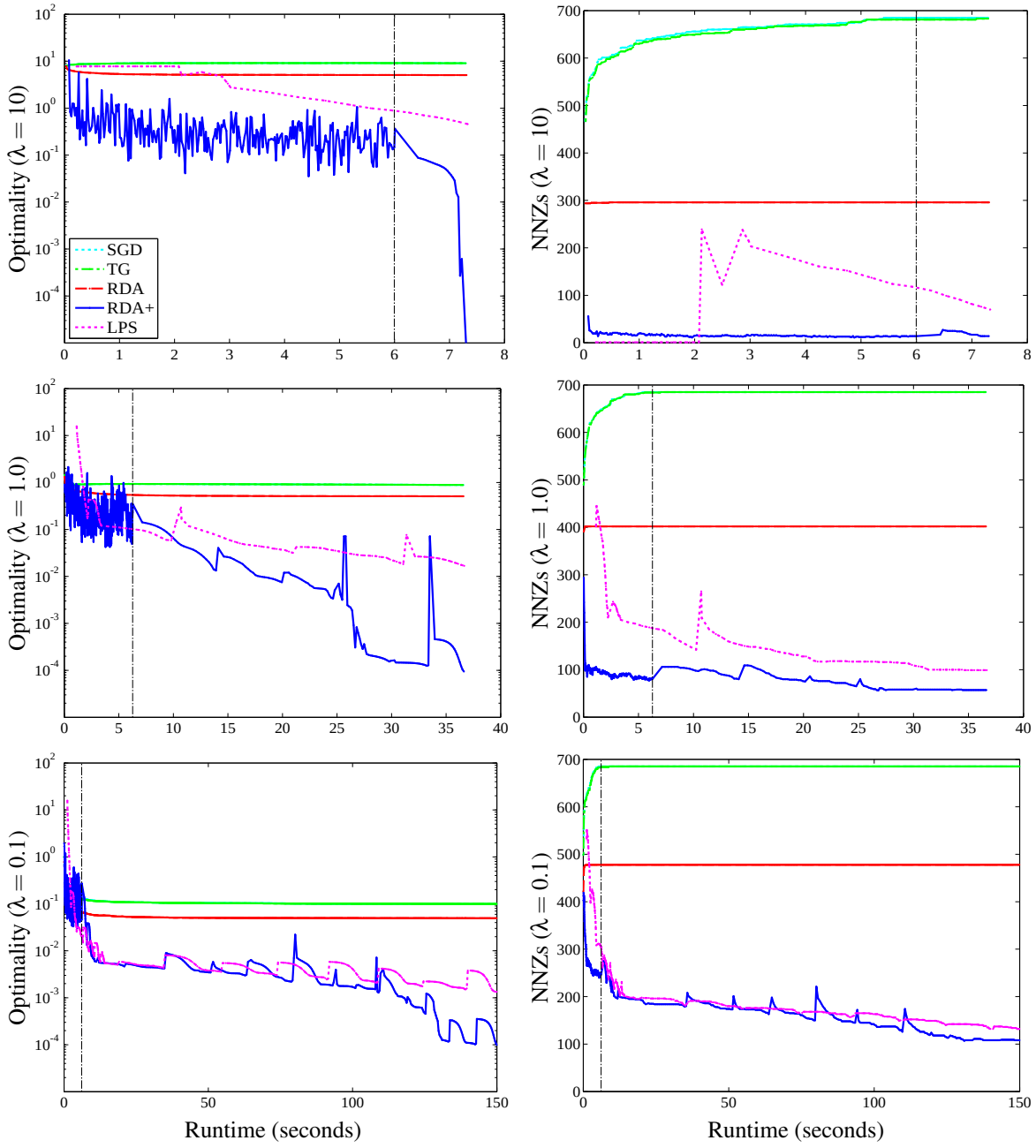

Figure 2: Convergence of averaged iterates (for SGD, TG, and RDA) and original iterates (for RDA$^+$ and LPS) on the ℓ_1-regularized logistic regression function constructed from the digits 6 and 7 in the MNIST data set. Convergence is measured in terms of the optimality measure (left) and the number of nonzero components (excluding intercepts) 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 vertical lines indicate the event of phase switching in RDA$^+$. The vertical axes on the left are in logarithmic scale.

