SCISPACE
formerly Typeset

Open access • Journal Article • DOI:10.1080/10556788.2011.591398

On the convergence of an active-set method for $\ell 1$ minimization — Source link

Zaiwen Wen, Wotao Yin, Hongchao Zhang, Donald Goldfarb

Institutions: Shanghai Jiao Tong University, Rice University, Louisiana State University, Columbia University

Published on: 01 Dec 2012 - Optimization Methods & Software (Taylor & Francis)

Topics: Rate of convergence, Active set method, Line search and Stationary point

Related papers:

- A Fast Algorithm for Sparse Reconstruction Based on Shrinkage, Subspace Optimization, and Continuation
- A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems
- Compressed sensing
- Sparse Reconstruction by Separable Approximation
- Fixed-Point Continuation for ℓ_1-Minimization: Methodology and Convergence

Share this paper:

View more about this paper here: https://typeset.io/papers/on-the-convergence-of-an-active-set-method-for-l1-453le5zwsv

ON THE CONVERGENCE OF AN ACTIVE SET METHOD FOR ℓ_1 MINIMIZATION

ZAIWEN WEN †, WOTAO YIN ‡, HONGCHAO ZHANG §, AND DONALD GOLDFARB ¶

Abstract. We analyze an abridged version of the active-set algorithm FPC_AS proposed in [18] for solving the l_1-regularized problem, i.e., a weighted sum of the l_1-norm $\|x\|_1$ and a smooth function $f(x)$. The active set algorithm alternatively iterates between two stages. In the first "nonmonotone line search (NMLS)" stage, an iterative first-order method based on "shrinkage" is used to estimate the support at the solution. In the second "subspace optimization" stage, a smaller smooth problem is solved to recover the magnitudes of the nonzero components of x. We show that NMLS itself is globally convergent and the convergence rate is at least R-linearly. In particular, NMLS is able to identify of the zero components of a stationary point after a finite number of steps under some mild conditions. The global convergence of FPC_AS is established based on the properties of NMLS.

Key words. l_1-minimization, basis pursuit, compressed sensing, subspace optimization, active set, continuation

AMS subject classifications. 49M29, 65K05, 90C25, 90C06

1. Introduction. In this paper, we consider the convergence properties of a two-stage active set algorithm for the l_1-regularized minimization problem

$$\min_{x\in\mathbb{R}^n} \psi_\mu(x) := \mu\|x\|_1 + f(x), \tag{1.1}$$

where $\mu > 0$ and $f(x) : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. This work is motivated by an efficient algorithm proposed in [18] for sparse reconstruction, which is applied to $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $A \in \mathbb{R}^{m\times n}$. Note that problem (1.1) is differentiable except at points where at least one component of x is zero. We can define the active set $\mathcal{A}(x)$ to be the set of indices corresponding to the zero components and the inactive set $\mathcal{I}(x)$ to be the support of x, i.e.,

$$\mathcal{A}(x) := \{i \in [1, n] \,|\, |x_i| = 0\} \text{ and } \mathcal{I}(x) := \{i \in [1, n] \,|\, |x_i| > 0\}. \tag{1.2}$$

In our framework, a first-order method based on the so-called "shrinkage" operation is used to identify an active set and then a second-order method is used to explore a smooth subproblem based on this active set. These two operations are iterated until convergence criteria are satisfied. Similar ideas can be found in nonlinear programming. For example, gradient projection and conjugate gradient steps have been combined to solve problems with bound constraints or linear constraints in [2, 3, 12, 15, 16], and linear programming and quadratic programming subproblems have been used to solve general nonlinear programs in [4, 5]. However, our algorithm is different from the active set algorithm [14] on how the working index set is chosen. Thanks to the solution sparsity, our approach is more aggressive and effective.

Our contributions are as follows. First, we interpret shrinkage as a gradient projection method for smooth problems with simple bound constraints [6, 12, 19], since problem (1.1) can be reformulated as the

†Department of Mathematics and Institute of Natural Sciences, Shanghai Jiaotong University (zw2109@sjtu.edu.cn). Research supported in part by NSF DMS-0439872 through UCLA IPAM.

‡Department of Computational and Applied Mathematics, Rice University. (wotao.yin@rice.edu). Research supported in part by NSF CAREER Award DMS-07-48839, ONR Grant N00014-08-1-1101, the U. S. Army Research Laboratory and the U. S. Army Research Office grant W911NF-09-1-0383 and an Alfred P. Sloan Research Fellowship.

§Department of Mathematics and Center for Computational & Technology at Louisiana State University. (hozhang@math.lsu.edu).

¶Department of Industrial Engineering and Operations Research, Columbia University. (goldfarb@columbia.edu). Research supported in part by NSF Grant DMS 10-16571, ONR Grant N00014-8-1-1118 and DOE Grant DE-FG02-08-25856.

minimization of a smooth objective function over a simple convex set (as explained later in (2.3)). Hence, the analysis in [6, 12, 19] can be extended naturally to explain the behavior of shrinkage. This point of view for shrinkage offers us the opportunity of utilizing research results in nonlinear programming to study problem (1.1). Second, we provide another successful example of how to combine powerful algorithmic components with different purposes together coherently. While subspace optimization ("de-biasing") has been proved to be useful as a post-processing procedure in GPSR [11] and FPC [13], it is not easy to integrate it into the main iterations as in our algorithm. Moreover, we can guarantee certain convergence results although the analysis becomes very difficult as our algorithm alternates between different phases.

This paper is organized as follows. In section 2, we present some useful properties of shrinkage. In section 3, we briefly review our two-stage active set algorithm. We analyze our first stage algorithm NMLS in subsection 4.1. We show that NMLS is globally convergent and the convergence rate is at least R-linear. The identification of the active set is presented in subsection 4.2. Global convergence of the active set algorithm is presented in subsection 4.3.

2. Properties of shrinkage. We now briefly review the iterative shrinkage procedure for solving (1.1). Given an initial point x^0, the algorithm iteratively computes

$$x^{k+1} := \mathcal{S}\left(x^k - \lambda g^k, \mu\lambda\right), \tag{2.1}$$

where $g^k := \nabla f(x^k)$, $\mu, \lambda > 0$, and for $y \in \mathbb{R}^n$ and $\nu \in \mathbb{R}$, the shrinkage operator is defined as

$$\mathcal{S}(y, \nu) := \operatorname{sgn}(y) \odot \max\left\{|y| - \nu, \mathbf{0}\right\}. \tag{2.2}$$

The convergence of the iterative shrinkage operation (2.1) has been studied in [7, 9, 13] under suitable conditions on λ and the Hessian $\nabla^2 f$. An appealing feature proved in [13] is that (2.1) yields the support and the signs of the minimizer x^* of (1.1) after a finite number of steps under favorable conditions. For more references related to shrinkage, the reader is referred to [18].

The scheme (2.1) exhibits characteristics of the gradient projection method for bound constrained optimization. In fact, (1.1) can be reformulated as

$$\min f(x) + \mu\xi, \quad \text{subject to } (x, \xi) \in \Omega := \{(x, \xi) \mid \|x\|_1 \le \xi\}, \tag{2.3}$$

and (2.2) yields the solution of $\min_{x\in\mathbb{R}^n} \nu\|x\|_1 + \frac{1}{2}\|x - y\|_2^2$, which is equivalent to

$$\min \frac{1}{2}\|x - y\|_2^2 + \nu\xi, \quad \text{subject to } (x, \xi) \in \Omega. \tag{2.4}$$

Hence, (2.1) and (2.2) performs like a gradient projection operator. Denote by $d^{(\lambda)}(x)$ the search direction generated by shrinkage (2.1), i.e.,

$$d^{(\lambda)}(x) := x^+ - x \text{ and } x^+ = \mathcal{S}(x - \lambda g, \mu\lambda). \tag{2.5}$$

We can obtain results similar to Proposition 2.1 in [12].

LEMMA 2.1. *We have the following properties:*

P1. $(\mathcal{S}(x, \nu) - x)^\top (y - \mathcal{S}(x, \nu)) + \nu(\xi - \|\mathcal{S}(x, \nu)\|_1) \ge 0$ *for all* $x \in \mathbb{R}^n$ *and* $(y, \xi) \in \Omega$ *and* $\nu > 0$.

P2. $(\mathcal{S}(x, \nu) - \mathcal{S}(y, \nu))^\top (x - y) \ge \|\mathcal{S}(x, \nu) - \mathcal{S}(y, \nu)\|_2^2$ *for all* $x, y \in \mathbb{R}^n$ *and* $\nu > 0$.

P3. $\|\mathcal{S}(x,\nu)-\mathcal{S}(y,\nu)\| \le \|x-y\|$ *for all* $x,y\in\mathbb{R}^n$ *and* $\nu>0$.

P4. $\|d^{(\lambda)}(x)\|$ *is nondecreasing in* $\lambda>0$ *for any* $x\in\mathbb{R}^n$*, i.e.,* $\|d^{(\lambda_1)}(x)\| \ge \|d^{(\lambda_2)}(x)\|$ *for* $\lambda_1\ge\lambda_2>0$.

P5. $\frac{\|d^{(\lambda)}(x)\|}{\lambda}$ *is nonincreasing in* $\lambda>0$ *for any* $x\in\mathbb{R}^n$*, i.e.,* $\frac{\|d^{(\lambda_1)}(x)\|}{\lambda_1} \le \frac{\|d^{(\lambda_2)}(x)\|}{\lambda_2}$ *for* $\lambda_1\ge\lambda_2>0$.

P6. $g^\top d^{(\lambda)}(x)+\mu(\|x^+\|_1-\|x\|_1) \le -\frac{1}{\lambda}\|d^{(\lambda)}(x)\|_2^2$ *for any* $x\in\mathbb{R}^n$ *and* $\lambda>0$.

