# ACCELERATED BLOCK-COORDINATE RELAXATION FOR REGULARIZED OPTIMIZATION[*]

STEPHEN J. WRIGHT[†]

**Abstract.** We discuss minimization of a smooth function to which is added a separable regularization function that induces structure in the solution. A block-coordinate relaxation approach with proximal linearized subproblems yields convergence to critical points, while identification of the optimal manifold (under a nondegeneracy condition) allows acceleration techniques to be applied on a reduced space. The work is motivated by experience with an algorithm for regularized logistic regression, and computational results for the algorithm on problems of this type are presented.

**Key words.** regularized optimization, block coordinate relaxation, active manifold identification

**AMS subject classifications.** 49K40, 49M27, 90C31

## 1. Introduction.

We discuss an algorithm for solving the problem

$$\min_{x \in \mathbf{R}^n} \phi_\tau(x) := f(x) + \tau P(x), \tag{1.1}$$

where $f$ is smooth (at least locally Lipschitz continuously differentiable in the region of interest) and $\tau > 0$ is a parameter. The regularization function (or "regularizer") $P$ has the following separable structure:

$$P(x) = \sum_{q \in Q} P_q(x_{[q]}), \tag{1.2}$$

where each $P_q$ is a closed, proper, extended-valued, convex function; $[q]$ denotes a subset of $\{1, 2, \ldots, n\}$; and $\{x_{[q]} \,|\, q \in Q\}$ is a partition of the components of $x$. Not all components of $x$ need be involved in the regularization function; we may have $P_q \equiv 0$ for some $q$.

Problems of the form (1.1) are appearing in many applications. Several (overlapping) problems in this class have been particularly well studied in recent times.

- Compressed sensing, where in the $\ell_2$-$\ell_1$ formulation, $f(x) = (1/2)\|Ax - y\|_2^2$ for some $A \in \mathbf{R}^{m \times n}$ with $m \ll n$ and special properties such as restricted isometry, and $P(x) = \|x\|_1$.
- Regularized logistic regression, where $f$ is a log-likelihood function obtained from labeled training data and the regularizer $P(x)$ is either $\|x\|_1$ [23, 21, 22] (possibly modified by omission of one or more terms from the norm) or a group-$\ell_2$ regularizer [17].
- Regularized least squares, where again $f$ is a linear least-squares function while $P$ could be $\|\cdot\|_1$ (leading to the LASSO estimator [24]), a group-$\ell_2$ regularizer (with $P_q(z) = \|z\|_2$ in the notation above; see [35, 7]), or a group $\ell_\infty$ regularizer (with $P_q(z) = \|z\|_\infty$ [30]). The group regularizers allow variables to be partitioned into subsets of closely related effects, where each subset is selected or deselected as a group.

[*] Version of September 2, 2011. Research supported by NSF Grants DMS-0914524 and DMS-0906818, NIH Grant EY009946, and DOE Grant DE-SC0002283.

[†] Computer Sciences Department, 1210 W. Dayton Street, University of Wisconsin, Madison, WI 53706, USA. swright@cs.wisc.edu

Other works [12, 13, 34] consider algorithms for both logistic and least-squares loss functions jointly.

In the examples above, the regularizer $P$ is nonsmooth, but such is not always the case. For example, in one formulation of the matrix completion problem of estimating a low-rank matrix $X \in \mathbb{R}^{m\times p}$ such that $\mathcal{A}(X) \approx b$, for noisy observations $b \in \mathbb{R}^l$ and a given linear operator $\mathcal{A} : \mathbb{R}^{m\times p} \to \mathbb{R}^l$, we can explicitly model $X$ as a product of factors $L$ and $R^T$ (for $L \in \mathbb{R}^{m\times r}$ and $R \in \mathbb{R}^{p\times r}$) and solve the following problem:

$$\min_{L,R} \frac{1}{2}\|\mathcal{A}(LR^T) - b\|_2^2 + \tau(\|L\|_F^2 + \|R\|_F^2).$$

(This formulation is an obvious consequence of [19, Subsection 5.3].)

In this paper we examine a block coordinate proximal linearization algorithm for solving (1.1), in which at the current iterate $x$ we select a subset $Q' \subset Q$ and solve the following subproblem for some $\mu \geq \mu_{\min} > 0$:

$$\min_d \sum_{q\in Q'} \nabla_{[q]}f(x)^T d_{[q]} + \frac{\mu}{2}|d|^2 + \tau \sum_{q\in Q'} P_q(x_{[q]} + d_{[q]}), \tag{1.3}$$

where $\nabla_{[q]}f$ denotes the gradient subvector corresponding to the components of $x$ that belong to the partition $[q]$, and here and throughout $|\cdot|$ denotes the Euclidean norm. We refer to $Q'$ as the *relaxation set* at $x$. Clearly, the solution of (1.3) has $d_{[q]} = 0$ for all $q \notin Q'$. Existence and uniqueness of a solution to (1.3) is immediate for $\mu > 0$, given our assumptions on the functions $P_q$, $q \in Q$. The optimality conditions are

$$0 \in \nabla_{[q]}f(x) + \mu d_{[q]} + \tau\partial P_q(x_{[q]} + d_{[q]}), \qquad \text{for all } q \in Q', \tag{1.4}$$

where $\partial$ denotes the subdifferential of a convex function.

If the solution of (1.3) produces a "sufficient decrease" in the objective $\phi_\tau$, the step is accepted. Otherwise $\mu$ is increased and (1.3) is re-solved, with the same relaxation set $Q'$; this process is repeated until a valid step is found. At iteration $k$, we denote this step by $d^k$ and refer to it as a *prox-descent step.* We then optionally compute an *enhanced step* $\tilde{d}^k$ that is required to improve on $d^k$ and to satisfy its own sufficient decrease conditions.

Denoting the relaxation set $Q'$ at iteration $k$ by $Q_k$, we require the following "generalized Gauss-Seidel" condition to be satisfied (see [28]): For some integer $T \geq 1$ and all $k \geq T-1$, we have

$$Q_k \cup Q_{k-1} \cup \cdots \cup Q_{k-T+1} = Q. \tag{1.5}$$

This condition ensures that every component of $x$ is "touched" by the relaxation scheme at least once every $T$ iterations.

Effectiveness of the algorithm depends strongly on whether the subproblems (1.3) can be formulated and solved efficiently. Such is the case in the applications mentioned above. When $P_q(z) = \|z\|_1$, the solution can be obtained by the "soft thresholding" operator at a cost that is linear in the total number of components in $d_{[q]}$, $q \in Q'$. The costs are similar when $P_q(z) = \|z\|_2$ or $P_q(z) = \|z\|_2^2$.

The basic approach has been well studied in the case of $Q_k = Q$ for all $k$; see for example the SpaRSA approach for compressed sensing [33] and the ProxDescent framework of [16]. The block coordinate technique often gives better practical performance when $Q_k$ encompasses only a small fraction of the components of $x$ and where the cost of computing the corresponding partial gradient is small compared to the cost of the full gradient. For the regularized logistic regression problem of [23], for example, the cost of evaluating the partial gradient is approximately linear in the number of components of the gradient required. We note that "Gauss-Southwell" rules for choosing $Q_k$ (see [28]) are less appealing than (1.5) because they appear to require knowledge of the full gradient.

This paper analyzes the convergence properties of a simple algorithm based on (1.3), in which the relaxation sets are required only to satisfy the mild condition (1.5). We describe in particular the global convergence properties of the method and its ability to identify the partly smooth manifold on which a critical point $x^*$ lies (provided that the $P_q$ satisfy appropriate properties at $x^*$). This identification property leads us to propose a specific acceleration strategy for choosing the enhanced step $\tilde{d}^k$: a reduced Newton method on the partly smooth manifold. We prove a superlinear convergence result for this strategy. We discuss how the algorithm can be implemented on several specific problems of the type (1.1), and finally focus on the important case of $\ell_1$-regularized logistic regression. An implementation is described in some detail and illustrative computational results are presented.

### 1.1. Related Work.

We discuss here some recent work on related algorithms and applications. Other related work can be found in the bibliographies of the papers mentioned here.

The SpaRSA approach for compressed sensing [33] solves subproblems of the form (1.3) in which $f$ is a linear least-squares objective, $P(x) = \|x\|_1$, each partition $[i]$ contains the single index $i$ (for $i = 1, 2, \ldots, n$), and $Q' = Q$. (In many applications of compressed sensing, there is little to be saved by a partial gradient evaluation.) SpaRSA has no "enhanced" step analogous to $\tilde{d}^k$ in Algorithm 1. In a more general setting, Lewis and Wright [16] describe an algorithm for (1.1) in which $P$ is allowed to be non-separable and prox-regular (rather than separable and convex, as here). The subproblems have the form (1.3) but again are solved on the full space, and enhanced steps are not considered in any detail. (Global convergence results are proved and there is some discussion of manifold identification.)

Another related line of work begins with the paper of Tseng and Yun [28], which considers the same problem (1.1) as this paper and solves block-relaxation subproblems like (1.3), but with a more general scaling in the quadratic term. A line search is performed along the direction $d^k$. (The algorithm described in the current paper is simpler in that it uses the quadratic term in (1.3) to modify both the direction and length of the step.) Enhanced steps are not considered explicitly in [28], though second-order information could eventually be used in the quadratic term to enhance the final convergence rate.

In later work [29], Tseng and Yun described a block relaxation approach for minimization of a smooth function subject to linear constraints (but without the separable regularization term of (1.1)). This problem is not unrelated to (1.1); we can always enforce bound constraints, for example, by defining a separable $P(x)$ that takes on the value 0 when $x$ satisfies its bounds and $\infty$ otherwise. Special attention is paid to the quadratic programming formulation of support vector machines. Another paper by the same authors [27] considers (1.1) with the addition of linear equality constraints, and describes a similar method. An application to bi-level optimization is described. One of Tseng's final papers [25] mentions the problem (1.1) again and outlines the algorithms presented in the earlier works, leaving open the question of whether accelerated first-order methods can be applied in a block relaxation context.

The current paper was motivated by the need to provide theoretical support for the algorithm described in Shi et al. [23] for $\ell_1$-regularized logistic regression. In the event, significant modifications were needed to make the method fit the framework of the present paper. We describe this application in detail in Section 5, and give computational results obtained with an implementation of the algorithm presented here.

There is a recent and extensive literature on least-squares and logistic regression problems with nonsmooth regularizers. We mention a few such contributions here, with a focus on methods that are suitable for large problems.

Shevade and Keerthi [21] present a relaxation algorithm for $\ell_1$-regularized logistic regression which selects one component at a time for relaxation — the one that violates the optimality conditions maximally. Friedman, Hastie, and Tibshirani [7] discuss a coordinate descent method for the $\ell_1$-regularized least-squares loss (with an additional penalty term involving $\|x\|_2^2$ – the "elastic net" formulation). They describe a cyclic coordinate relaxation method with exact line search along each direction. For logistic-regression loss, they propose a sequential-quadratic-programming outer loop, solving the quadratic subproblem with the method for least-squares loss. Koh, Kim, and Boyd [14] reformulate $\ell_1$-regularized logistic regression as a linearly constrained smooth optimization problem, and apply an interior-point method that uses conjugate gradients to compute the steps. The method of Shi et al. [22] for $\ell_1$-regularized logistic regression computes steps as in (1.3), but for the full set of components and for a fixed (large) value of $\mu_k$. An Armijo line search is added, as is a continuation strategy in the parameter $\tau$ (discussed further in Section 5). When the correct nonzero set appears to have been identified, the algorithm uses second-order steps like those in [14]. Yuan et al. [34] present an extensive and useful survey of various methods for $\ell_1$-regularized regression, using least-squares, logistic loss functions, and various extensions. Computational comparisons between different approaches are reported in detail.

An early reference to group-$\ell_2$ regularizers is the thesis of Bakin [1]. Yuan and Lin [35] discuss group-$\ell_2$ regularizer with least-squares loss function. An extension of LARS [4] is proposed to solve it, but no convergence analysis is presented. Kim, Kim, and Kim [13] describe a gradient projection approach for the formulation in which regularization is imposed as a constraint of the form $P(x) \leq \beta$, rather than included in the objective, showing that the projection can be performed efficiently when $P$ is the group-$\ell_2$ regularizer. Meier, van de Geer, and Bühlmann [17] consider logistic regression with a group-$\ell_2$ regularizer, applying variants of coordinate descent and the method from [26] to a problem in DNA splice site detection.

Turlach, Venables, and Wright [30] describe an application and computational results for an interior-point method for a least-squares objective with a group-$\ell_\infty$ regularizer, and give a statistical analysis for a special case involving orthonormal coefficient matrices.

**1.2. Outline.** We outline briefly the remainder of the paper. Section 2 reviews the relevant optimality and nondegeneracy conditions, discusses manifolds and their characterization, and defines partial smoothness. The relationship of second-order optimality conditions to strong local minimizers is explored in Subsection 2.3; this topic is useful when we introduce the reduced-Newton acceleration scheme in Subsection 3.4.

The proximal block coordinate relaxation algorithm is introduced and analyzed in Section 3. Global convergence results are obtained in Subsection 3.2, and identification of the (assumed) partly smooth manifold on which the limit point lies is analyzed in Subsection 3.3. The reduced-Newton acceleration scheme is described and analyzed in Subsection 3.4.

Section 4 outlines how the algorithm could be applied to several nonsmooth regularization functions that have been proposed in the recent literature. In Section 5, we describe an application to $\ell_1$-regularized logistic regression, giving details on the implementation and presenting computational results on three test problems with somewhat different properties. These tests suggest that there is much to be gained, computationally speaking, from using higher-order acceleration on the apparently optimal manifold, and from judicious implementation of a continuation strategy in the regularization parameter $\tau$. The code and test data used in Section 4 is available at `http://www.cs.wisc.edu/~swright/LPS/`, to allow reproduction of the tables in this paper (up to the effects of randomness in the algorithm, which are significant).