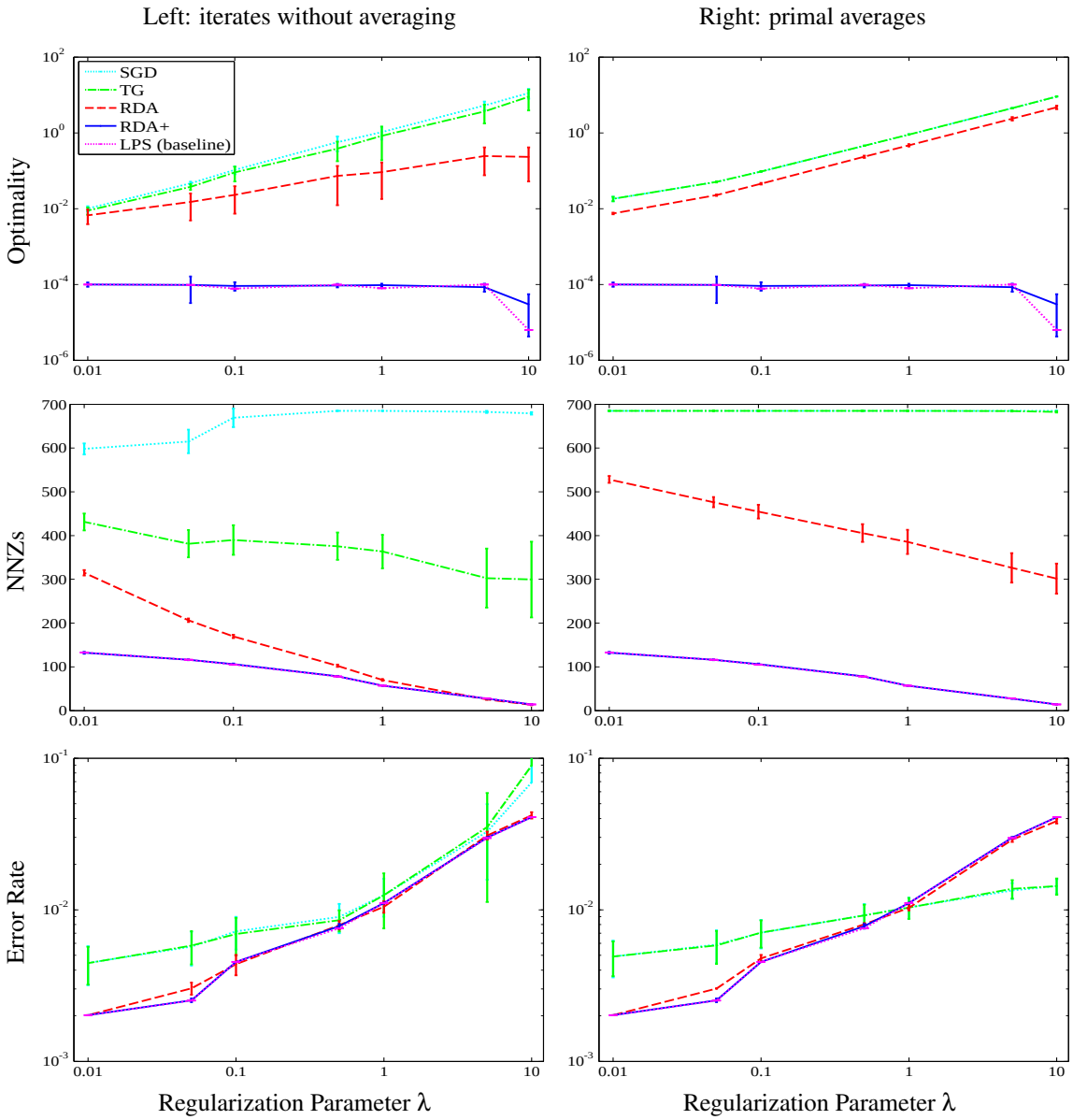

Figure 3: Quality of solutions (MNIST 6 vs. 7) in terms of the optimality, the number of nonzero components (without intercepts), and the test error rate, measured for 100 different random permutations of the training set. The plots on the left show the results for the iterates without averaging, and those on the right show averaged primal iterates for algorithms SGD, TG, and RDA, and non-averaged iterates for RDA$^+$ and LPS. The SGD, TG, and RDA algorithms are run up to the time taken for RDA$^+$ to achieve a 10^{-4} threshold in the measure (55), whereas LPS is run to convergence. Axes in the first and third rows are in logarithmic scale.