P7. *For any* $x^*\in\mathbb{R}^n$ *and* $\lambda>0$, $d^{(\lambda)}(x^*)=\mathbf{0}$ *if and only if* x^* *is a stationary point for* (1.1).

P8. *Suppose* x^* *is a stationary point for* (1.1). *If for some* $x\in\mathbb{R}^n$*, there exist scalars* $\omega>0$ *and* $L>0$ *such that*

$$(g(x)-g(x^*))^\top(x-x^*) \ge \omega\|x-x^*\|_2^2, \tag{2.6}$$

$$\|g(x)-g(x^*)\| \le L\|x-x^*\|. \tag{2.7}$$

Then we have $\|x-x^*\| \le \left(\frac{1+\lambda L}{\lambda\omega}\right)\|d^{(\lambda)}(x)\|$.

In fact, **P1** is simply the optimality conditions for (2.4). **P2** and **P3** are related to the nonexpansiveness of shrinkage. **P4** and **P5** show the relationship between the parameter λ and the norm of the search direction. **P6** will be used in the derivation of our nonmonotone line search condition. **P7** gives a useful alternative characterization of stationarity. And **P8** is useful in our convergence analysis. A proof of Lemma 2.1 is given in Appendix A.

3. An active-set algorithm. In order to simplify the writing and theoretical analysis, we study an abridged version of the two-stage active set algorithm proposed in [18]. The approach is motivated as follows. While shrinkage is very effective in obtaining a support superset, it can take a lot of iterations to recover the signal values. On the other hand, if one imposes the signs of the components of the variable x that are the same as those of the exact solution, problem (1.1) reduces to a small smooth optimization problem, which can be relatively easily solved to obtain x. Consequently, the key components are the identification of a "good" support set by using shrinkage and the construction of a suitable approximate smooth optimization problem.

In the first stage, we accelerate (2.1) by using a non-monotone line search method based on a strategy in [19]. Specifically, the new points are generated iteratively in the form $x^{k+1}=x^k+\alpha_k d^k$, where $d^k := d^{(\lambda^k)}(x^k)$ is the search direction defined by (2.5) for some $\lambda^k>0$, $\alpha_\rho>0$, $\alpha_k=\alpha_\rho\rho^h$ and h is the smallest integer that satisfies

$$\psi_\mu(x^k+\alpha_\rho\rho^h d^k) \le C_k+\sigma\alpha_\rho\rho^h\Delta^k. \tag{3.1}$$

Here, $\sigma>0$,

$$\Delta^k := (g^k)^\top d^k+\mu\|x^{k+}\|_1-\mu\|x^k\|_1,$$

$C_0=\psi_\mu(x^0)$, and let the next reference value C_{k+1} be taken as a convex combination of C_k and $\psi_\mu(x^{k+1})$, i.e.,

$$C_{k+1}=(\eta Q_k C_k+\psi_\mu(x^{k+1}))/Q_{k+1}, \tag{3.2}$$

where $\eta\in(0,1)$, $Q_{k+1}=\eta Q_k+1$ and $Q_0=1$.

We now describe the subspace optimization in the second stage. The active indices are further subdivided

into two sets

$$\mathcal{A}_{\pm}(x) := \{i \in \mathcal{A}(x) \,|\, |g_i(x)| < \mu\} \text{ and } \mathcal{A}_0(x) := \{i \in \mathcal{A}(x) \,|\, |g_i(x)| \geq \mu\}. \tag{3.3}$$

Note that $|g_i(x^*)| = \mu$ for $i \in \mathcal{A}_0(x^*)$ if x^* is an optimal solution of (1.1). We switch to subspace optimization, if for some fixed constant $\delta > 0$, either one of the two conditions

$$\lambda^{k-1}\|g^k_{\mathcal{I}(x^k)}\| > \delta\|d^{(\lambda^{k-1})}(x^{k-1})\|_2 \text{ and } \|(|g(x^k)| - \mu)_{\mathcal{I}(x^k)\cup\mathcal{A}_0(x^k)}\|_\infty \leq \|d^{(1)}(x^k)\|/\delta \tag{3.4}$$

$$|\psi^k_\mu - \psi^{k-1}_\mu| \leq \epsilon \max(|\psi^k_\mu|, |\psi^{k-1}_\mu|, 1) \tag{3.5}$$

is satisfied during the shrinkage stage (The justification for (3.4) and (3.5) will be specified in subsection 4.3). Then, we explore the face defined by the support of x^k by solving the problem:

$$\min \quad \varphi_\mu(x) := \mu \operatorname{sgn}(x^k_{\mathcal{I}^k})^\top x_{\mathcal{I}^k} + f(x), \text{ s.t. } x \in \Omega(x^k), \tag{3.6}$$

where

$$\Omega(x^k) := \left\{x \in \mathbb{R}^n : \operatorname{sgn}(x^k_i)x_i \geq 0, i \in \mathcal{I}(x^k) \text{ and } x_i = 0, i \in \mathcal{A}(x^k)\right\}. \tag{3.7}$$

To ensure convergence of the active set algorithm, we require that the iterates for solving subspace optimization satisfy:

CONDITION 3.1. *Denote by $x^{k,j}$ the j-th iteration for solving (3.6) starting from $x^{k,0} = x^k$. Then, the iterates are feasible and monotone decreasing, i.e., $\varphi_\mu(x^{k,j+1}) \leq \varphi_\mu(x^{k,j})$ for each j.*

The detailed description of our approach is presented in Algorithm 1 (FPC_AS).

Algorithm 1: FPC_AS Algorithm

Choose $\mu > 0$ and x^0. Set $\mathcal{I} = \emptyset$.
Set $\alpha_\rho > 0$, $\sigma, \eta \in (0,1)$, $\delta, \gamma > 1$, $0 < \lambda_m < \lambda_M < \infty$, $\Gamma > 1$, $C_0 = \psi_\mu(x^0)$, $Q_0 = 1$, $k = 0$.
while *not converge* **do**
  NMLS Compute $\lambda^k \in [\lambda_m, \lambda_M]$ and $d^k = x^{k+1} - x^k$, where $x^{k+1} = \mathcal{S}(x^k - \lambda^k g^k, \mu\lambda^k)$.
  Select α_k satisfying the Armijo conditions (3.1).
  Set $x^{k+1} = x^k + \alpha_k d^k$, $Q_{k+1} = \eta Q_k + 1$, and $C_{k+1} = (\eta Q_k C_k + \psi_\mu(x^{k+1}))/Q_{k+1}$.
  Sub *Do subspace optimization:* set `do_sub` $= 0$.
  if $\mathcal{I}(x^{k+1})$ *is not equal to* $\mathcal{I}$ **then**
    if (3.4) *is satisfied* **then** set `do_sub` $= 1$ and $\delta = \gamma\delta$.
    else if (3.5) *is satisfied* **then** set `do_sub` $= 1$.
  if `do_sub = 1` **then**
    Solve the subproblem (3.6) using at most Γ iterations to obtain a solution x^{k+2}.
    Set $\mathcal{I} = \mathcal{I}(x^{k+1})$, $C_{k+2} = \psi_\mu(x^{k+2})$, $Q_{k+2} = 1$ and $k := k + 2$.
  else set $k := k + 1$.

REMARK 3.2. *The parameter λ^k is chosen by the Barzilai-Borwein method [1]:*

$$\lambda^k = \max\left\{\lambda_m, \min\left\{\frac{(s^{k-1})^\top s^{k-1}}{(s^{k-1})^\top y^{k-1}}, \lambda_M\right\}\right\} \text{ or } \lambda^k = \max\left\{\lambda_m, \min\left\{\frac{(s^{k-1})^\top y^{k-1}}{(y^{k-1})^\top y^{k-1}}, \lambda_M\right\}\right\}, \tag{3.8}$$

where $s^{k-1} = x^k - x^{k-1}$, $y^{k-1} = g^k - g^{k-1}$ and $0 < \lambda_m < \lambda_M < \infty$.

REMARK 3.3. *In [18], instead of solving problem (1.1) with a given μ, a continuation procedure is*

implemented to solve a sequence of problems $x^*_{\mu_k} := \arg\min_{x\in\mathbb{R}^n} \psi_{\mu_k}(x)$*, one by one, where* $\mu_0 > \mu_1 > \cdots > \mu$ *and* μ_k *goes to* μ*. In this procedure, the solution (or approximate solution)* $x^*_{\mu_{k-1}}$ *is fed into the next problem as an initial solution. Numerical experiments have shown that continuation can significantly improve the overall efficiency. The convergence of the continuation procedure is a simple extension of the convergence of* $\min_{x\in\mathbb{R}^n} \psi_\mu(x)$*.*

REMARK 3.4. *The interested reader is referred to [18] for the efficiency of our algorithm and its ability to identify the optimal active set in practice.*

4. Convergence Analysis. Our analysis in this section is divided into three parts: the convergence of the shrinkage phase, the identification of the active set and the convergence of the overall algorithm. We should point out that, although most of the results can be adapted from these analysis in [6, 12, 19] for smooth minimization, they seem to be new in terms of ℓ_1-minimization and it is meaningful to present them coherently.

4.1. Convergence results of the shrinkage phase. Let Algorithm NMLS be a special case of Algorithm 1 without the subspace optimization phase, i.e., NMLS consists of shrinkage and the nonmonotone line search using condition (3.1). We study the convergence properties of NMLS by directly extending the results of the nonmonontone line search in [19] for minimizing differentiable functions to the nondifferetiable problem (1.1). We assume:

ASSUMPTION 4.1. *Define the level set* $\mathcal{L} := \{x \in \mathbb{R}^n : \psi_\mu(x) \le \psi_\mu(x^0)\}$*.*

1. $f(x)$ *is bounded from below on* $\mathcal{L}$ *and* $\varpi = \sup_k \|d^k\| < \infty$*.*
2. *If* $\widetilde{\mathcal{L}}$ *is the collection of* $x \in \mathbb{R}^n$ *whose distance to* $\mathcal{L}$ *is at most* ϖ*, then* ∇f *is Lipschitz continuous on* $\widetilde{\mathcal{L}}$*, i.e., there exists a constant* $L > 0$ *such that* $\|g(x) - g(y)\| \le L\|x - y\|$ *for all* $x, y \in \{x \in \mathbb{R}^n : f(x) \le f(x^0)\}$*.*

We first show that every iteration of the NMLS algorithm is well defined; namely, there exists a step size satisfying the Armijo condition (3.1).