**2. Optimality and Nondegeneracy Conditions.** In this section, we discuss properties of the objective function $\phi_\tau$ in the neighborhood of a solution $x^*$. Subsection 2.1 discusses criticality conditions. Subsection 2.2 discusses manifolds and their characterization, and defines the partially smooth manifolds of Lewis [15]. Subsection 2.3 discusses second-order conditions and shows how the notion of a strong local minimizer $x^*$ of $\phi_\tau$ is tied to the second-order sufficient conditions of the restriction of this function to the partly smooth manifold containing $x^*$.

Provided that $\nabla f$ is locally Lipschitz at a point $x$, $\phi_\tau$ is prox-regular at $x$. In fact, we can find $\sigma > 0$ such that for all $x'$ and $x''$ close to $x$, we have

$$\phi_\tau(x'') \geq \phi_\tau(x') + g^T(x'' - x') - \sigma|x'' - x'|^2, \qquad \text{for each } g \in \partial\phi_\tau(x'), \tag{2.1}$$

where, by smoothness of $f$, we have $\partial\phi_\tau(x) = \nabla f(x) + \tau\partial P(x)$. Note that in (2.1), in contrast to general prox-regular functions, $\sigma$ is independent of $g$; it can simply be chosen as the local Lipschitz constant for $\nabla f$ near $x$.

**2.1. Criticality and Optimality.** We say that $z^*$ is a *strong local minimizer* of a function $h : \mathbf{R}^r \to \bar{\mathbf{R}}$ with modulus $c > 0$ if

$$h(z) \geq h(z^*) + c|z - z^*|^2 + o(|z - z^*|^2) \quad \text{for all } z \text{ near } z^*. \tag{2.2}$$

(We use $\bar{\mathbf{R}}$ throughout to denote the extended reals $[-\infty, +\infty]$.) The criticality condition for $h$ is $0 \in \partial h(z^*)$, and the nondegeneracy condition is $0 \in \operatorname{ri} \partial h(z^*)$, where ri denotes the relative interior of a set. For the particular function $\phi_\tau$ of (1.1), we have

$$g \in \partial\phi_\tau(x) \quad \Leftrightarrow \quad g_{[q]} \in \nabla_{[q]} f(x) + \tau\partial P_q(x_{[q]}), \;\; \text{for all } q \in Q.$$

(This claim follows from smoothness of $f$, the closed proper convex nature of each $P_q$, the fact that $\{[q] \,|\, q \in Q\}$ is a partition of $\{1, 2, \dots, n\}$, and elementary results from Rockafellar [20], especially Theorem 23.8.) Thus, the criticality condition $0 \in \partial\phi_\tau(x^*)$ can be stated equivalently as follows:

$$0 \in \nabla_{[q]} f(x^*) + \tau\partial P_q(x^*_{[q]}), \qquad \text{for all } q \in Q, \tag{2.3}$$

while nondegeneracy at $x^*$ can be written in a similar partitioned form:

$$0 \in \operatorname{ri} \left[\nabla_{[q]} f(x^*) + \tau\partial P_q(x^*_{[q]})\right], \;\; \text{for all } q \in Q. \tag{2.4}$$

**2.2. Manifolds and Partial Smoothness.** We start our discussion of manifolds by repeating some definitions from Hare and Lewis [9, Definition 2.3] and Lewis and Wright [16, Definition 1.2]. We also use some notation and terminology regarding manifolds from Vaisman [31].

A set $\mathcal{M} \subset \mathbf{R}^m$ is a *manifold about* $\bar{z} \in \mathbf{R}^m$ 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 $F : \mathbf{R}^m \to \mathbf{R}^k$ that is $\mathcal{C}^p$ around $\bar{z}$ with $\nabla F(\bar{z})^T \in \mathbf{R}^{k \times m}$ surjective and such that points $z$ near $\bar{z}$ lie in $\mathcal{M}$ if and only if $F(z) = 0$. The *normal space* to $\mathcal{M}$ at $\bar{z}$, denoted as usual by $N_{\mathcal{M}}(\bar{z})$, is then the range of $\nabla F(\bar{z})$, while the *tangent space* to $\mathcal{M}$ at $\bar{z}$ is the null space of $\nabla F(\bar{z})^T$.

It is convenient here and later to assume that $\nabla F(\bar{z})$ is a matrix with orthonormal columns. This assumption can be made without loss of generality, by performing a $QR$ factorization of $\nabla F(\bar{z})$ (with $Q$ being $m \times k$ orthonormal and $R$ being $k \times k$ upper triangular) and replacing $F(z)$ by $R^{-1}F(z)$ (thus replacing $\nabla F(z)$ by $\nabla F(z)R^{-1}$).

We state an elementary technical result on manifold parametrization, which is essentially proved in [31, Sections 1.4-1.5]. A simple proof appears for completeness in Appendix A.

Lemma 2.1. *Let the manifold $\mathcal{M} \subset \mathbf{R}^m$ containing $\bar{z}$ be characterized by a $C^p$ function $F : \mathbf{R}^m \to \mathbf{R}^k$ with the properties described above, where $p \geq 2$. Then there is a point $\bar{y} \in \mathbf{R}^{m-k}$ and a $C^p$ function $G$ mapping some neighborhood of $\bar{y}$ to $\mathbf{R}^m$ 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 $Y \in \mathbf{R}^{m \times (m-k)}$ is an orthonormal matrix whose columns span the tangent space to $\mathcal{M}$ at $\bar{z}$.*

The proof constructs $G$ using the implicit function theorem and sets $\bar{y} = 0$. In practice, we can often identify a suitable $G$ by making use of the structure of $P$ (as demonstrated later) but we do not know (or need to know) $\bar{y}$ until the solution of the problem is known. Hence, it is useful to state and use the results in this section without assuming $\bar{y} = 0$.

It follows immediately from Lemma 2.1 that $\nabla G(\bar{y})^T = Y$ and that $\begin{bmatrix} \nabla F(\bar{x}) & Y \end{bmatrix}$ is an $m \times m$ orthogonal matrix.

The next technical result shows how perturbations from a point at which $h$ is partly smooth can be decomposed according to the manifold characterization above. A proof appears in Appendix A.

Lemma 2.2. *Let the manifold $\mathcal{M} \subset \mathbf{R}^m$ be characterized in a neighborhood of $\bar{z} \in \mathcal{M}$ by $C^p$ mappings $F : \mathbf{R}^m \to \mathbf{R}^k$ and $G : \mathbf{R}^{m-k} \to \mathbf{R}^m$ and the point $\bar{y}$ described above. Then for all $z$ near $\bar{z}$, there are unique $y(z) \in \mathbf{R}^{m-k}$ and $v(z) \in \mathbf{R}^k$ such that $z = G(y(z)) + \nabla F(\bar{z})v(z)$, and moreover $(y(z), v(z))$ is a $C^p$ function of $z$.*

Partial smoothness can now be defined as follows [15, Section 2].

Definition 2.3. *A function $h : \mathbf{R}^m \to \bar{\mathbf{R}}$ is* partly smooth *at a point $\bar{z} \in \mathbf{R}^m$ relative to a set $\mathcal{M} \subset \mathbf{R}^m$ containing $\bar{z}$ if $\mathcal{M}$ is a manifold about $\bar{z}$ and the following properties hold:*

(i) (Smoothness) *The restricted function $h\,|_{\mathcal{M}}$ is $\mathcal{C}^2$ near $\bar{z}$;*

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

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

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

*We refer to $\mathcal{M}$ as the* active manifold.

Since $f$ is smooth, it does not complicate the definition of active manifolds for $\phi_\tau$ at $x^*$; we need examine only the function $P$. Additionally, the structure (1.2) of $P$

ensures that we can express the active manifold as a Cartesian product over the block components. That is, we can say that $P$ (and hence $\phi_\tau$) is partly smooth at $x^*$ with active manifold $\mathcal{M}$ if and only if each $P_{[q]}$ is partly smooth at $x^*_{[q]}$ with active manifold $\mathcal{M}_q$, where $\mathcal{M} = \otimes_{q\in Q}\mathcal{M}_q$. (For a proof of this claim, see [15, Proposition 4.5].)

Following [15, Definition 5.6], we say that $x^*$ is a *strong critical point* of $\phi_\tau$ relative to the active manifold $\mathcal{M}$, where $\phi_\tau$ is partly smooth with respect to $\mathcal{M}$ at a point $x^*$, if

(i) $x^*$ is a strong local minimizer of $\phi_\tau |_{\mathcal{M}}$ with some modulus $c > 0$, and
(ii) the nondegeneracy condition (2.4) holds.

**2.3. Second-Order Conditions.** We now discuss second-order conditions that ensure at least quadratic increase in the objective $\phi_\tau$ as we move away from a solution $x^*$. These conditions motivate the Newton-based acceleration techniques of Subsection 3.4.

The first result relates the strong local minimizer property for $\phi_\tau|_{\mathcal{M}}$ at $x^*$ to the second-order sufficient conditions for an explicit representation of this function along this manifold.

THEOREM 2.4. *Suppose that $\phi_\tau$ is partly smooth at $x^* \in \mathbb{R}^n$ relative to an active manifold $\mathcal{M} \subset \mathbb{R}^n$. Suppose that $\mathcal{M}$ is characterized by $C^2$ mappings $F : \mathbb{R}^n \to \mathbb{R}^k$ and $G : \mathbb{R}^{n-k} \to \mathbb{R}^n$ and a point $y^* \in \mathbb{R}^{n-k}$, such that $F(x) = 0$ for all $x \in \mathcal{M}$ near $x^*$, $\nabla F(x^*)$ is orthonormal, $G(y) \in \mathcal{M}$ for all $y$ near $y^*$, and $G(y) = x^* + Y(y - y^*) + O(|y - y^*|^2)$ for some matrix $Y$ such that $\begin{bmatrix}\nabla F(x^*) & Y\end{bmatrix}$ is orthogonal. Then $\phi_\tau|_{\mathcal{M}}$ has a strong local minimizer at $x^*$ with modulus $c > 0$ if and only if the function defined by*

$$\psi_\tau(y) = \phi_\tau(G(y)) \tag{2.5}$$

*is $C^2$ in a neighborhood of $y^*$ with $\nabla\psi_\tau(y^*) = 0$ and $\nabla^2\psi_\tau(y^*)$ positive definite, with minimum eigenvalue at least $2c$.*

*Proof.* By Definition 2.3(i), we can define a neighborhood $U$ of $x^*$ and a $C^2$ mapping $\rho : U \to \bar{\mathbb{R}}$ that agrees with $\phi_\tau$ on $\mathcal{M} \cap U$. Then from (2.5), we have $\psi_\tau = \rho \circ G$, which is a composition of two $C^2$ functions and is therefore itself $C^2$ in a neighborhood of $y^*$. Note too that

$$|G(y) - x^*| = |Y(y - y^*) + O(|y - y^*|^2)| = |y - y^*| + O(|y - y^*|^2).$$

Consider first the forward implication. Since $x^*$ is a local minimizer of $\phi_\tau|_{\mathcal{M}}$, we have that $\psi_\tau(y) - \psi_\tau(y^*) = \phi_\tau(G(y)) - \phi_\tau(x^*) \geq 0$ for all $y \in \mathbb{R}^{n-k}$ sufficiently close to $y^*$, from which it follows that $\nabla\psi_\tau(y^*) = 0$. Thus we have

$$\psi_\tau(y) - \psi_\tau(y^*) = \frac{1}{2}(y - y^*)^T\nabla^2\psi_\tau(y^*)(y - y^*) + o(|y - y^*|^2).$$

Since if $\phi_\tau|_{\mathcal{M}}$ has a strong local minimizer at $x^*$ with modulus $c$, we have

$$\begin{aligned}\psi_\tau(y) - \psi_\tau(y^*) &= \phi_\tau(G(y)) - \phi_\tau(x^*)\\ &\geq c|G(y) - x^*|^2 + o(|G(y) - x^*|^2) = c|y - y^*|^2 + o(|y - y^*|^2).\end{aligned}$$

The forward implication follows by combining these last two estimates.

For the reverse implication, we have similarly that

$$\begin{aligned}\phi_\tau(G(y)) - \phi_\tau(x^*) &= \psi_\tau(y) - \psi_\tau(y^*)\\ &\geq c|y - y^*|^2 + o(|y - y^*|^2) = c|G(y) - x^*|^2 + o(|G(y) - x^*|^2),\end{aligned}$$

giving the result. □

We now show that strong critical points are in fact strong local minimizers for $\phi_\tau$. The argument is similar to that of Wright [32, Theorem 3.2 (i)] in a different setting, but is somewhat more general. It allows nonconvexity in the smooth part of $\phi_\tau$ (namely, $f$). We note that the result is *not* true for general prox-regular functions, as the example of [15, Section 7] attests.