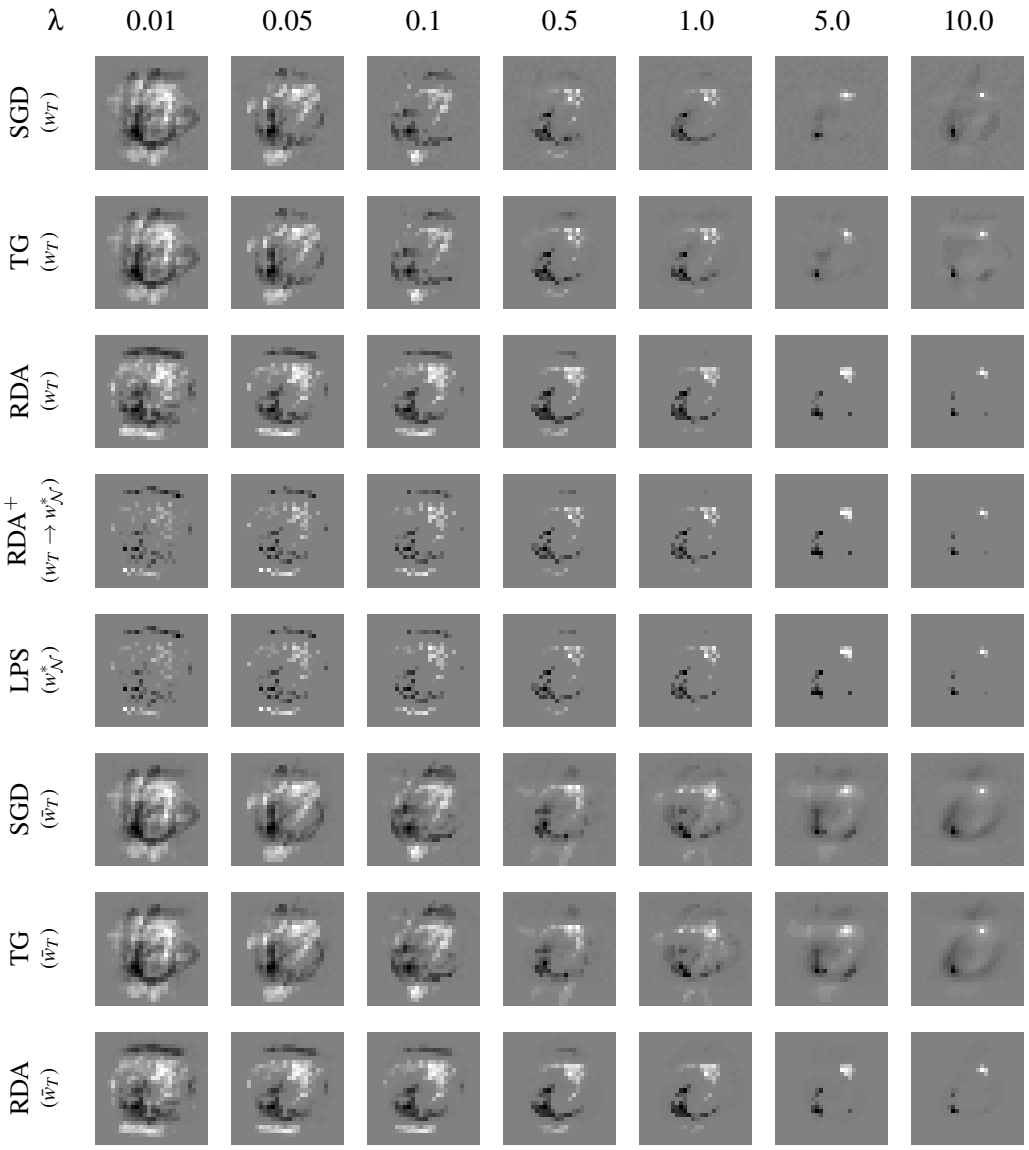

Figure 4: Sparsity patterns of the solutions for classification of the digits 6 and 7 in the MNIST data set. The regularization parameter λ is varied in the range of $[0.01, 10]$. The spots represent the positive (bright) and negative (dark) values, whereas the gray background represents zero. The top three rows show the solutions acquired without averaging, and the bottom three rows show those obtained with primal averaging. The two rows in the middle presents the solutions from RDA$^+$ and LPS. The algorithms SGD, TG, and RDA are run up to the time taken for RDA$^+$ to achieve a solution with 10^{-4} optimality value; the batch algorithm LPS is run without time limit. Note that for each value of λ, the sparsest solutions are obtained by RDA$^+$ and LPS.

complicated. Second, multiple samples could be used as in Dekel et al. (2012) to construct an approximate subgradient with reduced variance, which may thus lead to faster identification. Third, it is likely that the nondegeneracy assumption can be weakened, in which case a "super-manifold" of the optimal manifold would be identifiable. We leave such investigations to future work.

Acknowledgments

We thank our colleague Ben Recht for helpful comments and discussions. We are also grateful to the referees and the editor for their perceptive technical comments and their patient handling of the manuscript. We are particularly indebted to one of the anonymous referees, whose extensive, detailed comments and pointers led us to strengthen and simplify our analysis.

This research was supported in part by National Science Foundation Grant DMS-0914524, and in part by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project C1.