LEMMA 4.2. *Suppose that Assumption 4.1 holds. For the sequence* $\{x^k\}$ *generated by Algorithm* NMLS*, we have* $\psi^k_\mu \le C_k \le A_k$ *and* $Q_k \le 1/(1-\eta)$*, where* $A_k = \frac{1}{k+1}\sum_{i=0}^k \psi^k_\mu$*. If* $\Delta^k < 0$ *and* $\psi_\mu(x)$ *is bounded from below, there exists* α_k *satisfying the Armijo condition* (3.1)*.*

Proof. The inequalities $\psi^k_\mu \le C_k \le A_k$ and $Q_k \le 1/(1-\eta)$ follow directly from the proof of Lemma 1.1 and Theorem 2.2 in [19]. From Lemma 5 in [17], the Armijo condition

$$\psi_\mu(x^k + \alpha_k d^k) \le \psi_\mu(x^k) + \sigma\alpha_k\Delta^k. \tag{4.1}$$

is satisfied for any $\sigma \in (0,1)$, whenever $0 \le \alpha_k \le \min\{1, 2(1-\sigma)/(\lambda_M L)$. Since $\psi_\mu(x^k) \le C_k$, it follows that α_k can be chosen to satisfy (3.1) for each k. ∎

The next result provides lower and upper bounds for the step size α_k.

LEMMA 4.3. *Suppose that Assumption 4.1 holds. If the Armijo condition* (3.1) *is satisfied with* $0 < \alpha_\rho \le 1$*, then* $\alpha_\rho \ge \alpha_k \ge \tilde{\alpha} := \min\left\{\alpha_\rho, \frac{2\rho(1-\sigma)}{\lambda_M L}\right\}$*.*

Proof. If the initial step size α_ρ satisfies the Armijo condition (3.1), then $\alpha_k = \alpha_\rho$. Otherwise, since $\psi^k_\mu \le C_k$ and h_k is the smallest integer such that $\alpha_k = \alpha_\rho\rho^{h_k}$ satisfies (3.1), we have

$$\psi_\mu(x^k + \alpha_k d^k/\rho) > C_k + \sigma\alpha_k\Delta^k/\rho \ge \psi^k_\mu + \sigma\alpha_k\Delta^k/\rho. \tag{4.2}$$

From the convexity of the norm $\|\cdot\|_1$ and Assumption 4.1, we obtain for any $\alpha \in (0,1]$ that

$$
\begin{aligned}
(4.3)\qquad & \psi_\mu(x^k+\alpha d^k) - \psi_\mu(x^k) \\
&= \mu\|x^k+\alpha d^k\|_1 - \mu\|x^k\|_1 + f(x^k+\alpha d^k) - f(x^k) \\
&\le \alpha((g^k)^\top d^k + \mu\|x^k+d^k\|_1 - \mu\|x^k\|_1) + \int_0^1 (\nabla f(x^k+t\alpha d^k) - \nabla f(x^k))^\top(\alpha d^k)\,dt \\
&\le \alpha\Delta^k + \alpha^2\frac{L}{2}\|d^k\|_2^2,
\end{aligned}
$$

Since $0 < \alpha^k/\rho < 1$, it follows from (4.2) and (4.3) that

$$
\frac{\alpha_k}{\rho}\sigma\Delta^k \le \frac{\alpha_k}{\rho}\Delta^k + \left(\frac{\alpha_k}{\rho}\right)^2\frac{L}{2}\|d^k\|_2^2,
$$

which, after rearranging terms, together with **P6** of Lemma 2.1, gives us $\alpha_k \ge \frac{2\rho(1-\sigma)}{\lambda_M L}$. ∎

We now prove the global convergence of Algorithm NMLS.

THEOREM 4.4. *Suppose that Assumption 4.1 holds. Then the sequence $\{x^k\}$ generated by Algorithm* NMLS *satisfies* $\lim_{k\to\infty}\|d^k\| = 0$.

Proof. From the Armijo condition (3.1) and Lemma 4.3, we have

$$
\psi_\mu^{k+1} \le C_k - \zeta\|d^k\|_2^2, \tag{4.4}
$$

where $\zeta = \frac{\sigma\tilde{\alpha}}{\lambda_m}$. Then the proof is similar to the proof of Theorem 2.2 in [19]. From the updating rule (3.2) and (4.4), we obtain

$$
C_{k+1} = \frac{\eta Q_k C_k + \psi_\mu(x^{k+1})}{Q_{k+1}} \le \frac{\eta Q_k C_k + C^k - \zeta\|d^k\|_2^2}{Q_{k+1}} = C_k - \frac{\zeta\|d^k\|_2^2}{Q_{k+1}}.
$$

Since $\psi_\mu(x)$ is bounded from below and $\psi_k \le C_k$ for all k, we conclude that C_k is bounded from below. Hence, we obtain $\sum_{k=0}^{\infty}\frac{\|d^k\|_2^2}{Q_{k+1}} < \infty$, which together with the fact $Q_{k+1} \le 1/(1-\eta)$ from Lemma 4.2 implies that $\lim_{k\to\infty}\|d^k\| = 0$.∎

Let X^* be the set of stationary points of (1.1). We now state an assumption for proving the R-linear convergence of Algorithm NMLS instead of requiring the strong convexity assumption on $f(x)$ as in [19].

ASSUMPTION 4.5. *(Assumption 2, [17]) (a) $X^* \ne \emptyset$ and for any υ such that $\min\psi_\mu(x) \le \upsilon$, there exist scalars $\varrho > 0$ and $\epsilon > 0$ such that $dist(x, X^*) \le \varrho\|d^{(1)}(x)\|$, whenever $\psi_\mu(x) \le \upsilon, \|d^{(1)}(x)\| \le \epsilon$.*

(b) There exists a scalar $\delta > 0$ such that $\|x-y\| \ge \delta$ whenever $x, y \in X^*, \psi_\mu(x) \ne \psi_\mu(y)$.

LEMMA 4.6. *(Theorem 4, [17]) Suppose that $X^* \ne \emptyset$. Then Assumption 4.5(a) holds under any of the following conditions:*

1. *f is strongly convex and ∇f is Lipschitz continuous on $\mathbb{R}^n$.*
2. *f is quadratic.*
3. *$f(x) = g(Ex) + q^\top x$ for all $x \in \mathbb{R}^n$, where $E \in \mathbb{R}^{m\times n}$, $q \in \mathbb{R}^n$, and g is a strongly convex differentiable function on $\mathbb{R}^m$ with ∇g Lipschitz continuous on $\mathbb{R}^m$.*

In addition to Lemma 4.6, Assumption (4.5)(a) also holds under the strong second-order optimality sufficient conditions, i.e., for any stationary point $x^* \in X^*$, there exists $\omega > 0$ such that

$$
d^\top\nabla^2 f(x^*)d \ge \omega\|d\|^2, \text{ whenever } d_i = 0 \text{ for all } i \in \mathcal{A}_\pm(x^*). \tag{4.5}
$$

The proof of next lemma is a modification of Lemma 5.4 in [12].

LEMMA 4.7. *Suppose that Assumptions 4.1 holds. If f is twice-continuously differentiable near a stationary point x^* of (1.1) satisfying the strong second-order sufficient optimality conditions (4.5), then there exists $\rho > 0$ such that*

$$\|x - x^*\| \leq \varrho \|d^{(1)}(x)\|, \tag{4.6}$$

for all $x \in \mathcal{B}_\rho(x^*)$, where $\varrho = \sqrt{1 + \left(\frac{(1+L)^2}{0.5\omega}\right)^2}$ and $\mathcal{B}_\rho(x^*)$ is the ball centered at x^* with radius ρ.

Proof. By the continuity of the second derivative of f, it follows from (4.5) that for $\rho > 0$ sufficiently small,

$$(g(x) - g(x^*))^\top (x - x^*) \geq 0.5\omega \|x - x^*\|^2 \tag{4.7}$$

for all $x \in \mathcal{B}_\rho(x^*)$ with $x_i = 0$ for all $i \in \mathcal{A}_\pm(x^*)$. Choose ρ small enough if necessary so that

$$|x_i - g_i| \leq \mu \text{ and } |g_i| < \mu \text{ for all } i \in \mathcal{A}_\pm(x^*) \text{ and } x \in \mathcal{B}_\rho(x^*). \tag{4.8}$$

Define $\bar{x}$ as $\bar{x}_i = 0$ if $i \in \mathcal{A}_\pm(x^*)$, otherwise, $\bar{x}_i = x_i$. From (4.8), we obtain

$$\|x - \bar{x}\| \leq \|d^{(1)}(x)\| \tag{4.9}$$

for all $x \in \mathcal{B}_\rho(x^*)$ and

$$\mathcal{S}_i(\bar{x} - g, \mu) - \bar{x}_i = 0, \text{ and } d^{(1)}(x)_i = \mathcal{S}_i(x - g, \mu) - x_i = -x_i$$

for all $i \in \mathcal{A}_\pm(x^*)$, while

$$\mathcal{S}_i(\bar{x} - g, \mu) - \bar{x}_i = d^{(1)}(x)_i = \mathcal{S}_i(x - g, \mu) - x_i$$

for $i \notin \mathcal{A}_\pm(x^*)$. Hence, we obtain

$$\|\mathcal{S}(\bar{x} - g, \mu) - \bar{x}\| \leq \|d^{(1)}(x)\| \tag{4.10}$$

for all $x \in \mathcal{B}_\rho(x^*)$. From the Lipschitz continuity of g, (4.9), (4.10) and **P3** of Lemma 2.1, we have

$$\begin{aligned} \|d^{(1)}(\bar{x})\| &= \|\mathcal{S}(\bar{x} - \bar{g}, \mu) - \mathcal{S}(\bar{x} - g, \mu) + \mathcal{S}(\bar{x} - g, \mu) - \bar{x}\| \\ &\leq L\|\bar{x} - x\| + \|d^{(1)}(x)\| \\ &\leq (1 + L)\|d^{(1)}(x)\| \end{aligned} \tag{4.11}$$

for all $x \in \mathcal{B}_\rho(x^*)$. Therefore, we obtain from **P8** of Lemma 2.1, (4.7) and (4.11) that

$$\|\bar{x} - x^*\| \leq \left(\frac{1 + L}{0.5\omega}\right) \|d^{(1)}(\bar{x})\| \leq \left(\frac{(1 + L)^2}{0.5\omega}\right) \|d^{(1)}(x)\|.$$