THEOREM 2.5. *Suppose that $\phi_\tau$ is partly smooth at $x^*$ relative to $\mathcal{M}$, that $f$ is Lipschitz continuously differentiable at $x^*$, and that $x^*$ is a strong critical point. Then $x^*$ is in fact a strong local minimizer.*

*Proof.* Let the mappings $F$ and $G$, the matrix $Y$, and the point $y^* \in \mathbb{R}^{n-k}$ be defined as in Theorem 2.4, and recall that $\nabla G(y^*)^T = Y$. From Lemma 2.2, for all $x$ near $x^*$, we can find unique $y \in \mathbb{R}^{n-k}$ and $v \in \mathbb{R}^k$ with $|(y - y^*, v)| = O(|x - x^*|)$ such that $x = G(y) + \nabla F(x^*)v$. We thus have

$$\phi_\tau(x) - \phi_\tau(x^*) = [\phi_\tau(G(y) + \nabla F(x^*)v) - \phi_\tau(G(y))] + [\phi_\tau(G(y)) - \phi_\tau(x^*)] \,. \tag{2.6}$$

For the last bracketed term, we have from the strong local minimizer condition that there is $c > 0$ such that

$$\phi_\tau(G(y)) - \phi_\tau(x^*) \geq c|G(y) - x^*|^2 \tag{2.7}$$

for all $y$ near $y^*$. For the first bracketed term, note first that from Lemma A.1 there is $\epsilon > 0$ such that $\sup_{g \in \partial\phi_\tau(x^*)} g^T d \geq \epsilon|d|$ for all $d \in N_{\mathcal{M}}(x^*) = \operatorname{range} \nabla F(x^*)$. Second, from Definition 2.3 (iv), we have by choosing the neighborhood of $x^*$ sufficiently small that for all $\hat{g} \in \partial\phi_\tau(G(y))$, there is $g \in \partial\phi_\tau(x^*)$ such that $|\hat{g} - g| \leq \epsilon/2$. Third, recall that by Lipschitz continuity of $\nabla f$ at $x^*$, there is $\sigma > 0$ such that (2.1) holds in a neighborhood of $x^*$. Using all these facts, we have for all $y$ and $v$ such that $y - y^*$ and $v$ are sufficiently small that

$$\begin{aligned}\phi_\tau(G(y)+&\nabla F(x^*)v) - \phi_\tau(G(y)) \\ &\geq \sup_{\hat{g} \in \partial\phi_\tau(G(y))} \hat{g}^T \nabla F(x^*)v - \sigma|\nabla F(x^*)v|^2 \\ &\geq \sup_{g \in \partial\phi_\tau(x^*)} g^T \nabla F(x^*)v - (\epsilon/2)|\nabla F(x^*)v| - \sigma|\nabla F(x^*)v|^2 \\ &\geq \epsilon|\nabla F(x^*)v| - (\epsilon/2)|\nabla F(x^*)v| - \sigma|\nabla F(x^*)v|^2.\end{aligned}$$

By substituting this inequality and (2.7) into (2.6), we have that

$$\phi_\tau(x) - \phi_\tau(x^*) \geq (\epsilon/2)|\nabla F(x^*)v| - \sigma|\nabla F(x^*)v|^2 + c|G(y) - x^*|^2.$$

By choosing the neighborhood of $x^*$ small enough, we can ensure that $|\nabla F(x^*)v| < \epsilon/(4\sigma)$ and therefore

$$\begin{aligned}\phi_\tau(x) - \phi_\tau(x^*) &\geq \sigma|\nabla F(x^*)v|^2 + c|G(y) - x^*|^2 \\ &\geq \min(\sigma, c) \left[|\nabla F(x^*)v|^2 + |G(y) - x^*|^2\right] \\ &\geq \frac{1}{2}\min(\sigma, c)\,[|\nabla F(x^*)v| + |G(y) - x^*|]^2 \\ &\geq \frac{1}{2}\min(\sigma, c)|x - x^*|^2,\end{aligned}$$

as required. □

**Algorithm 1** Accelerated Proximal Block Coordinate Relaxation

```
Input: $\mu_{\rm top} > \mu_{\min} > 0$, $T > 1$, $\tau > 0$, $\eta > 1$, $\beta \geq 1$, $\gamma \in (0, .5)$, **tol** $> 0$;
for $k = 0, 1, 2, \ldots$ do
  Choose $Q_k \subset Q$ such that (1.5) is satisfied for the chosen $T$;
  Choose $\mu_k \in [\mu_{\min}, \mu_{\rm top}]$;
  Solve (1.3) for $d^k$;
  while $\phi_\tau(x^k + d^k) > \phi_\tau(x^k) - |d^k|^3$ do
    Set $\mu_k \leftarrow \eta\mu_k$;
    Solve (1.3) for $d^k$;
  end while
  {Try to improve on prox-descent step}
  Find $\tilde{d}^k$ with $\phi_\tau(x^k + \tilde{d}^k) \leq \phi_\tau(x^k + d^k)$ and $\phi_\tau(x^k + \tilde{d}^k) \leq \phi_\tau(x^k) - \gamma|\tilde{d}^k|^3$;
  Set $x^{k+1} \leftarrow x^k + \tilde{d}^k$;
end for
```

**3. Accelerated Proximal Block Coordinate Relaxation.** Algorithm 1 is the basic framework we consider in this paper. It is quite general, in that the sequence of relaxation sets need only satisfy (1.5), there are few restrictions on the choice of the initial $\mu_k$ at each iteration, and the successful proximal step $d^k$ can be replaced by any other step $\tilde{d}^k$ that improves the value of $\phi_\tau$ and satisfies another modest decrease condition. Subsection 3.2 proves a global convergence result for this framework (specifically, criticality of accumulation points), while Subsection 3.3 describes identification properties for the active manifold. A reduced-Newton acceleration scheme is described in Subsection 3.4, along with local convergence results.

The "sufficient decrease" criteria are key to the algorithm. From current iterate $x$, the prox-descent step $d$ is required to satisfy

$$\phi_\tau(x) - \phi_\tau(x + d) \geq |d|^3, \tag{3.1}$$

while the step $\tilde{d}$ actually taken must satisfy both

$$\phi_\tau(x) - \phi_\tau(x + \tilde{d}) \geq \gamma|\tilde{d}|^3, \tag{3.2}$$

for a given parameter $\gamma \in (0, .5)$, and

$$\phi_\tau(x + \tilde{d}) \leq \phi_\tau(x + d). \tag{3.3}$$

(The choice $\tilde{d} = d$ is obviously one option that satisfies both (3.2) and (3.3).)

Note that if the solution of the subproblem (1.3) is $d = 0$ for the first value tried (indicating that the optimality condition (2.3) is satisfied in the components of $Q_k$), the acceptance condition is satisfied and $\mu_k$ is not increased. This is a reasonable outcome, as the algorithm detects correctly that, to first order, no further progress can be made in this relaxation set.

We note that value of $\mu_k$ at iteration $k$, for relaxation set $Q_k$, may have little relevance for iteration $k + 1$, where the relaxation set $Q_{k+1}$ may be quite different. Algorithm 1 does not assume any "memory" in the choice of damping values. In our implementations, however, we have found that an effective initial choice of $\mu_{k+1}$ is some multiple (for example, .8) of the final value of $\mu_k$ from the previous iteration.

**3.1. Technical Results.** We start with two technical results about the dependence of the (unique) solution of subproblem (1.3) on $\mu$ and $x$. In the first result, we assume that $x$ and $Q'$ in (1.3) are given, and explore the dependence of the solution on $\mu$ alone.

LEMMA 3.1. *Suppose $P$ is finite at a point $x \in \mathbf{R}^n$ and denote the (unique) solution of (1.3) for any $\mu > 0$ and fixed relaxation set $Q'$ by $d(\mu)$. We have that $|d(\mu)|$ is a decreasing function of $\mu$ and that $d(\mu) \to 0$ as $\mu \uparrow \infty$.*

*Proof.* We give an elementary proof and assume for simplicity that $Q' = Q$. Supposing that $\tilde{\mu} > \mu > 0$, we have directly from (1.3) that

$$\nabla f(x)^T d(\tilde{\mu}) + \frac{\tilde{\mu}}{2}|d(\tilde{\mu})|^2 + \tau P(x + d(\tilde{\mu})) \le \nabla f(x)^T d(\mu) + \frac{\tilde{\mu}}{2}|d(\mu)|^2 + \tau P(x + d(\mu)),$$
$$\nabla f(x)^T d(\mu) + \frac{\mu}{2}|d(\mu)|^2 + \tau P(x + d(\mu)) \le \nabla f(x)^T d(\tilde{\mu}) + \frac{\mu}{2}|d(\tilde{\mu})|^2 + \tau P(x + d(\tilde{\mu})).$$

By adding these inequalities and rearranging, we obtain

$$\frac{1}{2}(\tilde{\mu} - \mu)|d(\tilde{\mu})|^2 \le \frac{1}{2}(\tilde{\mu} - \mu)|d(\mu)|^2,$$

which by $\tilde{\mu} - \mu > 0$ implies the first result.

Suppose now that $d(\mu) \not\to 0$ as $\mu \uparrow \infty$. By a compactness argument we can find an increasing, unbounded sequence $\mu_j$, $j = 1, 2, \dots$ and a limit point $\hat{d} \neq 0$ such that $d(\mu_j) \to \hat{d}$. Since 0 is a feasible point for (1.3), for each $j$, we have that

$$\nabla f(x)^T d(\mu_j) + \frac{\mu_j}{2}|d(\mu_j)|^2 + \tau P(x + d(\mu_j)) \le \tau P(x).$$

Rearranging, we obtain

$$P(x + d(\mu_j)) \le P(x) - \frac{1}{\tau}\left[\nabla f(x)^T d(\mu_j) + \frac{\mu_j}{2}|d(\mu_j)|^2\right].$$

By taking limits, we have $\lim_{j\to\infty} P(x + d(\mu_j)) = -\infty$. However since $d(\mu_j) \to \hat{d}$ and $P(x + \hat{d}) > -\infty$ (since $P$ is proper), and since closedness of $P$ implies lower semicontinuity at $x + \hat{d}$, we have a contradiction. □

In the next result, we assume only that $\mu \ge \mu_{\min} > 0$, and investigate the dependence of the solution of (1.3) on $x$, for $x$ in a neighborhood of a critical point.

LEMMA 3.2. *Suppose that $x^*$ is a critical point for $\phi_\tau$ and that $f$ is locally Lipschitz at $x^*$. Then there is a constant $\bar{L}$ such that provided that $\mu \ge \mu_{\min}$, we have $|d| \le \bar{L}|x - x^*|$ for all $x$ sufficiently close to $x^*$.*

*Proof.* We assume for simplicity that $Q' = Q$. (Similar logic holds for any $Q' \subset Q$.) Suppose for contradiction that there are sequences $\{x^l\}$ and $\{\mu_l\}$ with $x^l \to x^*$ and $\mu_l \ge \mu_{\min}$ for all $l = 1, 2, \dots$, such that

$$\lim_{l\to\infty} \frac{|x^l - x^*|}{|d^l|} = 0. \tag{3.4}$$

By criticality of $x^*$, we have

$$-\frac{1}{\tau}\nabla f(x^*) \in \partial P(x^*),$$

so by convexity of $P$ we have

$$P(x^l + d^l) \ge P(x^*) - \frac{1}{\tau}\nabla f(x^*)^T (x^l + d^l - x^*). \tag{3.5}$$

On the other hand, $d^l$ is optimal in (1.3) (better than the alternative step $x^* - x^l$), so we have

$$\nabla f(x^l)^T d^l + \frac{\mu_l}{2}|d^l|^2 + \tau P(x^l + d^l) \le \nabla f(x^l)^T (x^* - x^l) + \frac{\mu_l}{2}|x^l - x^*|^2 + \tau P(x^*). \tag{3.6}$$

By substituting from (3.5) for $P(x^l + d^l)$ into (3.6), we have after some rearrangement that

$$\begin{aligned} \frac{\mu_l}{2}|d^l|^2 &\le [\nabla f(x^*) - \nabla f(x^l)]^T (x^l + d^l - x^*) + \frac{\mu_l}{2}|x^l - x^*|^2 \\ &\le L|x^l - x^*|(|x^l - x^*| + |d^l|) + \frac{\mu_l}{2}|x^l - x^*|^2, \end{aligned}$$

where $L$ is the local Lipschitz constant for $\nabla f$ at $x^*$. Dividing both sides by $\mu_l|d^l|^2$, we have

$$\frac{1}{2} \le \frac{L}{\mu_l}\frac{|x^l - x^*|}{|d^l|}\left(\frac{|x^l - x^*|}{|d^l|} + 1\right) + \frac{1}{2}\frac{|x^l - x^*|^2}{|d^l|^2}.$$

By taking limits as $l \to \infty$, we have from $\mu_l \ge \mu_{\min}$ and (3.4) that the right-hand side approaches zero, giving a contradiction. □

**3.2. Global Convergence.** We now that accumulation points of Algorithm 1 are critical. Our first result verifies that the algorithm is well defined (in the sense that each inner loop eventually terminates) in the neighborhood of any point at which $\nabla f$ is locally Lipschitz.