Appendix A. Strong Minimizer Property

In this section we prove Theorem 5 and its corollaries stated in Section 2.5, based on results from manifold analysis and other elementary arguments. Our proof is similar to that of Wright (2012, Theorem 2.5), but simpler.

We first state an elementary result on manifold characterization, which is proved by Vaisman (1984, Sections 1.4-1.5) and Wright (2012, Appendix A).

Lemma 18 *Let the manifold $\mathcal{M} \subset \mathbb{R}^n$ containing $\bar{z}$ be characterized by a $\mathcal{C}^p$ ($p \geq 2$) function $H : \mathbb{R}^n \to \mathbb{R}^k$. Then there is $\bar{y} \in \mathbb{R}^{n-k}$ and a $\mathcal{C}^p$ function G mapping some neighborhood of $\bar{y}$ to $\mathbb{R}^n$ such that $G(y) \in \mathcal{M}$ for all y near $\bar{y}$. Moreover, $G(y) - \bar{z} = Y(y-\bar{y}) + O(\|y-\bar{y}\|^2)$, where $Y \in \mathbb{R}^{n\times(n-k)}$ is an orthonormal matrix whose columns span the tangent space to $\mathcal{M}$ at $\bar{z}$.*

The next result, from Wright (2012, Lemma 2.2), shows how perturbations from a point at which the objective function is partly smooth can be decomposed according to the manifold characterization above.

Lemma 19 *Let the manifold $\mathcal{M} \subset \mathbb{R}^n$ be characterized in a neighborhood of $\bar{z} \in \mathcal{M}$ by $\mathcal{C}^p$ mappings $H : \mathbb{R}^n \to \mathbb{R}^k$ and $G : \mathbb{R}^{n-k} \to \mathbb{R}^n$ and the point $\bar{y}$ described in Lemma 18. Then for all z near $\bar{z}$, there are unique vectors $y \in \mathbb{R}^{n-k}$ and $v \in \mathbb{R}^k$ with $\|(y^T - \bar{y}^T, v^T)\| = O(\|z-\bar{z}\|)$ such that $z = G(y) + \nabla H(\bar{z})v$.*

We also make use of a result from Wright (2012, Lemma A.1).