Since $\|x - \bar{x}\|^2 + \|\bar{x} - x^*\|^2 = \|x - x^*\|^2$, the inequality (4.6) is proved by squaring (4.9) and (4.11). ∎

We next present a relationship between $\psi_\mu(x^{k+1})$ and Δ^k with respect to the objective function value

at a stationary point. The proof here is adapted from the proof of Theorem 5.2 in [17] and it uses the relationship between $\|d^{(\lambda)}(x)\|$ and $\|d^{(1)}(x)\|$ derived from **P4** and **P5** of Lemma 2.1 as

$$\min(1, \lambda_m)\|d^{(1)}(x)\| \le \|d^{(\lambda)}(x)\| \le \max(1, \lambda_M)\|d^{(1)}(x)\|, \tag{4.12}$$

$$\min(1, \frac{1}{\lambda_M})\|d^{(\lambda)}(x)\| \le \|d^{(1)}(x)\| \le \max(1, \frac{1}{\lambda_m}\|d^{\lambda)}(x)\|, \tag{4.13}$$

for any λ such that $\lambda_m \le \lambda \le \lambda_M$.

LEMMA 4.8. *Suppose that Assumptions 4.1 and 4.5 hold. Then there exist ϑ, $\beta > 0$ and $\hat{k} > 0$ such that the sequence $\{x^k\}$ generated by Algorithm* NMLS *satisfies*

$$\psi_\mu(x^{k+1}) - \vartheta \le -\beta\Delta^k, \quad \forall k \ge \hat{k}. \tag{4.14}$$

Proof. Theorem 4.4 gives that $\lim_{k\to\infty} \|d^{(\lambda^k)}(x^k)\| = 0$. Hence, $\lim_{k\to\infty} \|d^{(1)}(x^k)\| = 0$, i.e., $\|d^{(1)}(x^k)\| \le \epsilon$ for $k > \bar{k}$ for $\bar{k}$ large enough, from (4.13). Since $\psi_\mu(x^{k+1}) \le C_k$ and C_{k+1} is a convex combination of C_k and $\psi_\mu(x^{k+1})$, we have $C_{k+1} \le C_k$ and

$$\psi_\mu(x^{k+1}) \le C_k \le C_{k-1} \le \cdots \le C_0 = \psi_\mu(x^0). \tag{4.15}$$

By Assumption 4.5(a), we have

$$\|x^k - \bar{x}^k\| \le \varrho\|d^{(1)}(x^k)\|, \quad \forall k \ge \bar{k}, \tag{4.16}$$

where $\varrho > 0$ and $\bar{x}^k \in X^*$ satisfies $\|x^k - \bar{x}^k\| = dist(x, X^*)$. Since $\lim_{k\to\infty} \|d^{(1)}(x^k)\| = 0$, we have $\lim_{k\to\infty} \|x^k - \bar{x}^k\| = 0$. Then it follows from $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$ and Assumption 4.5(b) that $\{\bar{x}^k\}$ eventually settles down at some isocost surface of $\psi_\mu(x)$, i.e., there exist an index $\hat{k} \ge \bar{k}$ and a scalar ϑ such that $\psi_\mu(\bar{x}^k) = \vartheta$ for all $k \ge \hat{k}$. By Theorem 2 in [17], $\liminf_{k\to\infty}\psi_\mu(x^k) \ge \vartheta$.

Using the mean value theorem with $\tilde{x}^k$ a point lying on the segment joining x^{k+1} with $\bar{x}^k$, we obtain

$$\begin{aligned}\psi_\mu(x^{k+1}) - \vartheta &= \nabla f(\tilde{x}^k)^\top(x^{k+1} - \bar{x}^k) + \mu\|x^{k+1}\|_1 - \mu\|\bar{x}^k\|_1 \\ &= (\nabla f(\tilde{x}^k) - g^k)^\top(x^{k+1} - \bar{x}^k) - \frac{1}{\lambda^k}(d^k)^\top(x^{k+1} - \bar{x}^k) \\ &\quad +(g^k + \frac{1}{\lambda^k}d^k)^\top(x^{k+1} - \bar{x}^k) + \mu\|x^{k+1}\|_1 - \mu\|\bar{x}^k\|_1.\end{aligned} \tag{4.17}$$

Since $\tilde{\alpha} \le \alpha^k \le \alpha_\rho \le 1$, Lemma 5 (a) in [17] implies that

$$(g^k + \frac{1}{\lambda^k}d^k)^\top(x^{k+1} - \bar{x}^k) + \mu\|x^{k+1}\|_1 - \mu\|\bar{x}^k\|_1 \le (\alpha^k - 1)\left(\frac{1}{\lambda^k}\|d^k\|^2 + \Delta^k\right) \le -(1 - \tilde{\alpha})\Delta^k,$$

which together with Lipschitz continuity of $g(x)$ and Cauchy-Schwartz inequality, yields from (4.17) that

$$\psi_\mu(x^{k+1}) - \vartheta \le L\|\tilde{x}^k - \bar{x}^k\|\,\|x^{k+1} - \bar{x}^k\| + \frac{1}{\lambda^k}\|d^k\|\,\|x^{k+1} - \bar{x}^k\| - (1 - \tilde{\alpha})\Delta^k \tag{4.18}$$

From (4.16), we obtain, for $k \ge \hat{k}$,

$$\|\tilde{x}^k - x^k\| \le \|x^{k+1} - x^k\| + \|x^k - \bar{x}^k\| \le \alpha^k\|d^k\| + \varrho\|d^{(1)}\| \le \delta_1\|d^k\|.$$

where $\delta_1 = \alpha_\rho + \varrho \min(1, 1/\lambda_m)$. Similarly, we have $\|x^{k+1} - \bar{x}^k\| \le \delta_1 \|d^k\|$ and $\|x^{k+1} - x^k\| \le \alpha_\rho \|d^k\|$. Therefore, we obtain from (4.18) that

$$\psi_\mu(x^{k+1}) - \vartheta \le \left(L\delta_1^2 + \frac{\delta_1}{\lambda_m} \right) \|d^k\|^2 - (1 - \tilde{\alpha})\Delta^k,$$

which gives (4.14) with $\beta = \left(L\delta_1^2 + \frac{\delta_1}{\lambda_m} \right) \lambda_M + (1 - \tilde{\alpha})$. □

Finally, we are able to establish the R-linear convergence result.

THEOREM 4.9. *(R-linear convergence) Suppose that Assumptions 4.1 and 4.5 hold. Then there exist ϑ, $0 < \tau < 1$ and $\hat{k} > 0$ such that the sequence $\{x^k\}$ generated by algorithm* NMLS *satisfies*

$$\psi_\mu(x^k) - \vartheta \le \tau^{k-\hat{k}}((\psi_\mu(x^0) - \vartheta), \tag{4.19}$$

for each k. Moreover, the sequence $\{x^k\}$ converges at least R-linearly.

Proof. 1) We claim that for all $k > \hat{k}$

$$\psi_\mu(x^{k+1}) - \vartheta \le (1 - \sigma\tilde{\alpha} b_2)(C_k - \vartheta), \tag{4.20}$$

where $b_2 = 1/(\beta + \sigma\tilde{\alpha})$. Suppose $-\Delta^k \ge b_2(C_k - \vartheta)$, then the Armijo condition (3.1) gives

$$\begin{aligned} \psi_\mu(x^{k+1}) - \vartheta &= (C_k - \vartheta) + (\psi_\mu(x^{k+1}) - C_k) \le (C_k - \vartheta) + \sigma\tilde{\alpha}\Delta^k \\ &\le (1 - \sigma\tilde{\alpha} b_2)(C_k - \vartheta). \end{aligned}$$

Suppose $-\Delta^k < b_2(C_k - \vartheta)$. From (4.14) in Lemma 4.8, we obtain, for all $k > \hat{k}$, that

$$\psi_\mu(x^{k+1}) - \vartheta \le -\beta\Delta^k \le \beta b_2(C_k - \vartheta) = (1 - \sigma\tilde{\alpha} b_2)(C_k - \vartheta).$$

2) We now show that for each $k > \hat{k}$,

$$C_{k+1} - \vartheta \le \tau(C_k - \vartheta), \tag{4.21}$$

where $0 < \tau = 1 - (1 - \eta)\sigma\tilde{\alpha} b_2 < 1$. The proof is similar to Theorem 3.1 of [19]. From the update formula (3.2), we have

$$C_{k+1} - \vartheta = \frac{\eta Q_k (C_k - \vartheta) + (\psi_\mu(x^{k+1}) - \vartheta)}{1 + \eta Q_k}. \tag{4.22}$$

Hence, using (4.20) in (4.22) and using $Q_{k+1} \le 1/(1 - \eta)$ yields

$$\begin{aligned} C_{k+1} - \vartheta &\le \frac{(\eta Q_k + 1 - \sigma\tilde{\alpha} b_2)(C_k - \vartheta)}{1 + \eta Q_k} = \left(1 - \frac{\sigma\tilde{\alpha} b_2}{Q_{k+1}} \right) (C_k - \vartheta) \\ &\le (1 - (1 - \eta)\sigma\tilde{\alpha} b_2)(C_k - \vartheta), \end{aligned}$$

which proves (4.21). Therefore, we obtain (4.19) from (4.21) by using (4.15).