LEMMA 3.3. *Suppose that $\nabla f$ is locally Lipschitz in a neighborhood of a point $\bar{x}$. Then there are positive constants $\rho$ and $\hat{\mu}$ such that the solution $d$ of (1.3) evaluated at any $x$ with $|x - \bar{x}| \le \rho$ and any $\mu$ with $\mu \ge \hat{\mu}$ satisfies the sufficient decrease condition (3.1).*

*Proof.* Choose $\rho$ small enough that for some $L > 0$ we have

$$|x - \bar{x}| \le \rho \text{ and } |d| \le \rho \;\Rightarrow\; |\nabla f(x + d) - \nabla f(x)| \le L|d|. \tag{3.7}$$

Now suppose for contradiction that for some $x$ with $|x - \bar{x}| \le \rho$, there is a sequence $\mu_j \uparrow \infty$ such that for $d^j$ that solves (1.3) with $\mu = \mu_j$ and $Q' = Q_j \subset Q$, we have

$$\phi_\tau(x) - \phi_\tau(x + d^j) < |d^j|^3. \tag{3.8}$$

By taking a subsequence if necessary, we can assume that $Q_j \equiv \bar{Q} \subset Q$. Since $d^j \to 0$ from Lemma 3.1, we can assume further that $|d^j| \le \rho$ for all $j$. By optimality conditions (1.4) for (1.3), we have

$$-\frac{1}{\tau}\left[\nabla_{[q]} f(x) + \mu_j d^j_{[q]}\right] \in \partial P_q(x_{[q]} + d^j_{[q]}), \qquad \text{for all } q \in \bar{Q},$$

and therefore by convexity of $P_q$ we have

$$P_q(x_{[q]}) - P_q(x_{[q]} + d^j_{[q]}) \ge \frac{1}{\tau}(d^j_{[q]})^T \left[\nabla_{[q]} f(x) + \mu_j d^j_{[q]}\right], \qquad \text{for all } q \in \bar{Q}. \tag{3.9}$$

Thus we have for all $j$ that

$$\begin{aligned}
&\phi_\tau(x+d^j)-\phi_\tau(x)\\
&\quad= (d^j)^T\nabla f(x) + (d^j)^T\left[\nabla f(x+t_jd^j)-\nabla f(x)\right]\\
&\qquad + \tau(P(x+d^j)-P(x)) && \text{for some } t_j\in(0,1)\\
&\quad\le (d^j)^T\nabla f(x) + L|d^j|^2 - \sum_{q\in\bar{Q}}(d^j_{[q]})^T\left[\nabla_{[q]}f(x)+\mu_jd^j_{[q]}\right] && \text{from (3.7) and (3.9)}\\
&\quad= -\mu_j|d^j|^2+L|d^j|^2.
\end{aligned}$$

Hence for all $j$ sufficiently large, using again that $d^j\to 0$ and $\mu_j\uparrow\infty$, we have

$$\phi_\tau(x)-\phi_\tau(x+d^j)\ge(\mu_j-L)|d^j|^2>|d^j|^3,$$

contradicting (3.8) and proving the result. □

It follows immediately that when $\nabla f$ is locally Lipschitz in a neighborhood of an iterate $x^k$ of Algorithm 1, the inner loop eventually terminates with a value of $\mu_k$ that satisfies the sufficient decrease condition.

We next prove the result about criticality of accumulation points.

Theorem 3.4. *Suppose that Algorithm 1 generates an infinite sequence and that $x^*$ is an accumulation point of this sequence at which $\nabla f$ is locally Lipschitz continuous. Then $x^*$ is a critical point.*

*Proof.* Denote by $\mathcal{K}$ the subsequence such that $\lim_{k\in\mathcal{K}}x^k=x^*$. Since $\{\phi_\tau(x^k)\}$ is decreasing and bounded below, the sequence converges and in fact $\phi_\tau(x^k)\downarrow\phi_\tau(x^*)$. From (3.2), we have

$$\phi_\tau(x^k)-\phi_\tau(x^{k+1})\ge\gamma|\tilde{d}^k|^3,\qquad \text{for } k=0,1,2,\dots,$$

so that $\tilde{d}^k\to 0$ for the full sequence. Since $\lim_{k\in\mathcal{K}}|x^k-x^*|=0$, we have for all $j=1,\dots,T$ (where $T$ is the cycle length from (1.5)) that

$$0\le\lim_{k\in\mathcal{K}}|x^{k-j}-x^*|\le\lim_{k\in\mathcal{K}}|x^k-x^*|+\sum_{l=1}^{j}|\tilde{d}^{k-l}|=0,$$

and hence $\lim_{k\in\mathcal{K}}x^{k-j}=x^*$ for all $j=0,1,\dots,T$.

By local Lipschitz continuity, we can define positive constants $L>0$ and $\rho>0$ such that $|\nabla f(x+d)-\nabla f(x)|\le L|d|$ for all $x$, $d$ with $|x-x^*|\le\rho$ and $|d|\le\rho$. As in Lemma 3.3, we can identify $\hat{\mu}>0$ such that the sufficient decrease condition (3.1) is satisfied along with $|d|\le\rho$, when $|x-x^*|\le\rho$ and $d$ is obtained from (1.3) for $\mu\ge\hat{\mu}$, for any $Q'\subset Q$. After eliminating from the subsequence $\mathcal{K}$ all indices $k$ such that $k\le T$, and all indices $k$ such that $|x^{k-j}-x^*|>\rho$ for some $j=0,1,\dots,T$, we still have an infinite subsequence by the argument above. The mechanism of the algorithm ensures that $\mu_{k-j}\le\max(\eta\hat{\mu},\mu_{\text{top}})$ for all $k\in\mathcal{K}$ and all $j=0,1,\dots,T$. In particular, we have $\mu_{k-j}d^{k-j}\to 0$ where the limit is taken over all elements $k\in\mathcal{K}$ and $j=0,1,\dots,T$.

We now choose any $q\in Q$ and note from (1.5) that $q\in Q_{k-j}$ for some $j=0,1,\dots,T-1$ and for every $k\in\mathcal{K}$. Let $j_k=k-j$ for some such $j$. We have from subproblem optimality (1.4) that

$$0\in\nabla_{[q]}f(x^{j_k})+\mu_{j_k}d^{j_k}_{[q]}+\tau\partial P_q(x^{j_k}_{[q]}+d^{j_k}_{[q]}),\qquad\text{for all } k\in\mathcal{K}.$$

By taking limits of this expression as $k \in \mathcal{K}$ approaches $\infty$, noting from the previous paragraph that $\lim_{k\in\mathcal{K}} \mu_{j_k} d^{j_k} = 0$, and using outer semicontinuity of $\partial P_q$, we have

$$0 \in \nabla_{[q]} f(x^*) + \tau \partial P_q(x^*_{[q]}).$$

Since $q$ is any element of $Q$, we conclude that the criticality condition (2.3) holds at $x^*$, completing the proof. □

**3.3. Identification.** In this section, we prove results about identification of the active manifold by iterates of the algorithm. We assume throughout that the active manifold at the limit point $x^*$ is partly smooth and that the nondegeneracy condition (2.4) is satisfied. A characterization of identification from Hare and Lewis [9] is key to the analysis. This result has recently been applied to algorithms that take prox-descent steps in Hare [10] (for smooth constrained optimization) and Lewis and Wright [16] and Hare [11] (for regularized optimization). The novelty in the analysis is largely in the handling of the block-coordinate steps rather than full prox-descent steps involving all the coordinates at once.

The first result shows that the prox-descent step from (1.3) yields identification from any point $x$ sufficiently close to a nondegenerate critical point $x^*$, provided that the relaxation set $Q'$ encompasses all components $q$ for which $x_{[q]}$ does not lie on the correct manifold.

THEOREM 3.5. *Let the nondegenerate criticality condition (2.4) be satisfied at $x^*$ and suppose that $\nabla f$ is locally Lipschitz there. Suppose that each $P_q$ is partly smooth at $x^*_{[q]}$, with active manifold $\mathcal{M}_q$, for all $q \in Q$. Then there is a $\bar{\delta} > 0$ such that for any $x$ with $|x - x^*| < \bar{\delta}$, one step of Algorithm 1 starting from $x$ with relaxation set $Q(x)$ chosen such that $q \in Q(x)$ whenever $x_{[q]} \notin \mathcal{M}_q$ results in a step $d$ for which $x + d \in \mathcal{M}$.*

*Proof.* Suppose for contradiction that there are sequences $x^l \to x^*$ and $Q(x^l) \subset Q$ with $q \in Q(x^l)$ whenever $x^l_{[q]} \notin \mathcal{M}_q$, yet one iteration of Algorithm 1 yields a step $d^l$ such that $x^l + d^l \notin \mathcal{M}$. We can assume without loss of generality that $Q(x^l) \equiv Q'$ for some $Q' \subset Q$.

By the local Lipschitz property of $\nabla f$, we can use Lemma 3.3 and Theorem 3.4 to deduce existence of $\hat{\mu}$ such that the sufficient decrease condition (3.1) is satisfied for all $\mu \geq \hat{\mu}$ and all $l$ sufficiently large. Hence, the mechanism of the algorithm will choose $\mu_l$ with $\mu_{\min} \leq \mu_l \leq \max(\eta\hat{\mu}, \mu_{\text{top}})$, and set $d^l$ to the solution of (1.3) for $x = x^l$ and $\mu = \mu_l$. Moreover, from Lemma 3.2, we have $|d^l| = O(|x^l - x^*|)$. It follows that $\mu_l d^l \to 0$.

We have for $q \in Q'$ that

$$\begin{aligned}
&\left[\nabla f(x^l + d^l) + \tau \partial P(x^l + d^l)\right]_{[q]} \\
=&\nabla_{[q]} f(x^l + d^l) + \tau \partial P_q(x^l_{[q]} + d^l_{[q]}) \\
=& \left[\nabla_{[q]} f(x^l) + \mu_l d^l_{[q]} + \tau \partial P_q(x^l_{[q]} + d^l_{[q]})\right] + O(|d^l|) - \mu_l d^l_{[q]},
\end{aligned}$$

and so from subproblem optimality (1.4) we have

$$\begin{aligned}
&\text{dist}\left(0, \left[\nabla f(x^l + d^l) + \tau \partial P(x^l + d^l)\right]_{[q]}\right) \\
&\qquad\qquad \leq O(|d^l|) + \mu_l |d^l| \to 0, \qquad \text{for all } q \in Q'. \qquad (3.10)
\end{aligned}$$

We now consider $q \notin Q'$, for which $x^l_{[q]} \in \mathcal{M}_q$ and $d^l_{[q]} = 0$. Note first that we have

$$\nabla_{[q]} f(x^l + d^l) = \nabla_{[q]} f(x^*) + O(|x^l - x^*|) + O(|d^l|), \tag{3.11}$$

by Lipschitz continuity of $\nabla f$. Moreover, using partial smoothness of $P_q$ at $x^*$ and the property $-\nabla_{[q]} f(x^*) \in \tau \partial P_q(x^*)$, we have

$$\begin{aligned}
&\operatorname{dist}(-\nabla_{[q]} f(x^*), \tau \partial P_q(x^l_{[q]} + d^l_{[q]}) \\
&= \operatorname{dist}(-\nabla_{[q]} f(x^*), \tau \partial P_q(x^l_{[q]}) && \text{since } d^l_{[q]} = 0 \text{ for } q \notin Q' \\
&= \operatorname{dist}(-\nabla_{[q]} f(x^*), \tau \partial P_q(x^*_{[q]}) + o(1) && \text{by Definition 2.3(iv)} \\
&= o(1). && (3.12)
\end{aligned}$$

It follows by combining (3.11) and (3.12) that

$$\operatorname{dist}\left(0, \left[\nabla f(x^l + d^l) + \tau \partial P(x^l + d^l)\right]_{[q]}\right) \to 0, \qquad \text{for all } q \notin Q'. \tag{3.13}$$

By combining (3.10) and (3.13), we obtain

$$\operatorname{dist}\left(0, \nabla f(x^l + d^l) + \tau \partial P(x^l + d^l)\right) \to 0,$$

implying from [9, Theorem 5.3] that $x^l + d^l \in \mathcal{M}$ for all $l$ sufficiently large, giving the desired contradiction. □

In the next result we consider the behavior of Algorithm 1 in the neighborhood of a nondegenerate critical point, when the choice of relaxation sets satisfies the generalized Gauss-Seidel condition (1.5). Obviously, we cannot hope to identify the correct manifold until all the components have had their turn for inclusion in $Q_k$, which because of (1.5), happens at least once in each cycle of $T$ iterations. We also need to assume that the steps $\tilde{d}^k$ actually taken by the algorithm do not move away from the manifold identified by the prox-descent steps $d^k$.

Theorem 3.6. *Let the nondegenerate criticality condition (2.4) be satisfied at $x^*$ and suppose that $\nabla f$ is locally Lipschitz there. Suppose that each $P_q$ is partly smooth at $x^*_{[q]}$, with active manifold $\mathcal{M}_q$, for all $q \in Q$. Then there is a $\bar{\delta} > 0$ with the following property. If Algorithm 1 is started from any initial point $x^0$ such that $|x^0 - x^*| < \bar{\delta}$ and $\phi_\tau(x^0) - \phi_\tau(x^*) \le \bar{\delta}$, with $x^k + \tilde{d}^k$ lying on the same manifold as $x^k + d^k$ for all $k = 0, 1, 2, \ldots, T-1$, then either $\phi_\tau(x^k) < \phi_\tau(x^*)$ for some $k = 0, 1, 2, \cdots$ (in which case Algorithm 1 cannot have an accumulation point at $x^*$) or else the $T$th iterate $x^T$ lies on the manifold $\mathcal{M} := \otimes_{q \in Q} \mathcal{M}_q$.*

*Proof.* Suppose for contradiction that there is no valid choice of $\bar{\delta}$. We can then define a sequence $\{x^{l,0}\}_{l=1,2,\ldots}$ with $|x^{l,0} - x^*| \le l^{-1}$ and $\phi_\tau(x^{l,0}) - \phi_\tau(x^*) \le l^{-1}$ such that $\phi_\tau(x^{l,k}) \ge \phi_\tau(x^*)$ for all iterates $\{x^{l,k}\}_{k=0,1,2,\ldots}$ of Algorithm 1 starting from $x^{l,0}$, and after $T$ steps of the algorithm, the $T$th iterate satisfies $x^{l,T} \notin \mathcal{M}$. Since $\phi_\tau(x^*) + l^{-1} \ge \phi_\tau(x^{l,0}) \ge \phi_\tau(x^{l,k}) \ge \phi_\tau(x^*)$, we have from (3.1) and (3.2) that $|d^{l,k}|$ and $|\tilde{d}^{l,k}|$ are both bounded by a multiple of $l^{-1/3}$ for all $k = 0, 1, \ldots, T-1$.

