MATHEMATICS OF COMPUTATION
Volume 87, Number 311, May 2018, Pages 1283–1305
http://dx.doi.org/10.1090/mcom/3238
Article electronically published on August 3, 2017

GRADIENT-BASED METHOD WITH ACTIVE SET STRATEGY FOR ℓ_1 OPTIMIZATION

WANYOU CHENG AND YU-HONG DAI

Abstract. In this paper, we propose an identification function and develop an active set identification technique for solving the ℓ_1 optimization problem. Such a technique has a strong ability to accurately identify the zero components in a neighbourhood of an isolated stationary point without strict complementarity conditions. Based on the active set identification technique, we propose a gradient-based method for the ℓ_1 optimization problem. To accelerate the algorithm, a subspace Barzilai-Borwein steplength and a subspace exact steplength are developed, respectively. Under appropriate conditions, we show that the method with the nonmonotone line search technique is globally convergent. Numerical experiments with compressive sensing problems show that our approach is competitive with several known methods for the standard ℓ_2-ℓ_1 problem.

1. Introduction

In this paper, we consider the optimization problem

$$\min \phi(x) := f(x) + \mu\|x\|_1, \tag{1.1}$$

where f is continuously differentiable, $\mu > 0$ and $\|\cdot\|_1$ is the one-norm. A special case of (1.1), that has attracted much attention in signal/image denoising and data mining/classification, is the well-known ℓ_2-ℓ_1 problem

$$\min_{x\in\mathcal{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_1, \tag{1.2}$$

where $A \in \mathcal{R}^{m\times n}$ is dense (usually $m \le n$), $b \in \mathcal{R}^m$ and n is large.

Recently, there have been many approaches for solving (1.1). One of the most popular methods for solving problem (1.2) is the class of iterative shrinkage-thresholding algorithms (ISTA), where each iteration involves a matrix-vector multiplication involving A and A^T followed by a shrinkage/soft-threshold step; see, e.g., [18, 24]. To accelerate the convergence, a two-step ISTA (TWISTA) algorithm was developed in [7] and the sequential subspace optimization techniques was added to ISTA [23] (nevertheless, global nonasymptotic rate of convergence has not been

Received by the editor November 28, 2014 and, in revised form, October 4, 2015, July 30, 2016, and October 23, 2016.

2010 *Mathematics Subject Classification.* Primary 90C06, 90C25, 65Y20, 94A08.

Key words and phrases. ℓ_1 minimization, compressive sensing, active set, Barzilai-Borwein method.

The authors were supported by the Chinese NSF Grant (nos. 11371154, 11331012 and 81173633), the Key Project of Chinese National Programs for Fundamental Research and Development (no. 2015CB856000), the China National Funds for Distinguished Young Scientists (no. 11125107) and by the Guangdong Province Outstanding Young Teacher Training Program (nos. 3XZ150603 and 2015KTSCX1).

established for the methods in [7, 23]). Beck and Teboulle [5] constructed a faster shrinkage-thresholding algorithm, called FISTA, that keeps its simplicity of ISTA but provides a better global convergence rate. To improve practical performances of the above methods, Wright et al. [40] introduced the sparse reconstruction by separable approximation (SpaRSA) algorithm for solving (1.1). The rules for choosing the parameter and the line search are quite different. Hager et al. [29] analyzed the convergence rate of SpaRSA and proposed an improved version of SpaRSA based on a cyclic version of the Barzilai-Borwein iteration [4] and an adaptive choice for the reference function value in the line search. Hale et al. [28] proposed a fixed point continuation (FPC) algorithm that embeds the soft-thresholding (ISTA) algorithm [18, 24] in a continuation strategy. Wen et al. [38] improved the performance of FPC by adding an active set (AS) step and proposed an abridged version of the active-set algorithm FPC_AS for solving (1.2). The global convergence of FPC_AS is shown in [39].

Other algorithms for the l_1 minimization include alternating direction method of multipliers SALSA [1, 9]; coordinatewise descent methods [36]; interior point methods [12, 30, 35]; projected gradient methods [25]; Bergman iterative regularization based methods [41]; gradient methods [32] for minimizing the more general function $J(x) + H(x)$, where J is nonsmooth, H is smooth, and both are convex; smoothed penalty algorithm (SPA) [2, 3]; second-order methods [26, 31]. We refer to papers [6, 10, 20] for recent advances in this area.