3) If $f(x)$ is strictly convex, then the R-linear convergence follows from (4.19) immediately. Otherwise,

we prove R-linear convergence as follows. From the defintion of C_{k+1} in (3.2), we have

$$C_k - \psi_\mu^{k+1} = (1+\eta Q_k)(C_k - C_{k+1}) \le \frac{1}{1-\eta}(C_k - C_{k+1}). \tag{4.23}$$

The Armijo condition (3.1) implies that

$$\psi_\mu^{k+1} - C_k \le \sigma\alpha_k\Delta^k \le \sigma\alpha_k\frac{\gamma-1}{\lambda^k}\|d^k\|_2^2 \le \frac{\sigma(\gamma-1)}{\lambda_M\alpha_\rho}\|x^{k+1}-x^k\|_2^2. \tag{4.24}$$

Rearranging terms of (4.24) and using (4.23) yield

$$\|x^{k+1}-x^k\| \le \sqrt{\frac{\lambda_M\alpha_\rho}{(1-\gamma)\sigma}\left(C_k - \psi_\mu^{k+1}\right)} \le \sqrt{\frac{\lambda_M\alpha_\rho}{(1-\gamma)(1-\eta_{\max})\sigma}\left(C_k - C_{k+1}\right)},$$

which implies that $\{x^k\}$ converges at least R-linearly since $\{C_k - C_{k+1}\}$ converges at least R-linearly because of (4.21). ∎

4.2. Identification of the active set by the shrinkage phase. We now discuss the identification of the active set by Algorithm NMLS. First, we state an elementary result in the lemma below.

LEMMA 4.10. *Assume that the sequence $\{x^k\}$ converges to x^*, i.e., $\lim_{k\to\infty} x^k = x^*$. Then, for any $\xi > 0$, there exists a finite number $\bar{k} > 0$ so that for all $k > \bar{k}$, 1) $\operatorname{sgn}(x_i^k) = \operatorname{sgn}(x_i^*)$ for all $i \in \mathcal{I}(x^*)$; 2. $|x_i^k - x_i^*| < \xi$ for $i \in \mathcal{I}(x^*)$ and $|x_i^k| < \xi$ for all $i \in \mathcal{A}(x^*)$.*

Proof. Suppose that the first part does not hold, there exists a $i \in \mathcal{I}(x^*)$ and a subsequence $\{x^k\}$, such that for k large enough $\operatorname{sgn}(x_i^k) \ne \operatorname{sgn}(x_i^*)$ which implies that $||x^k - x^*||_1 \ge |x_i^k - x_i^*| \ge |x_i^*|$, which is contradiction. The second part can be proved similarly. ∎

The next result uses the generalized directional derivative and the generalized gradient of $\psi_\mu(x)$ similar to the projected gradient method for linearly constrained problems in [6]. The generalized directional derivative of a function ψ at x in the direction $d \in \mathbb{R}^n$ is

$$\psi^\diamond(x;d) \equiv \lim_{t\downarrow 0}\sup_{y\to x}\frac{\psi(y+td)-\psi(y)}{t},$$

and it satisfies that

$$\psi^\diamond(x;d) = \max_{p\in\partial\psi(x)}\langle p, d\rangle. \tag{4.25}$$

The genearlized gradient of a function ψ is defined as

$$g_\Omega(x) := \arg\min_{p\in\partial\psi(x)}\|p\|.$$

Then it is straightforward to verify that the generalized gradient of the l_1-regularized function ψ_μ is

$$g_\Omega(x) := \arg\min_{p\in\partial\|x\|_1}\|g(x)+\mu p\|_2 = \begin{cases} g_i(x) + \mu\operatorname{sgn}(x_i), & \text{if } x_i \in \mathcal{I}(x),\\ g_i(x) - \mu\operatorname{sgn}(g_i(x)), & \text{if } x_i \in \mathcal{A}_0(x),\\ 0, & \text{if } x_i \in \mathcal{A}_\pm(x),\end{cases} \tag{4.26}$$

and Lemma 11.1.1 in [8] gives

$$-\|g_\Omega(x)\| = \min_{\|d\|\le 1} \psi^\diamond(x,d). \tag{4.27}$$

Next, we show that the active set of a stationary point can be identified after a finite number of steps under some conditions.

THEOREM 4.11. *Suppose that $f(x)$ is continuously differentiable and let $\{x^k\}$ be an arbitrary sequence converges to x^* such that $\lim_{k\to\infty}\|g_\Omega(x^k)\| = 0$. Then $\mathcal{A}_\pm(x^k) = \mathcal{A}_\pm(x^*)$ for all k sufficiently large.*

Proof. Since x^k converges to x^*, it is clear that $\mathcal{A}_\pm(x^k) \subseteq \mathcal{A}_\pm(x^*)$ for all k sufficiently large. Assume, however, that there is an infinite subsequence K_0 and an index l such that $l \in \mathcal{A}_\pm(x^*)$ but $x_l^k > 0$ (If not, choose a sequence such that $x_l^k < 0$), without loss of generality, for all $k \in K_0$. Let $d = e_l$, where e_l is the vector whose lth element equals to one and all other elements equal to zero. We obtain from (4.27) that

$$g_l^k + \mu = \max_{p\in\partial\|x^k\|_1} \langle g^k + \mu p, d\rangle \le \max_{\|d\|\le 1} \psi_\mu^\diamond(x^k,d) = \|g_\Omega(x^k)\| \tag{4.28}$$

Therefore, since $\lim_{k\to\infty}\|g_\Omega(x^k)\| = 0$, we obtain

$$g_l^* + \mu \le 0,$$

which contradicts the fact that $|g_l^*| < \mu$ as $l \in \mathcal{A}(x^*)$. ∎

Denote by $x^{k+} := \mathcal{S}(x^k - \lambda^k g^k, \mu\lambda^k)$ the point generated by shrinkage at x^k. We show that the generalized gradient at x^{k+} converges to zero similar to Theorem 3.2 in [6].

THEOREM 4.12. *Suppose Assumption 4.1 holds. Then if $\nabla f(x)$ is uniformly continuous in $\widetilde{\mathcal{L}}$, the sequence $\{x^k\}$ generated by Algorithm* NMLS *satisfies that*

$$\lim_{k\to\infty}\|g_\Omega(x^{k+})\| = 0.$$

Proof. Let $\epsilon > 0$ be given and choose a direction v^{k+} with $\|v^{k+}\| \le 1$ such that

$$\|g_\Omega(x^{k+})\| \le -\psi_\mu^\diamond(x^{k+}, v^{k+}) + \epsilon.$$

Using **P1** of Lemma 2.1 for any $z^{k+} = x^{k+} + t_{k+}v^{k+}$ and $t_{k+} > 0$, we have

$$(x^{k+} - x^k + \lambda^k g^k)^\top (z^{k+} - x^{k+}) + \mu\lambda^k(\|z^{k+}\|_1 - \|x^{k+}\|_1) \ge 0. \tag{4.29}$$

Taking t_{k+} small enough so that $\mathrm{sgn}(z_i^{k+}) = \mathrm{sgn}(x_i^{k+})$ for any $i \in \mathcal{I}(x^{k+})$ yields

$$\begin{aligned} |z_i^{k+}| - |x_i^{k+}| &= t_{k+}\begin{cases} \mathrm{sgn}(x_i^{k+})v_i^{k+} & \text{if } i \in \mathcal{I}(x^{k+}),\\ |v_i^{k+}| & \text{if } i \in \mathcal{A}(x^{k+}).\end{cases} \\ &\le \max_{p\in\partial|x_i^{k+}|} p_i v_i^{k+} \end{aligned} \tag{4.30}$$

Rearranging terms of (4.29) and using (4.30) give

$$t_{k+}(x^{k+} - x^k)^\top v^{k+} \ge -t_{k+}\lambda^k(g^k - g^{k+})^\top v^{k+} - t_{k+}\lambda^k \max_{p\in\partial\|x^{k+}\|_1} \langle g^{k+} + \mu p, v^{k+}\rangle. \tag{4.31}$$

Using (4.31), (4.25) and Cauchy-Schwartz inequality, we obtain

$$-\lambda^k \psi_\mu^\diamond(x^{k+}, v^{k+}) \le \|x^{k+} - x^k\| \, \|v^{k+}\| + \lambda^k \|g^k - g^{k+}\| \, \|v^{k+}\|.$$

Since λ^k is bounded, $\|x^{k+} - x^k\|$ converges to zero from Theorem 4.4 and ∇f is uniformly continuous, we conclude that

$$\limsup_{k\to\infty} -\psi_\mu^\diamond(x^{k+}, v^{k+}) \le 0,$$

which implies that

$$\limsup_{k\to\infty} \|g_\Omega(x^{k+})\| \le \epsilon.$$

Since ϵ is arbitrary, the proof is completed. ◻

The assumption that $g(x)$ is uniformly continuous in $\widetilde{\mathcal{L}}$ in Theorem 4.12 can be relaxed to the assumption that the sequence $\{x^k\}$ is bounded. The problem (1.1) is said to be *degenerate* at x^* if there exist some i such that $|g_i^*| = \mu$. Theorems 4.11 and 4.12 yield $\mathcal{A}_\pm(x^{k+}) = \mathcal{A}_\pm(x^*)$ for k sufficiently large. We have not established $\lim_{k\to\infty} \|g_\Omega(x^k)\| = 0$ for the sequence $\{x^k\}$ since we cannot show the inequality (4.30). However, we can still show that x_i^k, $i \in \mathcal{A}_\pm(x^*)$, converges to zero at least q-linearly.