Choosing any $q \in Q$, we denote by $k_l$ the largest iteration index in $0, 1, \ldots, T-1$ for which $q \in Q_{k_l}$. (Note that (1.5) guarantees existence of $k_l$.) Similarly to the proof of Theorem 3.5, we use Lipschitz continuity of $\nabla f$ at $x^*$ along with the reasoning in the proofs of Lemma 3.3 and Theorem 3.4 to deduce that the damping parameter

$\mu_{l,k_l}$ used at each of these steps is uniformly bounded over $l$. We now have

$$
\begin{aligned}
&\left[\nabla f(x^{l,T}) + \tau \partial P(x^{l,T})\right]_{[q]} \\
&= \nabla_{[q]} f\left(x^{l,k_l} + \sum_{k=k_l}^{T-1} \tilde{d}^{l,k}\right) + \tau \partial P_q \left(x_{[q]}^{l,k_l} + \sum_{k=k_l}^{T-1} \tilde{d}_{[q]}^{l,k}\right) \\
&= \nabla_{[q]} f\left(x^{l,k_l}\right) + \sum_{k=k_l}^{T-1} O(|\tilde{d}^{l,k}|) + \tau \partial P_q \left(x_{[q]}^{l,k_l} + d_{[q]}^{l,k_l}\right) + o(1) \\
&= \left[\nabla_{[q]} f(x^{l,k_l}) + \mu_{l,k_l} d_{[q]}^{l,k_l} + \tau \partial P_q (x_{[q]}^{l,k_l} + d_{[q]}^{l,k_l})\right] + \sum_{k=k_l}^{T-1} O(|\tilde{d}^{l,k}|) - \mu_{l,k_l} d_{[q]}^{l,k_l} + o(1),
\end{aligned}
$$

where the second equality follows from Definition 2.3(iv) (continuity of $\partial P_q$ along the manifold identified by $x_{[q]}^{l,k_l} + d_{[q]}^{l,k_l}$) and the fact that each subsequent iterate $x_{[q]}^{l,k}$, $k = k_l + 1, \dots, T$ lies on the manifold identified by $x^{l,k_l} + d^{l,k_l}$. Using subproblem optimality (1.4) and our estimates of $|d^{l,k_l}|$, $|\tilde{d}^{l,k_l}|$, and $\mu_{l,k_l}$, we have that

$$
\operatorname{dist}\left(0, \left[\nabla f(x^{l,T}) + \tau \partial P(x^{l,T})\right]_{[q]}\right) \to 0.
$$

Since this estimate can be derived for all $q \in Q$, we have that

$$
\operatorname{dist}\left(0, \nabla f(x^{l,T}) + \tau \partial P(x^{l,T})\right) \to 0,
$$

implying from [9, Theorem 5.3] that $x^{l,T} \in \mathcal{M}$ for all $l$ sufficiently large, giving the desired contradiction. □

We note that when each $P_q$ is everywhere finite valued, the function $P$ is Lipschitz on $\mathbf{R}^n$, and in this case we can dispense with the assumption that $\phi_\tau(x^0) - \phi_\tau(x^*) \le \bar{\delta}$ in Theorem 3.6.

**3.4. Reduced-Newton Acceleration.** We now describe a variant of Algorithm 1 in which the step $\tilde{d}^k$ is obtained from a reduced-Newton step on the current estimate of the optimal manifold $\mathcal{M}$. At iteration $k$, the procedure is as follows: Compute $d^k$ satisfying (3.1) and find a manifold $\mathcal{M}^k$ containing $x^k + d^k$ that is partly smooth at $x^k + d^k$. (If no such manifold can be conveniently identified, set $\tilde{d}^k = d^k$ and skip the acceleration step.) Now identify a mapping $G^k$ that parametrizes the manifold $\mathcal{M}^k$, in the sense of Lemma 2.1, and a point $y^k$ such that $G^k(y^k) = x^k + d^k$ and $G^k(y) \in \mathcal{M}^k$ for all $y$ in a neighborhood of $y^k$. Next, define $\psi_\tau^k(y) := \phi_\tau(G^k(y))$ (as in (2.5)) and compute a Newton step $w^k$ for $\psi_\tau^k$ from $y^k$. Finally, define the step to be taken as $\tilde{d}^k := G(y^k + w^k) - x^k$, if this step satisfies the acceptance conditions (3.2) and (3.3), and if $x^k + \tilde{d}^k$ lies on the same manifold $\mathcal{M}^k$ as $x^k + d^k$. Otherwise, set $\tilde{d}^k = d^k$.

We now prove a superlinear convergence for this procedure, building on the identification result of Theorem 3.6, the properties of manifolds and their parametrizations introduced in Subsection 2.2, and the standard properties of Newton's method.

THEOREM 3.7. *Suppose that $x^*$ is a strong critical point of $\phi_\tau$, where $f$ is $C^2$ at $x^*$ and each $P_q$ is partly smooth at $x_{[q]}^*$ with respect to an active manifold $\mathcal{M}_q$. Let $\{x^k\}$ be the sequence generated by the algorithm, with the acceleration procedure described above, and assume that $x^k \to x^*$. Then $x^k \in \mathcal{M} := \otimes_{q \in Q} \mathcal{M}_q$ for all $k$ sufficiently large, and the convergence of $\{x^k\}$ to $x^*$ is Q-quadratic.*

*Proof.* We can use Theorem 3.6 to deduce that $x^k + d^k \in \mathcal{M}$ for all $k$ sufficiently large, so the acceleration procedure ensures that in fact $x^{k+1} = x^k + \tilde{d}^k \in \mathcal{M}$ for all $k$ sufficiently large. Hence we can assume that $\mathcal{M}^k \equiv \mathcal{M}$ and $G^k \equiv G$ in the description of the acceleration procedure.

In the remainder of the proof, we identify a radius $\delta > 0$ such that the step $\tilde{d}^k$ obtained by the acceleration procedure is accepted for all $k$ large enough that $|x^k - x^*| < \delta$. For clarity, in the analysis below, we drop the superscript "$k$" and replace "$k + 1$" by "+".

Since the assumptions of Theorems 2.4 and 2.5 are satisfied, we have that $x^*$ is a strong local minimizer for some modulus of convexity $c > 0$. By constructing the parametrization $G$ of $\mathcal{M}$ as described above, and defining $\psi_\tau$ as in (2.5), we have that $\nabla^2\psi_\tau(\bar{y})$ is positive definite with smallest eigenvalue at least $2c$ at the solution $\bar{y}$ for which $G(\bar{y}) = x^*$, while $\nabla\psi_\tau(\bar{y}) = 0$.

By applying Lemma 3.2, and noting that $\mu \geq \mu_{\min}$, we have for $|x - x^*| \leq \delta$ (by decreasing $\delta$ if necessary) that the solution $d$ of (1.3) satisfies

$$|d| \leq c_2|x - x^*| \leq c_2\delta, \tag{3.14}$$

for some constant $c_2 > 0$. Hence, we have

$$|x + d - x^*| \leq (1 + c_2)\delta. \tag{3.15}$$

Consider now the reduced Newton step from $x + d$. From Lemma 2.1, we have

$$x + d - x^* = G(y) - x^* = Y(y - \bar{y}) + O(|y - \bar{y}|^2), \tag{3.16}$$

for an orthonormal matrix $Y$, so that by reducing $\delta$ further as needed, we have

$$\frac{1}{2}|y - \bar{y}| \leq |x + d - x^*| \leq 2|y - \bar{y}|. \tag{3.17}$$

Using this estimate together with (3.15), we see that $|y - \bar{y}| = O(\delta)$, so by reducing $\delta$ again if necessary and defining $\psi_\tau$ as in (2.5), the Newton step $w$ from $y$ is well defined, and we have

$$w = -[\nabla^2\psi_\tau(y)]^{-1}\nabla\psi_\tau(y). \tag{3.18}$$

Moreover, standard analysis of Newton's method yields that

$$|y + w - \bar{y}| \leq c_3|y - \bar{y}|^2 \tag{3.19}$$

for some $c_3 > 0$. We have in particular that

$$\frac{1}{2}|y - \bar{y}| \leq |w| \leq 2|y - \bar{y}|. \tag{3.20}$$

Defining $\tilde{d} = G(y + w) - x$, we have from Lemma 2.1 that

$$x + \tilde{d} - x^* = G(y + w) - x^* = Y(y + w - \bar{y}) + O(|y + w - \bar{y}|^2), \tag{3.21}$$

for the same orthonormal matrix $Y$ as in (3.16), so by decreasing $\delta$ further if necessary, we have

$$\frac{1}{2}|y + w - \bar{y}| \leq |x + \tilde{d} - x^*| \leq 2|y + w - \bar{y}|. \tag{3.22}$$

By comparing (3.16) with (3.21), we obtain

$$\tilde{d} - d = Yw + O(|y - \bar{y}|^2) + O(|y + w - \bar{y}|^2), \tag{3.23}$$

so from (3.20), we have

$$\tilde{d} - d = Yw + O(|w|^2) = Yw + O(|y - \bar{y}|^2). \tag{3.24}$$

By the usual argument, we thus have

$$\frac{1}{2}|d - \tilde{d}| \le |w| \le 2|d - \tilde{d}|. \tag{3.25}$$

Denoting the next iterate by $x^+$, defined by $x^+ := x + \tilde{d}$, we have

$$\begin{aligned} |x^+ - x^*| &\le 2|y + w - \bar{y}| && \text{from (3.22)} \\ &\le 2c_3|y - \bar{y}|^2 && \text{from (3.19)} \\ &\le 8c_3|x + d - x^*|^2 && \text{from (3.17)} \\ &\le 8c_3(1 + c_2)^2|x - x^*|^2 && \text{from (3.15)} \\ &\le 8c_3(1 + c_2)^2\delta|x - x^*| && \text{since } |x - x^*| \le \delta. \end{aligned} \tag{3.26}$$

By decreasing $\delta$ again if necessary, we have that $|x^+ - x^*| \le 0.5|x - x^*|$, so all the estimates obtained above and below for $x$ (and its corresponding steps $d$ and $\tilde{d}$) continue to apply at $x^+$ and indeed at all subsequent iterates. Note too that for this choice of $\delta$, we have

$$|\tilde{d}| \le |x^+ - x^*| + |x - x^*| \le 1.5|x - x^*| \le 1.5\delta. \tag{3.27}$$

We conclude by showing that $\tilde{d}$ satisfies the acceptance conditions (3.2) and (3.3). We have from standard Newton analysis, and using (3.25), that

$$\begin{aligned} \phi_\tau(x + \tilde{d}) = \psi_\tau(y + w) &\le \psi_\tau(y) - c_4|w|^2 \\ &= \phi_\tau(x + d) - c_4|w|^2 \le \phi_\tau(x + d) - \frac{1}{4}c_4|d - \tilde{d}|^2, \end{aligned} \tag{3.28}$$

for some $c_4 > 0$, so that (3.3) holds. Since $d$ satisfies (3.1), we have

$$\begin{aligned} \phi_\tau(x) - \phi_\tau(x + \tilde{d}) &= [\phi_\tau(x) - \phi_\tau(x + d)] + [\phi_\tau(x + d) - \phi_\tau(x + \tilde{d})] \\ &\ge |d|^3 + \frac{1}{4}c_4|\tilde{d} - d|^2. \end{aligned} \tag{3.29}$$

We consider two cases. First, when $|\tilde{d} - d| \le .2|\tilde{d}|$, we have $|d| \ge |\tilde{d}| - |d - \tilde{d}| \ge .8|\tilde{d}|$, so from (3.29) it follows that

$$\phi_\tau(x) - \phi_\tau(x + \tilde{d}) \ge |d|^3 \ge (.8)^3|\tilde{d}|^3 \ge \gamma|\tilde{d}|^3,$$

(since $\gamma \in (0, .5)$), so that (3.2) holds. Second, when $|\tilde{d} - d| > .2|\tilde{d}|$, we have immediately from (3.29) that

$$\phi_\tau(x) - \phi_\tau(x + \tilde{d}) \ge \frac{1}{4}c_4|\tilde{d} - d|^2 \ge .01c_4|\tilde{d}|^2.$$

By using (3.27) and reducing $\delta$ if necessary (to ensure that $|\tilde{d}| \le .01c_4/\gamma$), we have that (3.2) holds in this case too.

We conclude that it is possible to choose $\delta > 0$ such that if $|x^k - x^*| \le \delta$ for any iterate $x^k$ generated by Algorithm 1, and using the acceleration procedure outlined above, the enhanced step $\tilde{d}^l$ is accepted at all iterates $l \ge k$. Because of (3.26), the subsequent iterates converge Q-quadratically to $x^*$, as claimed. □

**4. Applications.** We discuss here implementation of the approaches of Section 3 to several particular cases of regularization function $P$ in (1.1).

We start with the case of $\ell_1$ regularized optimization, in which $f$ is smooth and $P(x) = \|x\|_1$. In the notation of (1.2), we have $Q = \{1, 2, \ldots, n\}$, $[q] = q$ for $q = 1, 2, \ldots, n$, and $P_q(x_{[q]}) = |x_q|$. Each subproblem (1.3) then becomes

$$\min_d \sum_{q \in Q'} \nabla_q f(x) d_q + \frac{\mu}{2} \sum_{q \in Q'} d_q^2 + \tau \sum_{q \in Q'} |x_q + d_q|.$$

It is well known that the solution $d$ can be evaluated in $O(|Q'|)$ operations via the "shrinkage operator." After rearrangement and a change of variables, we can write this problem equivalently as

$$\min_z \frac{1}{2}|z - g|^2 + \frac{\tau}{\mu}\|z\|_1,$$

for $z_q := x_q + d_q$ and $g_q := x_q - (1/\mu)\nabla_q f(x)$ $(q \in Q')$, whose explicit solution is

$$z_q = \operatorname{sign}(g_q) \max(|g_q| - \tau/\mu, 0), \qquad q \in Q'.$$

For each vector $x$ and each component $q$, there are three appropriate choices of component manifold $\mathcal{M}_q$: $\alpha < 0$, $\alpha > 0$, and $\alpha = 0$ for scalar $\alpha$, depending on whether $x_q$ is negative, positive, or zero, respectively. The estimate $\mathcal{M}$ for the active manifold at iterate $x + d$ is the Cartesian product of these component manifolds. We derive an explicit representation along the lines of Lemma 2.1 as follows. Define the sets $Q_0 := \{i : x_i + d_i = 0\}$, $Q_- := \{i : x_i + d_i < 0\}$, and $Q_+ := \{i : x_i + d_i > 0\}$ and the matrix $E$ whose columns are the columns of the $n \times n$ identity that correspond to indices in $Q_0$, while $Y$ is its complement. The mappings $F$ and $G$ described in Lemma 2.1 are thus

$$F(x) = E^T x, \qquad G(y) = Yy,$$

where the components of $y$ are indexed not sequentially but rather with the indices $i$ from $Q_-$ and $Q_+$.