Lemma 20 *Consider a function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, a point $\bar{z} \in \mathbb{R}^n$, and a manifold $\mathcal{M}$ containing $\bar{z}$ such that φ is partly smooth at $\bar{z}$ with respect to $\mathcal{M}$. If the nondegeneracy condition $0 \in \operatorname{ri} \partial\varphi(\bar{z})$ holds, then there exists $\varepsilon > 0$ such that*

$$\sup_{g \in \partial\varphi(\bar{z})} \langle g, d\rangle \geq \varepsilon\|d\|, \ \ \forall d \in N_{\mathcal{M}}(\bar{z}).$$

Proof (Theorem 5) We now proceed with the proof of the main result of Section 2.5. Recall the following assumptions:

(i) ϕ is partly smooth at w^* relative to the optimal manifold $\mathcal{M}$.

(ii) w^* is a locally strong minimizer of $\phi|_{\mathcal{M}}$ with modulus $c_{\mathcal{M}} > 0$ and radius $r_{\mathcal{M}} > 0$, and

(iii) the nondegeneracy condition (13) holds at w^*.

For the minimizer w^* of (1) and the optimal manifold $\mathcal{M}$ containing w^*, we consider the mappings H and G, the matrix Y, and the point $\bar{y} \in \mathbb{R}^{n-k}$ satisfying Lemma 18 and Lemma 19, associated with $\bar{z} = w^* \in \mathbb{R}^n$. From Lemma 19, for all w satisfying $\|w - w^*\| \leq \bar{r} \leq r_{\mathcal{M}}$ with small enough $\bar{r} > 0$, we can find unique vectors $y \in \mathbb{R}^{n-k}$ and $v \in \mathbb{R}^k$ with $\|(y^T - \bar{y}^T, v^T)\| = O(\|w - w^*\|)$ such that $w = G(y) + \nabla H(w^*)v$. Therefore we have

$$\phi(w) - \phi(w^*) = [\phi(G(y) + \nabla H(w^*)v) - \phi(G(y))] + [\phi(G(y)) - \phi(w^*)]. \tag{57}$$

From the locally strong minimizer property relative to $\mathcal{M}$ and the facts that $w^* \in \mathcal{M}$ and $G(y) \in \mathcal{M}$ for all y near $\bar{y}$, we have for the second bracketed term that

$$\phi(G(y)) - \phi(w^*) = \phi|_{\mathcal{M}}(G(y)) - \phi|_{\mathcal{M}}(w^*) \geq c_{\mathcal{M}} \|G(y) - w^*\|^2 \tag{58}$$

for all y near $\bar{y}$. Consider next the first bracketed term of (57). From Lemma 20, we have $\varepsilon > 0$ such that $\sup_{g \in \partial\phi(w^*)} \langle g, d \rangle \geq \varepsilon \|d\|$ for all $d \in N_{\mathcal{M}}(w^*)$. From the subcontinuity property (iv) of $\partial\phi(w^*)$ in Definition 4, we can choose a neighborhood of w^* sufficiently small that for all $G(y)$ in this neighborhood and $g \in \partial\phi(w^*)$, there exists $\hat{g} \in \partial\phi(G(y))$ such that $\|\hat{g} - g\| \leq \varepsilon/2$. These facts, together with convexity of ϕ, imply that for all y near $\bar{y}$ and v near 0 we have

$$\begin{aligned}
\phi(G(y) + \nabla H(w^*)v) - \phi(G(y)) &\geq \sup_{\hat{g} \in \partial\phi(G(y))} \langle \hat{g}, \nabla H(w^*)v \rangle \\
&\geq \sup_{g \in \partial\phi(w^*)} \langle g, \nabla H(w^*)v \rangle - (\varepsilon/2)\|\nabla H(w^*)v\| \\
&\geq (\varepsilon/2)\|\nabla H(w^*)v\|.
\end{aligned}$$