Corollary 4.13. *Suppose that Assumption 4.1 holds and the sequence $\{x^k\}$ is generated by Algorithm* NMLS. *If $\nabla f(x)$ is uniformly continuous in $\widetilde{\mathcal{L}}$, then x_i^k, $i \in \mathcal{A}_\pm(x^*)$, converges to zero either after a finite number of steps or at least q-linearly.*

Proof. 1) Assume that $\mathcal{A}_\pm(x^*)$ is nonempty. Since $f(x)$ is continuously differentiable, there exists a $\gamma > 0$ with the property that for all $x \in \mathcal{B}_\gamma(x^*)$ so that $|g_i(x)| < \mu$ if $i \in \mathcal{A}_\pm(x^*)$. Let k_+ be chosen large enough that $x_k \in \mathcal{B}_\gamma(x^*)$ for all $k > k_+$. Suppose that there exists $x_l^k = 0$ for $l \in \mathcal{A}_\pm(x^*)$ and $k \ge k_+$. Then the shrinkage gives $\mathcal{S}_l(x^k - \lambda^k g^k, \mu\lambda^k) = 0$ since $|x_l^k - \lambda^k g_l^k| = \lambda^k |g_l^k| < \mu\lambda^k$. Hence, $d_l^k = 0$. Consequently, when an index $l \in \mathcal{A}_\pm(x^*)$ becomes active, i.e., $x_l^k = 0$, at iterate x^k, $k \ge k_+$, it remains active for all the subsequent iterations.

2) We now focus on the nontrivial indices in $\mathcal{A}_\pm(x^*)$, i.e., there exists $l \in \mathcal{A}(x^*)$ and $x_l^k \ne 0$ for all $k \ge k_+$. Let ξ sufficiently small. There exists $\bar{k}$ sufficiently large so that $|x_i^{\bar{k}}| < \xi$ for $i \in \mathcal{A}(x^*)$ from Lemma 4.10 and $\mathcal{A}_\pm(x^{k+}) = \mathcal{A}_\pm(x^*)$ from Theorems 4.12 and 4.11 for all $k \ge \bar{k}$. Since $\tilde{\alpha} < \alpha_k < \alpha_\rho \le 1$ and $x_i^{k+1} = (1-\alpha_k)x^k + \alpha_k x^{k+}$, we obtain

$$|x_i^k| \le (1-\tilde{\alpha})^{k-\bar{k}} \xi$$

for any $i \in \mathcal{A}_\pm(x^*)$ and all $k \ge \bar{k}$. ◻

The efficiency of our active set algorithm depends on how fast the iterative shrinkage scheme can identify the correct support. Since the true zero components can be nonzero after a lot of iterations in practice and the size of the support $\mathcal{I}(x^k)$ decides the size of subspace optimization, we can use the identification function proposed in [10] for general nonlinear programming to identify an approximate support.

Definition 4.14. *A continuous function $\rho(x) : \mathbb{R}^n \to \mathbb{R}_+$ is called an identification function for x^* with respect to a sequence $\{x^k\}$ if $\rho(x^*) = 0$ and*

$$\lim_{x^k \to x^*, x \ne x^*} \frac{\rho(x^k)}{\|x^k - x^*\|} = +\infty. \tag{4.32}$$

Therefore, the active set $\mathcal{A}(x^k)$ and the support $\mathcal{I}(x^k)$ can be replaced approximately by sets

$$\mathcal{A}_\rho(x^k) := \{i \in [1,n] \,|\, |x_i^k| \le \rho(x^k)\}, \quad \mathcal{I}_\rho(x^k) := \{i \in [1,n] \,|\, |x_i^k| > \rho(x^k)\}. \tag{4.33}$$

We further divide the approximate active indices $\mathcal{A}_\rho(x^k)$ into two sets

$$\mathcal{A}_{\rho,\pm}(x^k) := \{i \in \mathcal{A}_\rho(x) \,|\, ||g_i^k| - \mu| \ge \rho(x^k)\}, \quad \mathcal{A}_{\rho,0}(x^k) := \{i \in \mathcal{A}_\rho(x^k) \,|\, ||g_i^k| - \mu| \le \rho(x^k)\}. \tag{4.34}$$

LEMMA 4.15. Let ρ be an identification function for x^* and the sequence $\{x^k\}$ converges to x^*. Then $\mathcal{A}_\rho(x^k) = A(x^*)$ and $\mathcal{A}_{\rho,0}(x^k) = \mathcal{A}_0(x^*)$ for k sufficiently large.

Proof. 1) If $i \in \mathcal{A}(x^*)$, we have

$$|x_i^k| \le \|x^k - x^*\|_2 \le \rho(x^k), \tag{4.35}$$

for k sufficiently large, so that, by (4.33), $i \in \mathcal{A}_\rho(x^k)$. On the other hand, there exists k sufficiently large so that $\rho(x^k) < \min\{|x_i^*|, i \in \mathcal{I}(x^*)\}$ and $||x_i^k| - \frac{1}{2}\rho(x^k)| > \frac{1}{2}\rho(x^k)$ for $i \in \mathcal{I}(x^*)$. Hence, if $|x_i^*| > 0$, we have $i \notin \mathcal{A}_\rho(x^k)$.

2) If $i \notin \mathcal{A}_\pm(x^*)$, then $|g_i(x^*)| = \mu$. From the Lipschitz continuity of $g(x)$, we have

$$||g_i(x^k)| - \mu| \le ||g_i(x^k)| - |g_i(x^*)|| \le |g_i(x^k) - g_i(x^*)| \le L\|x^k - x^*\| \le \rho(x^k),$$

for k sufficiently large. On the other hand, if $i \in \mathcal{A}_\pm(x^*)$, then $|g_i^*| < \mu$, hence, by continuity of $g(x)$, $i \in \mathcal{A}_{\rho,\pm}(x^k)$. Therefore, these facts together with 1) yield $\mathcal{A}_{\rho,0}(x^k) = \mathcal{A}_0(x^*)$ for k sufficiently large. ∎

Using the strong second-order sufficient optimality conditions (4.5) of problem (1.1), we will show that the function

$$\rho_1(x) := \sqrt{\|(|g| - \mu)_{\mathcal{I}(x)\cup\mathcal{A}_0(x)}\|}, \tag{4.36}$$

is an identification function for the sequence $\{x^k\}$ generated by Algorithm NMLS.

LEMMA 4.16. *Suppose Assumption 4.1 holds. If the sequence $\{x^k\}$ is generated by Algorithm* NMLS, *then for each $i \in \mathcal{A}_\pm(x^*)$ we have*

$$\limsup_{k\to\infty} \frac{|x_i^k|}{\|x^k - x^*\|^2} < \infty. \tag{4.37}$$

Proof. From the proof of Corrollary 4.13, an index $l \in \mathcal{A}_\pm(x^*)$ of x^k remains active for all the subsequent iterations once it becomes active for k sufficiently large. Hence, the inequality 4.37 holds in this case. Now, let us focus on the nontrivial indices in $\mathcal{A}_\pm(x^*)$. That is, suppose that there exists $l \in \mathcal{A}(x^*)$ and $x_l^k \ne 0$ for all $k \ge k_+$ and we assume $x_l^k > 0$ without loss of generality. If (4.37) does not hold, we can choose k sufficiently large, if necessary, so that

$$\frac{|x_l^k|}{\|x^k - x^*\|^2} \ge \frac{L(2 + \lambda_M L)^2}{2(1-\sigma)|g_l^k + \mu|}, \tag{4.38}$$

and

$$|x_l^k - \lambda^k g_l^k| < \mu\lambda^k, \tag{4.39}$$

for $\lambda^k \geq \lambda_m > 0$. since x^k converges to x^*, $x_l^* = 0$ and $|g_l(x^*)| < \mu$. Hence, we obtain from (4.39) that

$$d_l^k = \mathcal{S}_l(x_l^k - \lambda^k g_l^k, \mu\lambda^k) - x^k = -x_l^k. \tag{4.40}$$

Since the shrinkage operator is component separable, **P6** of Lemma 2.1 holds for each component, i.e.,

$$\Delta_i^k := g_i d_i + \mu(|x_i + d_i| - |x_i|) \leq -\frac{1}{\lambda}|d_i|^2. \tag{4.41}$$

Specifically, we have from (4.40) that

$$\Delta_l^k = -(g_l^k + \mu)x_l^k. \tag{4.42}$$

We now prove that $\alpha_k = 1$ is an acceptable step size. Using the Lipschitz continuity of $g(x)$, (4.41) and (4.42), we have

$$\begin{aligned} &\psi_\mu(x^k + d^k) - \psi_\mu(x^k) \\ &\leq ((g^k)^\top d^k + \mu\|x^k + d^k\|_1 - \mu\|x^k\|_1) + \int_0^1 (\nabla f(x^k + td^k) - \nabla f(x^k))^\top (d^k)\, dt \\ &\leq \sigma\Delta^k + (1-\sigma)\Delta^k + \frac{L}{2}\|d^k\|_2^2 \leq \sigma\Delta^k + (1-\sigma)\Delta_l^k + \frac{L}{2}\|d^k\|_2^2 \\ &= \sigma\Delta^k - (1-\sigma)(g_l^k + \mu)\, x_l^k + \frac{L}{2}\|d^k\|_2^2 \end{aligned} \tag{4.43}$$

From **P3**, **P7** of Lemma 2.1 and the Lipschitz continuity of ∇f, we obtain

$$\begin{aligned} \|d^{(\lambda^k)}(x^k)\| &= \|d^{(\lambda^k)}(x^k) - d^{(\lambda^k)}(x^*)\| \\ &= \|S(x^k - \lambda^k g^k, \mu\lambda^k) - x^k - (S(x^* - \lambda^k g^*, \mu\lambda^k) - x^*)\| \\ &\leq \|x^k - x^*\| + \|x^k - \lambda^k g^k - (x^* - \lambda^k g^*)\| \\ &\leq 2\|x^k - x^*\| + \lambda^k\|g^k - g^*\| \\ &\leq (2 + \lambda_M L)\|x^k - x^*\| \end{aligned}$$

Combining the upper bounds for $\|d^k\|$ and the lower bound for $|x_l^k|$ yields

$$\begin{aligned} \frac{L}{2}\|d^k\|_2^2 &\leq \frac{L}{2}(2 + \lambda_M L)^2\|x^k - x^*\|^2 \\ &\leq \left(\frac{(1-\sigma)x_l^k\,(g_l^k + \mu)}{\|x^k - x^*\|^2}\right)\|x^k - x^*\|^2 \\ &= (1-\sigma)x_l^k\,(g_l^k + \mu), \end{aligned}$$

which together with (4.43) implies that

$$\psi_\mu(x^k + d^k) \leq \psi_\mu(x^k) + \sigma\Delta^k \leq C^k + \sigma\Delta^k.$$

Hence, $\alpha_k = 1$ is an acceptable step size and $x_l^{k+1} = x_l^k + d_l^k$. Therefore, using (4.39), we obtain $x_l^{k+1} = 0$ which contradicts the fact that $x_l^k > 0$ for all $k \geq k_+$. ∎

LEMMA 4.17. *Suppose Assumption 4.1 holds and the sequence $\{x^k\}$ is generated by Algorithm* NMLS.