The optimality condition (2.3) for $x^*$ is thus

$$\nabla f(x^*) + \tau v = 0, \tag{4.1}$$

where

$$v_i \begin{cases} = -1 & \text{if } x_i^* < 0 \\ = 1 & \text{if } x_i^* > 0 \\ \in [-1, 1] & \text{if } x_i^* = 0. \end{cases} \tag{4.2}$$

The nondegeneracy condition (2.4) is the same, except that we require $v_i$ to be in the open interval $(-1, 1)$ when $x_i^* = 0$.

For this case, the function $\psi_\tau(y)$ defined in (2.5) is thus

$$\psi_\tau(y) = f(Yy) - \tau \sum_{i \in Q_-} y_i + \tau \sum_{i \in Q_+} y_i,$$

which is evidently as smooth as $f$. The Newton step for $\psi_\tau$ is easily calculated if $Y^T \nabla f$ and $Y^T(\nabla^2 f)Y$ are known. If $|Q_-| + |Q_+| \ll n$, it may be much less expensive to evaluate these quantities (or approximations to them) than to evaluate the full gradient and Hessian of $f$, as we discuss further in Section 5.

Consider next the case in which $P$ is a group-$\ell_2$ regularizer:

$$P(x) = \sum_{q \in Q} |x_{[q]}|,$$

where here each $x_{[q]}$ may be a subvector rather than a single component. The subproblem (1.3) is again separable in the subvectors $d_{[q]}$; we solve

$$\min_{d_{[q]}} \nabla_{[q]} f(x)^T d_{[q]} + \frac{\mu}{2} |d_{[q]}|^2 + \tau |x_{[q]} + d_{[q]}|, \qquad q \in Q'.$$

A closed-form solution is again available; see for example [33, Section II.D]. When $[q]$ contains at least two components, the most natural possibilities for the partial manifold $\mathcal{M}_q$ identified at $x_{[q]} + d_{[q]}$ are the two cases $x_{[q]} + d_{[q]} = 0$ and $x_{[q]} + d_{[q]} \neq 0$. The reduced function $\psi_\tau$ is thus obtained by zeroing out the components $[q]$ in the argument of $f$ for which $x_{[q]} + d_{[q]} = 0$, and omitting these same terms from the summation $\sum_{q \in Q'} \|x_{[q]}\|_2$. Note that both $\nabla \psi_\tau$ and $\nabla^2 \psi_\tau$ will have contributions from the regularization terms $q$ for which $x_{[q]} + d_{[q]} \neq 0$.

Finally, we mention the group-$\ell_\infty$ case in which $P_q(x_{[q]}) = \|x_{[q]}\|_\infty$. The subproblem (1.3) can be solved in time linear in the number of components by using duality to restate it in terms of projection onto an $\ell_1$-norm ball; see [33, Section II.D] and [3] for details. The estimate of active manifold $\mathcal{M}_q$ at a point $x_{[q]} \neq 0$ is

$$\mathcal{M}_q = \{x_{[q]} + z \,:\, z_i = t\,\mathrm{sign}(x_{[q]})_i \text{ for all } i \text{ with } |(x_{[q]})_i| = \|x_{[q]}\|_\infty \text{ and some } t \in \mathbf{R}\},$$

that is, the manifold is (locally) the set of vectors whose components that achieve the absolute maximum are the same as in $x_{[q]}$. It is easy to find linear mappings $F$ and $G$ corresponding to such manifolds, and the restriction of $\phi_\tau$ to $\mathcal{M}$ is as smooth as $f$, in a neighborhood of $x$. From the subgradient of $\|u\|_\infty$ at a point $u \neq 0$ defined by

$$\partial \|u\|_\infty = \begin{cases} [-1, 0] & \text{if } u_i = -\|u\|_\infty, \\ [0, 1] & \text{if } u_i = \|u\|_\infty, \\ 0 & \text{otherwise,} \end{cases}$$

we see that $\partial P_q(x_{[q]} + z)$ is in fact constant for all $x_{[q]} + z \in \mathcal{M}_q$ in a neighborhood of $x_{[q]}$. (When $x_{[q]} = 0$, we have simply $\mathcal{M}_q = 0 \in \mathbf{R}^{|[q]|}$ and $\partial P_q(0) = [-1, 1]^{|[q]|}$.)

## 5. Computational Example.

We present some computational results obtained with Algorithm 1 on $\ell_1$-regularized logistic regression. Our results suggest that the major algorithmic features considered in this paper — block-coordinate relaxation and reduced Newton-like steps — improve the efficienct of the basic prox-linear approach significantly. The results are illustrative rather than definitive; optimized implementations and exhaustive testing of the various algorithmic options on a variety of realistic data sets will be studied elsewhere.

We start by describing the logistic regression application, and give some detail of the algorithmic choices made in our implementation of Algorithm 1.

### 5.1. $\ell_1$-Regularized Logistic Regression.

Suppose we are given a "training set" of $m$ feature vectors $x_i \in \mathbf{R}^n$, $i = 1, 2, \ldots, m$ and corresponding binary labels $b_i \in \{-1, +1\}$, $i = 1, 2, \ldots, m$. Our goal is to learn a regression function $p : \mathbf{R}^n \to [0, 1]$

that predicts the chance of a given feature vector $x$ having label $+1$. (It follows that $1 - p(x)$ is the chance of $x$ having label $-1$.) We parametrize $p$ by a vector $z \in \mathbf{R}^n$, and assume that it has the following form:

$$p(x; z) = \frac{1}{1 + e^{z^T x}}. \tag{5.1}$$

Note that $1 - p(x; z) = 1/(1 + e^{-z^T x})$. We use the training set to find an appropriate value for $z$, ideally one for which $p(x_i; z)$ is close to 1 when $b_i = +1$ and close to zero when $b_i = -1$, for most $i = 1, 2, \dots, m$.

The log-likelihood function for the observed data is

$$\begin{aligned} \mathcal{L}(z) &:= \sum_{i:b_i=+1} \log p(x_i; z) + \sum_{i:b_i=-1} \log(1 - p(x_i; z)) \\ &= \sum_{i:b_i=-1} z^T x_i - \sum_{i=1}^{m} \log(1 + e^{z^T x_i}). \end{aligned} \tag{5.2}$$

In logistic regression, $z$ is chosen to maximize $\mathcal{L}(z)$. We can obtain a sparse $z$ (one with few nonzeros, their locations highlighting the most significant components of the feature vector) by incorporating a multiple of $\|z\|_1$ in the objective. The function to be minimized is thus

$$\phi_\tau(z) = -\frac{1}{m}\mathcal{L}(z) + \tau \|z\|_1. \tag{5.3}$$

(In our experiment below, we include an "intercept" in the regression by appending 1 to each feature vector but *not* including the corresponding additional component of $z$ in the regularization term. For simplicity, however, we omit this detail and base our description on formulation (5.3).)

We now outline the cost of evaluating the function $\mathcal{L}$ and its gradient and Hessian. To evaluate $\mathcal{L}$, we need to compute $z^T x_i$, $i = 1, 2, \dots, m$. Using $X$ to denote the $m \times n$ matrix whose rows are $x_i^T$, we see that the matrix-vector product $Xz$ is required, where $z$ is usually a vector of few nonzeros. If we assume no sparsity of the vectors $x_i$, the cost would be approximately $m$ times the number of nonzeros in $z$. An additional $O(m)$ exponentiations, logarithms, and basic arithmetic operations are required.

For the gradient, we have

$$\nabla \mathcal{L}(z) = -X^T w, \;\; \text{where} \;\; w_i = \begin{cases} -(1 + e^{z^T x_i})^{-1}, & \text{if } b_i = -1, \\ (1 + e^{-z^T x_i})^{-1}, & \text{if } b_i = +1. \end{cases} \tag{5.4}$$

Since the evaluation of $Xz$ has already been performed as part of the function evaluation, the additional costs here are (i) computation of $w$ ($O(m)$ operations); and (ii) computation of $(X^T w)_j$ for the desired components $j \in \mathcal{G} \subset \{1, 2, \dots, n\}$ of the gradient. For most data sets, (i) is dominated by (ii), and the cost of evaluating a partial gradient for indices in the set $\mathcal{G}$ is approximately $|\mathcal{G}|/n$ times the cost of a full gradient.

The Hessian of $\mathcal{L}$ is

$$\nabla^2 \mathcal{L}(z) = -X^T \mathrm{diag}(u) X, \qquad \text{where} \;\; u_i = \frac{e^{z^T x_i}}{(1 + e^{z^T x_i})^2}, \;\; i = 1, 2, \dots, m. \tag{5.5}$$

The cost of computing $u$ is just $O(m)$, as $Xz$ is known from the function evaluation. Hence the main cost of evaluating a principal submatrix $[\nabla^2 \mathcal{L}(z)]_{\mathcal{C}\mathcal{C}}$ of the Hessian, corresponding to the subset of variables $\mathcal{C} \subset \{1, 2, \dots, n\}$ used in the acceleration step, is essentially the cost of a weighted matrix multiplication of the column submatrix $X_{\cdot\mathcal{C}}$ by its transpose. To reduce cost further, we can use sampling of rows of $X$ to obtain an approximation to $[\nabla^2 \mathcal{L}(z)]_{\mathcal{C}\mathcal{C}}$, as in Byrd et al. [2]. After selecting the subset $\mathcal{S} \subset \{1, 2, \dots, m\}$ at random, the approximation is obtained by performing a weighted matrix-matrix multiplication of $X_{\mathcal{S}\mathcal{C}}$ and its transpose.

**5.2. Implementation.** We now discuss some key aspects of the implementation.

*Selection of $Q_k$.* We try two alternative techniques for selecting the relaxation set $Q_k$. In the first scheme, we set $Q_k$ to be some fixed (user-defined) fraction of the indices in $\{1, 2, \dots, n\}$, randomly chosen with equal probability. We refer to this as the "unbiased" scheme. In the alternative "biased" scheme, we include in $Q_k$ all components of $z$ that are nonzero at the current $z_k$, and add some fixed fraction of the other components, randomly chosen. In neither scheme do we check explicitly that the condition (1.5) is satisfied, though our random selection strategy makes it highly unlikely that any index will be overlooked indefinitely.

*Reduced Newton-like steps.* For computation of the accelerated step $\tilde{d}^k$, we use essentially the reduced Newton scheme described in Section 3.4. Optionally, we use a sampled approximation to the reduced Hessian, as described above. Since the reduced Hessian is ill-conditioned in many instances, we also add a damping term $\lambda_k I$ to the reduced Hessian, choosing $\lambda_k$ to be 10 times the norm of the smallest vector in the subdifferential of $\phi_\tau$ taken over the subset of components at which this vector was most recently computed. Finally, after computing the reduced (approximate) Newton step $\tilde{d}^k$, we scale it by a line search parameter $\alpha_k$, setting $\alpha_k$ to be the largest positive value for which none of the components of $x^k + \alpha_k \tilde{d}^k$ have different signs from the components of $x^k + d^k$. If this $\alpha_k$ is too small (below $10^{-2}$ in our implementation), we conclude that the truncation is too severe for the reduced Newton step to have much value, so we discard it without evaluating the function at this point. In the formula (3.2), we set the parameter $\gamma$ to $10^{-3}$, so that the modified step $\tilde{d}^k$ almost always satisfies this condition whenever (3.3) holds.

*Termination.* Testing for termination occurs when the smallest vector in the subdifferential of $\phi_\tau$ over the current relaxation set falls below a specified threshold (in our case, $10^{-6}$). When this occurs, we evaluate the subdifferential over *all* components and check whether it satisfies the same criterion, terminating the algorithm if it does so and continuing to iterate otherwise.

*Choice of $\mu_k$.* Algorithm 1 places few restrictions on the choice of $\mu_k$. In our implementations, we used a scheme rooted in the Levenberg-Marquardt method for nonlinear least squares. The initial choice of $\mu_k$ is $\max(\mu_{\min}, 0.8\mu_{k-1})$, where $\mu_{k-1}$ is the final value of this parameter from the previous iterate (after any increases that are required to satisfy the sufficient decrease condition). We set the parameter $\eta$ (the factor by which $\mu_k$ is increased at inner iterations) to be 2, while $\mu_{\min}$ is set somewhat arbitrarily to $10^{-3}$. We do not choose a value for $\mu_{\text{top}}$, but we do terminate the algorithm with an error message if the value of $\mu_k$ exceeds $10^{20}$.