By substituting this inequality and (58) into (57), we obtain

$$\phi(w) - \phi(w^*) \geq (\varepsilon/2)\|\nabla H(w^*)v\| + c_{\mathcal{M}} \|G(y) - w^*\|^2.$$

By further reducing $\bar{r}$ if necessary, we can choose the neighborhood of w^* small enough to ensure that $\|\nabla H(w^*)v\| \leq 1$, and therefore

$$\begin{aligned}
\phi(w) - \phi(w^*) &\geq (\varepsilon/2)\|\nabla H(w^*)v\|^2 + c_{\mathcal{M}} \|G(y) - w^*\|^2 \\
&\geq \min(\varepsilon/2, c_{\mathcal{M}}) \left[\|\nabla H(w^*)v\|^2 + \|G(y) - w^*\|^2 \right] \\
&\geq \frac{1}{2} \min(\varepsilon/2, c_{\mathcal{M}}) \Big[\|\nabla H(w^*)v\| + \|G(y) - w^*\| \Big]^2 \\
&\geq \frac{1}{2} \min(\varepsilon/2, c_{\mathcal{M}}) \|w - w^*\|^2.
\end{aligned}$$

(The third inequality follows from the elementary bound $(a^2 + b^2) \geq \frac{1}{2}(a+b)^2$, for any scalars a and b.) We have thus shown that w^* indeed is a local strong minimizer of ϕ, without the restriction to the manifold $\mathcal{M}$, with modulus $c := \frac{1}{2}\min(\varepsilon/2, c_{\mathcal{M}})$ and radius $\bar{r}$. ■

We follow with the proofs of the remaining results of Section 2.5.

Proof (Corollary 6) Given $w \in \mathcal{O}$ with $\|w - w^*\| > \bar{r}$, we have from the convexity of ϕ that

$$\phi\left(w^* + \bar{r}\frac{w-w^*}{\|w-w^*\|}\right) \leq \phi(w^*) + \frac{\bar{r}}{\|w-w^*\|}(\phi(w) - \phi(w^*)).$$

From the locally strong minimizer property (Theorem 5), we also have

$$\phi\left(w^* + \bar{r}\frac{w-w^*}{\|w-w^*\|}\right) - \phi(w^*) \geq c\left\|\left(w^* + \bar{r}\frac{w-w^*}{\|w-w^*\|}\right) - w^*\right\|^2 = c\bar{r}^2.$$

Collecting the above two inequalities leads to the claim. ■

Proof (Corollary 7) Given $w \in \mathcal{O}$, if $\|w-w^*\| \leq \bar{r}$, then the claim follows from (15). If $\|w-w^*\| > \bar{r}$, then we have from strong convexity of ϕ that

$$\begin{aligned}\phi\left(w^* + \bar{r}\frac{w-w^*}{\|w-w^*\|}\right) \leq \phi(w^*) + \frac{\bar{r}}{\|w-w^*\|}(\phi(w) - \phi(w^*)) \\ - \frac{\sigma}{2}\frac{\bar{r}}{\|w-w^*\|}\left(1 - \frac{\bar{r}}{\|w-w^*\|}\right)\|w-w^*\|^2.\end{aligned}$$

From the locally strong minimizer property (Theorem 5), we also have

$$\phi\left(w^* + \bar{r}\frac{w-w^*}{\|w-w^*\|}\right) - \phi(w^*) \geq c\bar{r}^2.$$

Combining the above two inequalities results in

$$\phi(w) - \phi(w^*) \geq \left[\sigma/2 + \frac{\bar{r}}{\|w-w^*\|}(c - \sigma/2)\right]\|w-w^*\|^2 \geq \min(c, \sigma/2)\|w-w^*\|^2.$$

■

Appendix B. Expected Error Bounds for Iterates of RDA

In this section we provide the background for the results of Section 3, regarding the iterates generated by the RDA algorithm.