If x^* satisfies the strong second-order sufficient optimality conditions, then there exists $\kappa^* > 0$ such that

$$\||g(x^k)| - \mu)_{\mathcal{I}(x^k)\cup\mathcal{A}_0(x^k)}\| \geq \kappa^* \|d^1(x^k)\| \tag{4.44}$$

for k sufficiently large.

Proof. Define $\bar{x}$ as $\bar{x}_i = 0$ if $i \in \mathcal{A}_\pm(x^*)$, otherwise, $\bar{x}_i = x_i$. Choose $\gamma > 0$, so that, for $x^k \in \mathcal{B}_\gamma(x^*)$, we have

$$\begin{aligned} \|d^{(1)}(x^k)\| &\leq \|d^{(1)}(x^k) - d^{(1)}(x^*)\| \\ &\leq \|d^{(1)}(x^k) - d^{(1)}(\bar{x}^k)\| + \|d^{(1)}(\bar{x}^k) - d^{(1)}(x^*)\| \\ &\leq (2+L)(\|x^k - \bar{x}^k\| + \|\bar{x}^k - x^*\|). \end{aligned} \tag{4.45}$$

From Lemma 4.16, there exists a constant ξ such that

$$\limsup_{k\to\infty} \frac{|x_{ki}|}{\|x^k - x^*\|^2} \leq \xi < \infty, \tag{4.46}$$

which implies, for k sufficiently large, that

$$\|\bar{x}^k - x^k\| \leq \sum_{i\in\mathcal{A}_\pm(x^*)} |x_i^k| \leq n\xi\|x^k - x^*\|^2 \leq n\xi\|x^k - x^*\|(\|x^k - \bar{x}^k\| + \|\bar{x}^k - x^*\|). \tag{4.47}$$

Hence, for any $\epsilon > 0$ and k sufficiently large, (4.47) yields

$$\|\bar{x}^k - x^k\| \leq \epsilon\|\bar{x}^k - x^*\|, \tag{4.48}$$

since x^k converges to x^*. Combining (4.45) and (4.48), there exists a constant $c > 0$ such that

$$\|d^{(1)}(x^k)\| \leq c\|\bar{x}^k - x^*\| \tag{4.49}$$

for k sufficiently large.

Let k be chosen large enough that

$$\|x^k - x^*\| < \min\{|x_i^*|, i \in \mathcal{I}(x^*)\}. \tag{4.50}$$

Suppose, in this case, that $i \in \mathcal{A}(x^k)$. If $|x_i^*| > 0$, then $\|x^k - x^*\| \geq |x_i^*|$, which contradicts (4.50). Hence $\bar{x}_i^k = x_i^* = 0$. Moreover, if $i \in \mathcal{A}_\pm(x^*)$, then by the definition of $\bar{x}^k$, $\bar{x}_i^k = x_i^* = 0$. In summary,

$$\begin{cases} \bar{x}_i^k = x_i^* = 0, & \text{for each } i \in \mathcal{A}(x^k) \cup \mathcal{A}_\pm(x^*), \\ |g_i(x^*)| = \mu, & \text{for each } i \in \mathcal{A}_\pm(x^*)^c \end{cases} \tag{4.51}$$

where $\mathcal{A}_\pm(x^*)^c$ is the complement of $\mathcal{A}_\pm(x^*)$. Define $\mathcal{Z} = \mathcal{A}(x^k)^c \cap \mathcal{A}_\pm(x^*)^c$.

By the strong second-order sufficient optimality conditions for x near x^*, we have

$$\frac{\omega}{2}\|\bar{x} - x^*\|^2 \leq (\bar{x} - x^*)^\top \int_0^1 \nabla^2 f(x^* + t(\bar{x} - x^*))\, dt(\bar{x} - x^*) = (\bar{x} - x^*)^\top (g(\bar{x}) - g(x^*)). \tag{4.52}$$

We substitute $x = x^k$ in (4.52) and utilize (4.51) to obtain

$$|(\bar{x}^k - x^*)^\top (g(\bar{x}^k) - g(x^*))| \le \sum_{i \in \mathcal{Z}} |(\bar{x}_i^k - x_i^*)| \cdot ||g_i(\bar{x}^k)| - \mu| \le \|\bar{x}^k - x^*\| \, \|(|g(\bar{x}^k)| - \mu)_{\mathcal{I}(x^k) \cup \mathcal{A}_0(x^k)}\| \tag{4.53}$$

for k sufficient large, since $\mathcal{Z} \subseteq \mathcal{A}_\pm(x^*)^c$ and $\mathcal{Z} \subseteq \mathcal{A}(x^k)^c = \mathcal{I}(x^k)$. Hence, we obtain

$$\frac{\omega}{2} \|\bar{x}^k - x^*\| \le \||g(x^k)| - \mu)_{\mathcal{I}(x^k) \cup \mathcal{A}_0(x^k)}\|. \tag{4.54}$$

Combining (4.49) and (4.54), the proof is complete. ◻

THEOREM 4.18. *Suppose Assumption 4.1 holds and the sequence $\{x^k\}$ generated by Algorithm* NMLS *converges x^*. If x^* satisfies the strong second-order sufficient optimality conditions, then $\rho_1(x^k)$ defined by* (4.36) *is an identification function.*

Proof. From Lemmas 4.7 and 4.17, we obtain

$$\lim_{x \to x^*, x \neq x^*} \frac{\rho_1(x)}{\|x - x^*\|} \ge \frac{\sqrt{\kappa^* \|d^1(x^k)\|}}{\varrho \|d^1(x^k)\|} \to +\infty$$

for k sufficiently large. Therefore, ρ_1 is an identification fucntion. ◻

REMARK 4.19. *From the proof of Lemma 4.16, in particular, the proof of the inequality* (4.43)*, the step size $\alpha = 1$ is acceptable if*

$$\lambda \le \frac{2(1-\sigma)}{L}. \tag{4.55}$$

Therefore, Theorems 4.12 and 4.11 yield $\mathcal{A}_\pm(x^k) = \mathcal{A}_\pm(x^*)$ for k sufficiently large.

4.3. Convergence results of the active set algorithm. We now study the subspace optimization stage of Algorithm 1. The justification for test (3.4) and (3.5) is based on the convergence properties of Algorithm NMLS. On the one hand, we want to start subspace optimization as soon as possible; on the other hand, we want the active set that defines the subspace optimization problem to be as accurate as possible. If there is at least one nonzero components of x^*, then $\|g^*_{\mathcal{I}^*}\| \ge \mu$ since $|g_i^*| = \mu$ for $i \in \mathcal{I}^*$ from the optimality conditions. Suppose the sequence $\{x^k\}$ generated by the first stage converges to an optimal solution x^* of (1.1). Then $g(x^k)$ converges $g(x^*)$ and $\|d^{(\lambda^k)}(x^k)\|_2$ converges to zero from **P7** of Lemma 2.1. Hence, the quantity $\lambda^{k-1}\|g^k_{\mathcal{I}(x^k)}\|/\|d^{(\lambda^{k-1})}\|_2$ tends to infinity and the first part of condition (3.4) will be satisfied after a finite number of iterations. However, the quantity $\lambda^{k-1}\|g^k_{\mathcal{I}(x^k)}\|/\|d^{(\lambda^{k-1})}\|_2$ cannot tell us whether the current point x^k is optimal or not. Hence, we also check the second part of condition (3.4) in which $\|(|g(x^k)| - \mu)_{\mathcal{I}(x^k) \cup \mathcal{A}_0(x^k)}\|_\infty$ is a measure of optimality. In fact, if $i \notin \mathcal{A}_\pm(x^*)$, then $|g_i(x^*)| = \mu$. By Lemma 4.7, we have

$$||g_i(x^k)| - \mu| \le ||g_i(x^k)| - |g_i(x^*)|| \le |g_i(x^k) - g_i(x^*)| \le L\|x^k - x^*\| \le L\varrho \|d^{(1)}(x^k)\|$$

Hence, there exists a constant δ such that $\|(|g(x^k)| - \mu)_{\mathcal{I}(x^k) \cup \mathcal{A}_0(x^k)}\|_\infty \le \delta \|d^{(1)}(x^k)\|$ for k sufficiently large. If it happens that the shrinkage phase converges slowly and cannot make sufficient progress after a lot of iterations, the relative change of the objective function value between two consecutive iterations usually will be small. Hence, satisfaction of (3.5) indicates that the Algorithm NMLS is stagnating. Therefore,

Algorithm FPC_AS is well-defined. We now analyze the global convergence of Algorithm FPC_AS similar to Theorem 4.1 in [12].