*Continuation in $\tau$.* A heuristic to perform continuation in the regularization parameter $\tau$ is essential to efficient performance, especially for problems with mild regularization (that is, those for which the number of nonzero components in $z$ is relatively large). When $\tau \geq \tau_{\max} := (1/m)\|\nabla \mathcal{L}(0)\|_\infty$, the minimizer of (5.3) is $z = 0$; this defines the maximum value of interest. Our strategy starts by setting $\tau$ to this value, and decreases successively by a constant factor until the target value is attained. The problem (5.3) is solved for each of these values of $\tau$, starting from a point that is an "adjustment" of the solution for the previous (larger) value of $\tau$. This adjustment is performed by applying a reduced Newton-like method for the current value of $\tau$, *holding the zero components for this previous solution fixed at zero.* If any of the Newton-like steps cause nonzero values to change sign, the adjustment is discarded and we simply use the previous solution as the starting point. Apart from its advantages of efficiency, continuation is useful from the application point of view, as an appropriate value of $\tau$ is usually not known a priori. It is useful to inspect the solutions for a range of values and to use statistical procedures or domain knowledge to select the most appropriate ones. The use of continuation heuristics is common in compressed sensing (see, for example, [8, 6, 33]), but there have been few attempts to provide theoretical support.

**5.3. Results.** The algorithm was implemented in Matlab, by modifying an earlier version of the LPS code at `http://www.cs.wisc.edu/~swright/LPS/`. Computational results were obtained on a four-core 64-bit Linux system with 2.27 GHz Intel Xeon processors and 6 GB main memory.

In reporting the results, we show the total number of function evaluations and the total CPU time summed over all the cores in the system (which may of course exceed wall-clock time). We also show an "equivalent" number of full gradient evaluations, calculated by summing the total number of gradient components evaluated during the run and dividing by $n$. (This statistic captures fairly the effect that evaluation of, say 10% of the elements of the gradient requires about 10% of the effort of a full gradient evaluation.)

Ten continuation steps are used for all data sets, with the final target value being .25 times $\tau_{\max}$. Termination is declared for each value of $\tau$ when the norm of the smallest subgradient vector falls below $10^{-6}$.

We constructed test data sets like those of [23], as follows. Given dimensions $m$ and $n$, each feature vector $x_i \in \mathbb{R}^n$ is completely dense, with elements chosen randomly to be $+1$ or 0 with equal probability. A "true" coefficient vector $\bar{z}$ is selected, and an intercept of $-3$ is introduced, so that the "true" odds function is $p(x;\bar{z}) = (1 + \exp(x^T\bar{z} - 3))^{-1}$. The label is chosen to be $b_i = +1$ with probability $p(x_i, \bar{z})$, and $-1$ otherwise.

In our first data set `bigdata2.mat`, there are $n = 20000$ features and $m = 4000$ training points. The target coefficient vector $\bar{z}$ is selected to have 10 nonzero components, each chosen from $N(0,1)$. The second data set `bigdata11.mat` selects $\bar{z}$ in the same way but has markedly different dimensions: $n = 1000$ and $m = 100000$. In the third data set `bigdata13.mat`, we have $n = 1000$ and $m = 50000$, but $\bar{z}$ has 100 nonzeros with values $10^{\xi}$, where $\xi$ is chosen from $N(0,1)$, independently for each nonzero component. For the training sets `bigdata2.mat` `bigdata11.mat`, the solution for the final value of $\tau$ has 9 and 8 nonzeros, respectively — similar to the 10 nonzeros in the target vector $\bar{z}$. For `bigdata13.mat`, the final solution has 64 nonzeros, which would be sufficient to capture the most significant components from the 100 nonzeros in $\bar{z}$.

Tables 5.1, 5.2, and 5.3 show results for `bigdata2.mat`, `bigdata11.mat`, and `bigdata13.mat`, respectively, for different choices of the parameters governing the size of the partial gradient set $\mathcal{G}$, the density of Hessian sampling $|\mathcal{S}|/m$, and the two different strategies for choice of the relaxation set $Q_k$. Results for the variants that use no reduced Newton acceleration (that is, set $\tilde{d}^k = d^k$ for all $k$) are also shown,

Table 5.1

*Computational results for the* `bigdata2` *set* ($n = 20000$, $m = 4000$).

(a) Performance of the method described in Section 5, showing number of function evaluations (`nf`), number of equivalent gradient evaluations (`ng`), and CPU time (seconds).

| | | unbiased $Q_k$ | | | biased $Q_k$ | | |
|---|---|---|---|---|---|---|---|
| $\lvert\mathcal{G}\rvert/n$ | $\lvert\mathcal{S}\rvert/m$ | `nf` | `ng` | CPU | `nf` | `ng` | CPU |
| 1.00 | 1.00 | 139 | 46.0 | 38.2 | | | |
| 1.00 | 0.05 | 179 | 37.0 | 28.7 | not applicable | | |
| 1.00 | 0.01 | 324 | 71.1 | 36.6 | | | |
| 1.00 | none | 794 | 567.1 | 298.6 | | | |
| 0.20 | 1.00 | 166 | 23.8 | 25.4 | 182 | 27.7 | 28.9 |
| 0.20 | 0.05 | 213 | 22.8 | 25.9 | 226 | 24.5 | 26.7 |
| 0.20 | 0.01 | 361 | 30.7 | 17.4 | 367 | 29.3 | 16.7 |
| 0.20 | none | 961 | 207.1 | 112.5 | 809 | 130.9 | 74.8 |
| 0.05 | 1.00 | 159 | 18.7 | 18.6 | 157 | 17.8 | 16.7 |
| 0.05 | 0.05 | 212 | 18.0 | 19.8 | 240 | 18.9 | 20.2 |
| 0.05 | 0.01 | 374 | 24.1 | 14.3 | 409 | 20.3 | 12.6 |
| 0.05 | none | 1140 | 157.4 | 90.5 | 764 | 43.3 | 30.2 |
| 0.01 | 1.00 | 153 | 16.3 | 15.7 | 157 | 16.4 | 15.1 |
| 0.01 | 0.05 | 200 | 15.4 | 16.8 | 228 | 16.5 | 16.8 |
| 0.01 | 0.01 | 338 | 16.7 | 10.2 | 365 | 16.8 | 10.5 |
| 0.01 | none | 1031 | 130.1 | 74.9 | 718 | 20.9 | 17.7 |

(b) Profile of a single run from Table 5.1(a): $\lvert\mathcal{G}\rvert/n = .01$, $\lvert\mathcal{S}\rvert/m = .01$, for unbiased $Q_k$.

| $\tau$ | nonzeros | iterations | `nf` | `ng` | CPU |
|---|---|---|---|---|---|
| 4.58e-02 | 1 | 2 | 5 | 1.0 | 0.52 |
| 3.98e-02 | 4 | 10 | 21 | 2.1 | 1.19 |
| 3.47e-02 | 4 | 2 | 24 | 1.0 | 0.60 |
| 3.02e-02 | 6 | 13 | 36 | 2.1 | 1.30 |
| 2.63e-02 | 6 | 2 | 27 | 1.0 | 0.62 |
| 2.29e-02 | 7 | 11 | 39 | 2.1 | 1.27 |
| 1.99e-02 | 7 | 2 | 23 | 1.0 | 0.60 |
| 1.73e-02 | 8 | 25 | 70 | 3.2 | 2.13 |
| 1.51e-02 | 8 | 3 | 28 | 1.0 | 0.64 |
| 1.31e-02 | 8 | 3 | 28 | 1.0 | 0.65 |
| 1.14e-02 | 8 | 3 | 37 | 1.0 | 0.68 |
| total | | | 338 | 16.7 | 10.2 |

on the lines with the entry "none" in the column headed $\lvert\mathcal{S}\rvert/m$. Note that when the full gradient is evaluated ($\lvert\mathcal{G}\rvert/n = 1$) there is no distinction between the "biased" and "unbiased" sampling strategies, since $Q_k \equiv \{1, 2, \ldots, n\}$. Hence, the top right box is not filled.

Tables 5.1(b), 5.2(b), and 5.3(b) show a profile for a particular run of the code for a single set of parameter choices. We note that all variants produced the same solutions (in terms of the final number of nonzeros, for each value of $\tau$). These tables show how the work was distributed between the sequence of $\tau$ values used in the

Table 5.2

*Computational results for the* bigdata11 *set* ($n = 1000$, $m = 100000$).

(a) Performance of the method described in Section 5, showing number of function evaluations (`nf`), number of equivalent gradient evaluations (`ng`), and CPU time (seconds).

| | | unbiased $Q_k$ | | | biased $Q_k$ | | |
|---|---|---|---|---|---|---|---|
| $\|\mathcal{G}\|/n$ | $\|\mathcal{S}\|/m$ | `nf` | `ng` | CPU | `nf` | `ng` | CPU |
| 1.00 | 1.00 | 139 | 40.4 | 44.5 | | | |
| 1.00 | 0.05 | 139 | 39.4 | 45.1 | not applicable | | |
| 1.00 | 0.01 | 145 | 39.4 | 45.6 | | | |
| 1.00 | none | 580 | 500.3 | 356.5 | | | |
| 0.20 | 1.00 | 164 | 24.1 | 39.8 | 159 | 24.5 | 37.9 |
| 0.20 | 0.05 | 177 | 27.0 | 43.0 | 183 | 27.2 | 43.9 |
| 0.20 | 0.01 | 175 | 25.5 | 42.1 | 183 | 25.7 | 43.1 |
| 0.20 | none | 1130 | 283.7 | 249.0 | 573 | 114.6 | 114.0 |
| 0.05 | 1.00 | 161 | 19.3 | 34.5 | 160 | 19.2 | 35.6 |
| 0.05 | 0.05 | 161 | 18.3 | 34.5 | 181 | 21.0 | 39.7 |
| 0.05 | 0.01 | 173 | 20.4 | 37.8 | 181 | 19.5 | 38.0 |
| 0.05 | none | 1352 | 222.7 | 226.2 | 542 | 44.5 | 67.3 |
| 0.01 | 1.00 | 163 | 18.8 | 34.9 | 160 | 17.9 | 34.6 |
| 0.01 | 0.05 | 167 | 18.8 | 34.9 | 182 | 18.1 | 39.1 |
| 0.01 | 0.01 | 164 | 17.8 | 33.7 | 180 | 19.0 | 38.0 |
| 0.01 | none | 1133 | 167.0 | 177.0 | 540 | 26.1 | 56.8 |

(b) Profile of a single run from Table 5.2(a): $|\mathcal{G}|/n = .01$, $|\mathcal{S}|/m = .01$, for unbiased $Q_k$.

| $\tau$ | nonzeros | iterations | `nf` | `ng` | CPU |
|---|---|---|---|---|---|
| 1.89e-02 | 1 | 2 | 5 | 1.0 | 0.87 |
| 1.65e-02 | 2 | 8 | 17 | 2.1 | 2.43 |
| 1.43e-02 | 2 | 2 | 11 | 1.0 | 1.31 |
| 1.25e-02 | 3 | 7 | 20 | 2.1 | 3.77 |
| 1.09e-02 | 3 | 2 | 11 | 1.0 | 2.15 |
| 9.46e-03 | 3 | 2 | 11 | 1.0 | 2.05 |
| 8.23e-03 | 4 | 7 | 19 | 2.1 | 4.64 |
| 7.17e-03 | 5 | 6 | 19 | 2.1 | 4.86 |
| 6.24e-03 | 6 | 5 | 16 | 2.1 | 3.96 |
| 5.43e-03 | 9 | 9 | 22 | 2.2 | 4.90 |
| 4.73e-03 | 9 | 2 | 13 | 1.1 | 2.78 |
| total | | | 164 | 17.8 | 33.7 |

continuation heuristic, for one of the parameter combinations that worked fairly well across all data sets.

We can make several general observations about the performance reported here.

1. The use of reduced Newton acceleration generally yields a vast performance improvement over the method that uses only first-order information. In a few cases, however, the first-order method is competitive, for example in the bottom right of Tables 5.1(a) and 5.2(a), where we evaluate only 1% of the gradient and use the “biased” scheme for selecting the relaxation set.

Table 5.3
*Computational results for the* `bigdata13` *set* ($n = 1000$, $m = 5000$).