Proof (Lemma 9) For the general convex case, with $\{\beta_t\}$ chosen by (20), we consider the expected regret up to time t with respect to w^*, and obtain

$$\begin{aligned}
\mathbb{E}[R_t(w^*)] &= \mathbb{E}\left[\sum_{j=1}^{t}(F(w_j;\xi_j)+\Psi(w_j)) - \sum_{j=1}^{t}(F(w^*;\xi_j)+\Psi(w^*))\right] \\
&= \sum_{j=1}^{t}\mathbb{E}\left[\mathbb{E}\left\{(F(w_j;\xi_j)+\Psi(w_j)-F(w^*;\xi_j)-\Psi(w^*))\,|\,\xi_{[j-1]}\right\}\right] \\
&= \sum_{j=1}^{t}\mathbb{E}\left[f(w_j)+\Psi(w_j)-f(w^*)-\Psi(w^*)\right] \\
&= \sum_{j=1}^{t}\mathbb{E}\left[\phi(w_j)-\phi(w^*)\right]. \qquad (59)
\end{aligned}$$

Under Assumptions 1 and 2, there exists $c>0$ and $\bar{r}>0$ that satisfy (15), which is

$$\phi(w)-\phi(w^*) \geq c\|w-w^*\|^2, \text{ for all } w \text{ with } \|w-w^*\| \leq \bar{r},$$

as proved in Theorem 5. Noting that $I_{(\|w_j-w^*\|\leq\bar{r})}+I_{(\|w_j-w^*|>\bar{r})}=1$, we can split the right-hand side of (59) into two sums and obtain

$$\mathbb{E}[R_t(w^*)] = \sum_{j=1}^{t}\mathbb{E}\left[I_{(\|w_j-w^*\|\leq\bar{r})}\{\phi(w_j)-\phi(w^*)\}\right] + \sum_{j=1}^{t}\mathbb{E}\left[I_{(\|w_j-w^*\|>\bar{r})}\{\phi(w_j)-\phi(w^*)\}\right]. \qquad (60)$$

Note that both terms on the right-hand side of (60) are nonnegative. For the first term, we have by using the regret bound (22) and the locally strong minimizer property (15) that

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

proving the first inequality (24). For the second inequality, we have from (60), the regret bound (22), and Corollary 6 that

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

thus proving (25).

When Ψ is strongly convex with the modulus $\sigma>0$ and $\{\beta_t\}$ chosen by (21), we apply the other regret bound (23) to (59), resulting in

$$\frac{G^2}{2\sigma}(6+\ln t) \geq \mathbb{E}R_t(w^*) \geq \sum_{j=1}^{t}\mathbb{E}\{\phi(w_j)-\phi(w^*)\} \geq \min(c,\sigma/2)\sum_{j=1}^{t}\mathbb{E}\left[\|w_j-w^*\|^2\right],$$

where for the last inequality we use the fact that w^* is a (global) strong minimizer with the modulus $\min(c, \sigma/2)$, as shown in Corollary 7. This proves (26).

■

Proof (Theorem 10) We start with the general convex case. From the Cauchy-Schwartz inequality $\|z\|_1 \leq \sqrt{m}\|z\|_2$ for a vector $z \in \mathbb{R}^m$ and Jensen's inequality, we have

$$\begin{aligned}\frac{1}{t}\sum_{j=1}^{t}\mathbb{E}\left[I_{(\|w_j-w^*\|\leq\bar{r})}\|w_j-w^*\|\right] &\leq \frac{\sqrt{t}}{t}\left[\sum_{j=1}^{t}\left\{\mathbb{E}\left[I_{(\|w_j-w^*\|\leq\bar{r})}\|w_j-w^*\|\right]\right\}^2\right]^{1/2}\\ &\leq \left[\frac{1}{t}\sum_{j=1}^{t}\mathbb{E}\left[I_{(\|w_j-w^*\|\leq\bar{r})}\|w_j-w^*\|^2\right]\right]^{1/2}\\ &\leq \frac{1}{\sqrt{c}}\left(\gamma D^2+\frac{G^2}{\gamma}\right)^{1/2}t^{-1/4},\end{aligned}$$