THEOREM 4.20. *(Global convergence) Suppose Assumption 4.1 holds and Condition 3.1 is satisfied. Then Algorithm FPC_AS either terminates in a finite number of iterations at a stationary point, or we have*

$$\liminf_{k\to\infty}\|d^{(\lambda)}(x^k)\| = 0. \tag{4.56}$$

Proof. Since we terminate each subspace optimization after at most Γ iterations, either only NMLS is performed for large k or NMLS is restarted an infinite number of times. If only NMLS is performed for large k, then (4.56) follows from Theorem 4.4. Suppose that NMLS is restarted an infinite number of times at $k_1 < k_2 < \cdots$ and that it terminates at $k_1 + l_1 < k_2 + l_2 < \cdots$, respectively. Thus $k_i < k_i + l_i \le k_{i+1}$ for each i. If (4.56) does not hold, then there exist ϵ such that $\|d^{(1)}(x^k)\| \ge \epsilon$. It follows from (4.4) and (4.15) that

$$\psi_\mu^{k_i+l_i} \le C^{k_i+l_i-1} - \zeta\|d^k\|^2 \le \psi_\mu(x^{k_i}) - \zeta\epsilon\min(1,\lambda_m). \tag{4.57}$$

From the definition of subspace optimization, we obtain $\varphi_\mu(x^{k_i+l_i}) = \psi_\mu(x^{k_i+l_i})$. Since subspace optimization will not make a zero component in $\mathcal{A}(x^{k_i+l_i})$ nonzero, we obtain $\mathcal{I}(x^{k_{i+1}}) \subseteq \mathcal{I}(x^{k_i+l_i})$ and

$$\varphi_\mu(x^{k_{i+1}}) = \psi_\mu(x^{k_{i+1}}).$$

By Condition 3.1, we have $\varphi_\mu(x^{k_{i+1}}) \le \varphi_\mu(x^{k_i+l_i})$; hence $\psi_\mu(x^{k_{i+1}}) \le \psi_\mu(x^{k_i+l_i})$. This together with (4.57) gives

$$\psi_\mu(x^{k_{i+1}}) \le \psi_\mu(x^{k_i}) - \zeta\epsilon\min(1,\lambda_m),$$

which contradicts the assumption that $\psi_\mu(x)$ is bounded from below. ∎

5. Conclusions. We have presented a two-stage active-set algorithm for the l_1-norm regularized optimization in which the iterative shrinkage scheme is used to estimate the support at the solution and then a subspace optimization problem is solved to recover the magnitudes of the components in the estimated support. The difficulty is to integrate shrinkage and subspace optimization coherently to guarantee convergence. We show the convergence of the first stage algorithm NMLS by noting that shrinkage operator exhibits many characteristics similar to those of the gradient projection for the bounded constrained problem. In particular, NMLS is able to identify of the zero components of a stationary point after a finite number of steps under some mild conditions. The overall convergence of FPC_AS is enforced by decreasing the original objective function after the subspace optimization phase.

Appendix A. Proof of Lemma 2.1 .

Proof. 1) The first-order optimality conditions for a stationary point x^* is

$$\nabla f(x^*)(x - x^*) + \mu(\xi - \|x^*\|_1) \ge 0, \text{ for all } (x,\xi) \in \Omega, \tag{A.1}$$

since $\xi^* = \|x^*\|_1$. Applying (A.1) to problem (2.4) gives **P1**.

2) Replacing y with $\mathcal{S}(y,\nu)$ and ξ with $\|\mathcal{S}(y,\nu)\|_1$ in **P1** gives

$$(\mathcal{S}(x,\nu)-x)^\top(\mathcal{S}(y,\nu)-\mathcal{S}(x,\nu))+\nu(\|\mathcal{S}(y,\nu)\|_1-\|\mathcal{S}(x,\nu)\|_1)\geq 0.$$

A similar inequality is obtained if x and y is exchanged. Adding these two inequalities gives **P2**.

3) **P3** is the nonexpansive property of shrinkage operator given by Lemma 3.2 in [13].

4) **P4** and **P5** are given by Lemma 3 in [17].

5) Replacing x with $x-\lambda g$, y with x, ξ with $\|x\|_1$ and $\nu=\mu\lambda$ in **P1** gives

$$(\mathcal{S}(x-\lambda g,\mu\lambda)-(x-\lambda g))^\top(x-\mathcal{S}(x-\lambda g,\mu\lambda))+\mu\lambda(\|x\|_1-\|\mathcal{S}(x-\lambda g,\mu\lambda)\|_1)\geq 0,$$

which is equivalent to

$$(d^{(\lambda)}(x)+\lambda g)^\top(-d^{(\lambda)}(x))+\mu\lambda(\|x\|_1-\|x^+\|_1)\geq 0,$$

which further gives **P6** after rearranging terms. An alternative proof is given in Lemma 2.1 in [17].

6) If x^* is a stationary point, replacing x with x^* in **P6** and together with the optimality conditions (A.1), we obtain $\|d^{(\lambda)}(x^*)\|=0$. On the contrary, if $\|d^{(\lambda)}(x^*)\|=0$, then $\mathcal{S}(x^*-\lambda g^*,\mu\lambda)=x^*$. Replacing x by $x^*-\lambda g^*$ in **P1**, we obtain

$$(\mathcal{S}(x^*-\lambda g^*,\mu\lambda)-(x^*-\lambda g^*))^\top(y-\mathcal{S}(x^*-\lambda g^*,\mu\lambda))+\mu\lambda(\xi-\|\mathcal{S}(x^*-\lambda g^*,\mu\lambda)\|_1)\geq 0,$$

which gives (A.1). An alternative proof is given in Lemma 1 in [17].

7) Replacing x with $x-\lambda g(x)$ and replacing y with x^* in **P1** gives

$$(\mathcal{S}(x-\lambda g,\mu\lambda)-(x-\lambda g))^\top(x^*-\mathcal{S}(x-\lambda g,\mu\lambda))+\mu\lambda(\|x^*\|_1-\|\mathcal{S}(x-\lambda g,\mu\lambda)\|_1)\geq 0,$$

which is equivalent to

$$(d^{(\lambda)}(x)+\lambda g)^\top(x^*-x^+)+\mu\lambda(\|x^*\|_1-\|x^+\|_1)\geq 0. \tag{A.2}$$

Since $(x^+,\|x^+\|_1)\in\Omega$, the optimality conditions (A.1) gives

$$\lambda(g^*)^\top(x^+-x^*)\geq\mu\lambda(\|x^*\|_1-\|x^+\|_1),$$

which together with (A.2) gives

$$(d^{(\lambda)}(x)+\lambda(g-g^*))^\top(x^*-x^+)\geq 0. \tag{A.3}$$

Expanding (A.3) and rearranging terms, we obtain

$$d^{(\lambda)}(x)^\top(x^*-x)-\|d^{(\lambda)}(x)\|^2+\lambda(g^*-g)^\top d^{(\lambda)}(x)\geq\lambda(g-g^*)^\top(x-x^*). \tag{A.4}$$

Using the Schwartz inequality, inequalities (2.6) and (2.7) in (A.3), we obtain

$$(1+\lambda L)\|d^{(\lambda)}(x)\|\,\|x^*-x\|-\|d^{(\lambda)}(x)\|^2\geq\lambda\omega\|x^*-x\|^2,$$

which proves **P8**. □

REFERENCES

[1] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.

[2] J. V. Burke and J. J. Moré, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211.

[3] ———, *Exposing constraints*, SIAM J. Optim., 4 (1994), pp. 573–595.

[4] R. H. Byrd, N. I. M. Gould, J. Nocedal, and R. A. Waltz, *An algorithm for nonlinear optimization using linear programming and equality constrained subproblems*, Math. Program., 100 (2004), pp. 27–48.

[5] ———, *On the convergence of successive linear-quadratic programming algorithms*, SIAM J. Optim., 16 (2005), pp. 471–489.

[6] P. H. Calamai and J. J. Moré, *Projected gradient methods for linearly constrained problems*, Math. Programming, 39 (1987), pp. 93–116.

[7] P. L. Combettes and J.-C. Pesquet, *Proximal thresholding algorithm for minimization over orthonormal bases*, To appear in SIAM Journal on Optimization, (2007).

[8] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-region methods*, MPS/SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[9] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Communications in Pure and Applied Mathematics, 57 (2004), pp. 1413–1457.

[10] F. Facchinei, A. Fischer, and C. Kanzow, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1999), pp. 14–32.

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

[12] W. W. Hager and H. Zhang, *A new active set algorithm for box constrained optimization*, SIAM J. Optim., 17 (2006), pp. 526–557.

[13] E. T. Hale, W. Yin, and Y. Zhang, *Fixed-point continuation for l_1-minimization: methodology and convergence*, SIAM J. Optim., 19 (2008), pp. 1107–1130.

[14] H. Lee, A. Battle, R. Raina, and A. Ng, *Efficient sparse coding algorithms*, in Advances in Neural Information Processing Systems, B. Schölkopf, J. Platt, and T. Hoffman, eds., vol. 19, MIT Press, Cambridge, MA, 2007, pp. 801–808.

[15] J. J. Moré and G. Toraldo, *Algorithms for bound constrained quadratic programming problems*, Numer. Math., 55 (1989), pp. 377–400.

[16] ———, *On the solution of large quadratic programming problems with bound constraints*, SIAM J. Optim., 1 (1991), pp. 93–113.

[17] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program., 117 (2009), pp. 387–423.

[18] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang, *A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization and continuation*, SIAM Journal on Scientific Computing, 32 (2010), pp. 1832–1857.

[19] H. Zhang and W. W. Hager, *A nonmonotone line search technique and its application to unconstrained optimization*, SIAM J. Optim., 14 (2004), pp. 1043–1056.