(a) Performance of the method described in Section 5, showing number of function evaluations (`nf`), number of equivalent gradient evaluations (`ng`), and CPU time (seconds).

| | | unbiased $Q_k$ | | | biased $Q_k$ | | |
|---|---|---|---|---|---|---|---|
| $\lvert\mathcal{G}\rvert/n$ | $\lvert\mathcal{S}\rvert/m$ | `nf` | `ng` | CPU | `nf` | `ng` | CPU |
| 1.00 | 1.00 | 338 | 173.9 | 140.1 | not applicable | | |
| 1.00 | 0.05 | 304 | 127.2 | 95.2 | | | |
| 1.00 | 0.01 | 409 | 166.2 | 127.2 | | | |
| 1.00 | none | 13158 | 12994.3 | 4739.5 | | | |
| 0.20 | 1.00 | 332 | 49.4 | 83.8 | 323 | 49.2 | 81.2 |
| 0.20 | 0.05 | 398 | 56.8 | 81.9 | 409 | 57.7 | 85.0 |
| 0.20 | 0.01 | 521 | 65.8 | 107.2 | 556 | 77.9 | 118.3 |
| 0.20 | none | 9118 | 1925.0 | 1026.1 | 12376 | 2914.2 | 1508.8 |
| 0.05 | 1.00 | 262 | 35.2 | 62.3 | 343 | 40.5 | 83.7 |
| 0.05 | 0.05 | 353 | 38.3 | 66.2 | 401 | 41.8 | 73.6 |
| 0.05 | 0.01 | 474 | 50.7 | 88.1 | 561 | 56.1 | 107.4 |
| 0.05 | none | 9472 | 1107.3 | 795.9 | 12418 | 1172.7 | 961.5 |
| 0.01 | 1.00 | 258 | 31.2 | 57.9 | 285 | 34.3 | 67.8 |
| 0.01 | 0.05 | 300 | 36.5 | 54.3 | 368 | 36.7 | 65.8 |
| 0.01 | 0.01 | 411 | 40.9 | 72.3 | 487 | 42.2 | 89.1 |
| 0.01 | none | 3019 | 463.7 | 285.5 | 12255 | 703.9 | 817.7 |

(b) Profile of a single run from Table 5.3(a): $|\mathcal{G}|/n = .01$, $|\mathcal{S}|/m = .01$, for unbiased $Q_k$.

| $\tau$ | nonzeros | iterations | `nf` | `ng` | CPU |
|---|---|---|---|---|---|
| 3.18e-02 | 1 | 2 | 5 | 1.0 | 0.42 |
| 2.77e-02 | 9 | 11 | 22 | 3.2 | 3.40 |
| 2.41e-02 | 14 | 8 | 26 | 2.3 | 3.12 |
| 2.10e-02 | 15 | 9 | 29 | 2.3 | 4.72 |
| 1.83e-02 | 21 | 11 | 32 | 2.4 | 4.88 |
| 1.59e-02 | 32 | 15 | 43 | 3.8 | 6.86 |
| 1.38e-02 | 38 | 21 | 53 | 5.2 | 8.85 |
| 1.20e-02 | 43 | 23 | 56 | 5.5 | 10.23 |
| 1.05e-02 | 45 | 10 | 40 | 4.3 | 8.66 |
| 9.13e-03 | 56 | 23 | 57 | 5.8 | 10.89 |
| 7.95e-03 | 64 | 14 | 48 | 5.1 | 10.23 |
| total | | | 411 | 40.9 | 72.3 |

2. There is some performance benefit from using partial gradient evaluations, but it is not very significant on these data sets. One reason is that to test termination for each value of $\tau$, we evaluate one full gradient vector $\nabla\mathcal{L}(z)$. In many cases (for examples the runs reported in Tables 5.2(b), and 5.3(b)) this is by far the most expensive operation for each $\tau$; the reduced gradient and Hessian evaluations needed to perform the iterations cost little by comparison. (Note that many entries in the `ng` column of these tables are barely more than the level of 1.0 needed to perform the convergence test.) By applying

techniques like those proposed recently in [5], we can potentially avoid this bottleneck and obtain a rigorous convergence test without evaluating a full gradient. We leave details to later work.
3. The simplest variant of Algorithm 1 — full gradient evaluation, no reduced-Newton acceleration — gives for all three data sets the poorest performace by far.
4. The benefits from using sampled approximations to the reduced Hessians are not very significant, possibly because the reduced Hessian are such small matrices in these computations that the time spent evaluating them is a relatively small part of the computation.
5. There is no clear benefit to be gained from using the "biased" technique for choosing $Q_k$.
6. The "adjustment" strategy (for obtaining a warm start for each $\tau$ by applying a reduced Newton method to the solution from the previous $\tau$ value) led to significantly faster run times than simply using the previous solution as the warm start. This is especially true for full-gradient variants, which are uncompetitive if adjustment is not performed.

We repeat that our results are mainly illustrative in nature. The conclusions may be quite different in other data regimes — for example if the dimensions were much larger, if the data matrix $X$ were sparse and derived from a real application (rather than a random test model), or if the number of nonzero elements for the final value of $\tau$ were larger.

**5.4. Reproducibility.** In the interests of reproducibility, we have placed the code at `http://www.cs.wisc.edu/~swright/LPS/` (see version 2.2 of LPS), along with a routine `TestTables.m` for generating the data in the tables of this paper. Because of the random nature of the algorithm (both in the selection of gradient components and in the selection of training points to use in the Hessian estimate), results will usually differ from run to run, and of course according to the capablities of the computer on which they execute. The LPS distribution also contains data for smaller standard test problems arising from real applications, and a routine to run the code on these examples. They generally solve in a few seconds.

We include in the distribution the data sets number of smaller test problems, mostly following those reported in [22], along with code to run these problems.

**Acknowledgments.** I am grateful to Jon Borwein for interesting discussions and support during a visit to CARMA (University of Newcastle), where some of this work was performed. I also acknowledge several interesting conversations with Jorge Nocedal on algorithms for logistic regression and support vector machines. I thank Sangkyun Lee for a close reading of the manuscript and important suggestions. Finally, I thank the editor and referees for their helpful comments.

## Appendix A. Manifold Characterization: Proofs.

*Proof.* (Lemma 2.1) We prove the result constructively via the implicit function theorem; see for example [18, Theorem A.2]. Assuming WLOG that $\nabla F(\bar{z})$ is orthonormal and defining $Y$ as in the statement of the theorem, we observe that the $m \times m$ matrix $\begin{bmatrix} \nabla F(\bar{z}) & Y \end{bmatrix}$ is nonsingular, in fact orthogonal. We now consider the map $\Phi : \mathbb{R}^m \times \mathbb{R}^{m-k} \to \mathbb{R}^m$ defined as follows:

$$\Phi(z, y) = \begin{bmatrix} F(z) \\ Y^T(z - \bar{z}) - (y - \bar{y}) \end{bmatrix}.$$

Note first that $\Phi(\bar{z}, \bar{y}) = 0$. Second, we have

$$\nabla_z \Phi(z, y) = \begin{bmatrix} \nabla F(z) & Y \end{bmatrix},$$

which is nonsingular at the point $(z, y) = (\bar{z}, \bar{y})$, as noted above. Third, $\Phi$ is $p$ times continuously differentiable in a neighborhood of $(\bar{z}, \bar{y})$, by the assumed properties of $F$. Thus, by applying the implicit function theorem, we have that $z$ is implicitly a $C^p$ function of $y$. We identify $G(y)$ with $z$ to obtain the main result.

For the final statements, note from the implicit function theorem that

$$\nabla G(y) = -\nabla_y \Phi(z, y)[\nabla_z \Phi(z, y)]^{-1} = \begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} \nabla F(z) & Y \end{bmatrix}^{-1},$$

and thus at $y = \bar{y}$, using $\begin{bmatrix} \nabla F(\bar{z}) & Y \end{bmatrix}^{-1} = \begin{bmatrix} \nabla F(\bar{z}) & Y \end{bmatrix}^T$, we have $\nabla G(\bar{y}) = Y^T$. Thus, by Taylor's theorem, we have

$$G(y) - \bar{z} = G(y) - G(\bar{y}) = Y(y - \bar{y}) + O(|y - \bar{y}|^2).$$

□

*Proof.* (Lemma 2.2) We again invoke the implicit function theorem to prove the claim. Defining the function $\Psi : \mathbf{R}^{m-k} \times \mathbf{R}^k \times \mathbf{R}^m \to \mathbf{R}^m$ by

$$\Psi(y, v; z) := G(y) + \nabla F(\bar{z})v - z,$$

we note first that $\Psi(\bar{y}, 0; \bar{z}) = 0$. Second, we have

$$\nabla_{(y,v)} \Psi(y, v; z) = \begin{bmatrix} \nabla G(y) \\ \nabla F(z)^T \end{bmatrix},$$

so that

$$\nabla_{(y,v)} \Psi(\bar{y}, 0; \bar{z}) = \begin{bmatrix} \nabla G(\bar{y}) \\ \nabla F(\bar{z})^T \end{bmatrix} = \begin{bmatrix} Y^T \\ \nabla F(\bar{z})^T \end{bmatrix},$$

which is an orthogonal (and hence nonsingular) matrix. Third, since $G$ is a $C^p$ function of $y$ in a neighborhood of $\bar{y}$, we have that $\Psi$ is also $p$ times continuously differentiable in a neighborhood of $(\bar{y}, 0, \bar{z})$. Defining $(y, v)$ to be the solution of $\Psi(y, v; z) = 0$ for a given $z$, the conclusion follows immediately from the implicit function theorem. □

Finally, we include the following result for completeness.

LEMMA A.1. *Suppose we have a function $h : \mathbf{R}^m \to \bar{\mathbf{R}}$, a point $\bar{z}$, and a manifold $\mathcal{M}$ with $\bar{z} \in \mathcal{M} \subset \mathbf{R}^m$ such that $h$ is partly smooth at $\bar{z}$ with respect to $\mathcal{M}$. Suppose in addition that the nondegenerate criticality condition $0 \in \operatorname{ri} \partial h(\bar{z})$ holds. Then there is $\epsilon > 0$ such that for all $d \in N_{\mathcal{M}}(\bar{z})$, we have*

$$\sup_{g \in \partial h(\bar{z})} g^T d \geq \epsilon |d|.$$

*Proof.* Since $\operatorname{aff} \partial h(\bar{z}) = N_{\mathcal{M}}(\bar{z})$ and $0 \in \operatorname{ri} \partial h(\bar{z})$, there is $\epsilon > 0$ such that $g \in \partial h(\bar{z})$ for all $g \in N_{\mathcal{M}}(\bar{z})$ with $|g| \leq \epsilon$. Thus for any $d \in N_{\mathcal{M}}(\bar{z})$, we have $\epsilon d/|d| \in \partial h(\bar{z})$, and the result follows immediately. □

## REFERENCES

[1] S. Bakin, *Adaptive regression and model selection in data mining problems*, PhD thesis, Australian National University, 1999.

[2] R. H. Byrd, G. M. Chin, W. Neveitt, and J. Nocedal, *On the use of stochastic hessian information in unconstrained optimization*, Technical Report, Northwestern University, June 2010.

[3] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, *Efficient projections on the l1-ball for learning in higher dimensions*, in Proceedings of the International Conference on Machine Learning - ICML, Helsinki, 2008.

[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Annals of Statistics, 32 (2004), pp. 407–499.

[5] L. El Ghaoui, V. Viallon, and T. Rabbani, *Safe feature elimination for the LASSO and sparse supervised learning problems*, Technical Report, EECS Department, University of California-Berkeley, 2011. arXiv:1009.4219v2.

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

[7] J. Friedman, T. Hastie, and R. Tibshirani, *Regularization paths for generalized inear models via coordinate descent*, Tech. Report April, Department of Statistics, Stanford University, 2009.

[8] E. T. Hale, W. Yin, and Y. Zhang, *A fixed-point continuation method for $\ell_1$-minimization: Methodology and convergence*, SIAM Journal on Optimization, 19 (2008), pp. 1107–1130.

[9] W. Hare and A. Lewis, *Identifying active constraints via partial smoothness and prox-regularity*, Journal of Convex Analysis, 11 (2004), pp. 251–266.

[10] W. L. Hare, *A proximal method for identifying active manifolds*, Computational Optimization and Applications, 43 (2009), pp. 295–306.

[11] ———, *Identifying active manifolds in regularization problems*, tech. report, Mathematics Department, UBC O, 2010. To appear in *Proceedings of the Interdisciplinary Workshop on Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, B.I.R.S.

[12] J. Kim, Y. Kim, and Y. Kim, *Gradient LASSO algorithm*, technical report, Seoul National University, 2006.

[13] Y. Kim, J. Kim, and Y. Kim, *Blockwise sparse regression*, Statistica Sinica, 16 (2006), pp. 375–390.

[14] K. Koh, S.-J. Kim, and S. Boyd, *An interior-point method for large-scale $\ell_1$-regularized logistic regression*, Journal of Machine Learning Research, 8 (2007), pp. 1519–1555.

[15] A. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM Journal on Optimization, 13 (2003), pp. 702–725.

[16] A. S. Lewis and S. J. Wright, *A proximal method for composite minimization*, Optimization Technical Report, University of Wisconsin-Madison, August 2008.

[17] L. Meier, S. van de Geer, and P. Buhlmann, *The group lasso for logistic regression*, Journal of the Royal Statistical Society B, 70 (2008), pp. 53–71.

[18] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, second ed., 2006.

[19] B. Recht, M. Fazel, and P. Parrilo, *Guaranteed minimum-rank solutions to linear matrix equations via nuclear norm minimization*, SIAM Review, 52 (2010), pp. 471–501.

[20] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[21] S. K. Shevade and S. S. Keerthi, *A simple and efficient algorithm for gene selection using sparse logistic regression*, Bioinformatics, 19 (2003), pp. 2246–2253.

[22] J. Shi, W. Yin, S. Osher, and P. Sajda, *A fast hybrid algorithm for large scale $\ell_1$-regularized logistic regression*, Journal of Machine Learning Research, 11 (2010), pp. 713–741.

[23] 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 (2008), pp. 137–153.

[24] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, Journal of the Royal Statistical Society B, 58 (1996), pp. 267–288.

[25] P. Tseng, *Approximation accuracy, gradient methods, and error bound for structured convex optimization*, technical report, Deparment of Mathematics, University of Washington, 2009. To appear in Mathematical Programming, Series B.

[26] ———, *Further results on a stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Transactions on Information Theory, 55 (2009), pp. 888–899.

[27] P. Tseng and S. Yun, *A block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization*, Journal of Optimization Theory and Applications, 140 (2009), pp. 513–535.

[28] ———, *A coordinate gradient descent method for nonsmooth separable minimization*, Mathematical Programming, Series B, 117 (2009), pp. 387–423.

[29] ———, *A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training*, Computational Optimization and Applications, 47 (2010), pp. 179–206.

[30] B. Turlach, W. N. Venables, and S. J. Wright, *Simultaneous variable selection*, Technometrics, 47 (2005), pp. 349–363.

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

[32] S. J. Wright, *Identifiable surfaces in constrained optimization*, SIAM Journal on Control and OptimizationOptimization, 31 (1993), pp. 1063–1079.

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

[34] G.-X. Yuan, K.-W. Chang, C.-J. Hsieh, and C. Lin, *A comparison of optimization methods for large-scale l1-regularized linear classification*, technical report, Department of Computer Science, National Taiwan University, 2009.

[35] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. Royal Statistical Society B, 68 (2006), pp. 49–67.