where the last inequality is from (24). By combining this result with (25) and the definition of $\hat{t}$ in (28), and using our standing assumption that $\bar{r} \in (0,1]$, we have for $t \geq \hat{t}$ that

$$\begin{aligned}\frac{1}{t}\sum_{j=1}^{t}\mathbb{E}\|w_j-w^*\| &= \frac{1}{t}\sum_{j=1}^{t}\mathbb{E}\left[I_{(\|w_j-w^*\|\leq\bar{r})}\|w_j-w^*\|\right]+\frac{1}{t}\sum_{j=1}^{t}\mathbb{E}\left[I_{(\|w_j-w^*\|>\bar{r})}\|w_j-w^*\|\right]\\ &\leq \frac{1}{\sqrt{c}}\left(\gamma D^2+\frac{G^2}{\gamma}\right)^{1/2}t^{-1/4}+\frac{1}{\bar{r}c}\left(\gamma D^2+\frac{G^2}{\gamma}\right)t^{-1/2}\\ &\leq \frac{1}{\sqrt{c}}\left(\gamma D^2+\frac{G^2}{\gamma}\right)^{1/2}t^{-1/4}+\frac{1}{\sqrt{\bar{r}c}}\left(\gamma D^2+\frac{G^2}{\gamma}\right)^{1/2}t^{-1/4}\\ &\leq \mu t^{-1/4}.\end{aligned}$$

for μ defined in (28).

For the strongly convex case, we have from Cauchy-Schwarz and Jensen's inequalities that

$$\frac{1}{t}\sum_{j=1}^{t}\mathbb{E}\|w_j-w^*\| \leq \left[\frac{1}{t}\sum_{j=1}^{t}\left\{\mathbb{E}\|w_j-w^*\|\right\}^2\right]^{1/2} \leq \left[\frac{1}{t}\sum_{j=1}^{t}\mathbb{E}\|w_j-w^*\|^2\right]^{1/2}.$$

Applying the bound in (26) to the last line leads to (29). ■

References

M. Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000.

K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967.

P. L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems 20*, pages 65–72, 2008.

L. Bottou. Stochastic learning. In *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Artificial Intelligence*, pages 146–168. Springer Verlag, 2004.

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

K. L. Chung. On a stochastic approximation method. *Annals of Mathematical Statistics*, 25(3): 463–483, 1954.

O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165–202, 2012.

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 2009.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the 23rd Annual Conference on Learning Theory*, 2010.

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

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

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23(3):462–466, 1952.

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.

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

A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13: 702–725, 2003.

A. S. Lewis and S. J. Wright. A proximal method for composite minimization. Technical report, University of Wisconsin-Madison, August 2008.

Q. Lin, X. Chen, and J. Peña. A sparsity preserving stochastic gradient method for composite optimization. Working paper, Tepper School of Business, Carnegie Mellon University, March 2011. Revised May 2012.

A. Nemirovski and D. Yudin. On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. *Soviet Mathematics-Doklady*, 19, 1978.

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

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.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

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

C. Oberlin and S. J. Wright. Active set identification in nonlinear programming. *SIAM Journal on Optimization*, 17(2):577–605, 2006.

B. T. Polyak. New stochastic approximation type procedures. *Avtomatica i Telemekhanika*, 7: 98–107, 1990.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

J. Sacks. Asymptotic distribution of stochastic approximation procedures. *Annals of Mathematical Statistics*, 29(2):373–405, 1958.

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*, pages 807–814, 2007.

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

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

S. J. Wright. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31(4):1063–1079, 1993.

S. J. Wright. Accelerated block-coordinate relaxation for regularized optimization. *SIAM Journal on Optimization*, 22(1):159–186, 2012.

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

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

T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

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