In this paper, we shall propose an identification function and an active set identification technique. Such a technique has a strong ability to accurately identify the zero components in a neighbourhood of an isolated stationary point of (1.1) without strict complementarity conditions. Based on this active set identification technique, we propose a gradient-based method for solving (1.1). Specifically, at each iteration, the active variables and free variables are defined by the identification technique; we take $d^k = -x^k$ to update some active variables, while using a gradient-based method to update the free variables and some active variables. Hence the method is distinct from the existing methods [38, 39] in that it uses a first-order method to obtain a working index set and utilizes a second-order method to solve a smooth subproblem defined by the working index set. To accelerate the convergence of the algorithm, a subspace Barzilai-Borwein steplength and a subspace exact steplength are provided, respectively. The use of the gradient method reduces the storage requirement of our method. Hence, the method can be used to solve large-scale problems. In addition, the method has the following advantages: (a) The method is suitable for solving a more general problem of (1.1); (b) Rapid changes in the active set are allowed; (c) The method based on a nonmonotone line search technique [27] is showed to be globally convergent; (d) The main computational burden at each iteration involves matrix-vector multiplication involving A and A^T; (e) Preliminary numerical experiments show that the method is effective and competitive with the state-of-the-art methods.

The remainder of this paper is organized as follows. We propose the algorithm and the steplength strategy in Section 2. In Section 3, we establish the global convergence of the algorithm. Some numerical results are reported in Section 4 and conclusions are made in the last section.

Throughout the paper, $\|.\|$ denotes the Euclidean norm of vectors. A_T and x_T denote the collections of columns and entries of A and x, whose indices are in an index set $T \subset \{1, 2, 3, \cdots, n\}$, respectively.

2. Motivation and properties

In this section, we propose a new active-set algorithm for solving (1.1). We begin with some notations. Let $\bar{x}$ be a stationary point of (1.1). We define the active set $\mathcal{G}(\bar{x})$ to be the set of indices corresponding to the zero components and the inactive set $\mathcal{T}(\bar{x})$ to be support of $\bar{x}$, i.e.,

$$\mathcal{G}(\bar{x}) = \{i : \bar{x}_i = 0\} \text{ and } \mathcal{T}(\bar{x}) = \{i : \bar{x}_i \neq 0\}.$$

The active set is further subdivided into two sets

$$\mathcal{G}_+(\bar{x}) = \{i \in \mathcal{G}(\bar{x}) : |g_i(\bar{x})| < \mu\} \text{ and } \mathcal{G}_0(\bar{x}) = \{i \in \mathcal{G}(\bar{x}) : |g_i(\bar{x})| \geq \mu\},$$

where $g_i(\bar{x})$ is the i-th component of the gradient vector of f at $\bar{x}$. Then the first-order necessary condition for $\bar{x}$ to solve (1.1) is equivalent to the following system:

$$\begin{cases} g_i(\bar{x}) + \alpha_i \mu = 0, & \text{if } i \in \mathcal{G}(\bar{x}), \\ g_i(\bar{x}) + \mu \operatorname{sgn}(\bar{x}_i) = 0, & \text{if } i \in \mathcal{T}(\bar{x}), \end{cases}$$

where $\alpha_i \in$[-1,1] is a constant and $\operatorname{sgn}(t) : \mathcal{R} \to \mathcal{R}$ is the sign function defined by

$$\operatorname{sgn}(t) = \begin{cases} 1, & \text{if } t > 0, \\ 0, & \text{if } t = 0, \\ -1, & \text{if } t < 0. \end{cases}$$

In general, the efficiency of an active-set algorithm heavily depends on how fast the correct support of the solution point can be identified.

In what follows, we introduce an active-set identification technique to approximate the correct support. To this aim, we briefly recall the general definition of identification function [11].

Definition 2.1. A continuous function $\rho(x) : \mathcal{R}^n \to \mathcal{R}_+$ is called an identification function for $\bar{x}$ with respect to a sequence $\{x^k\}$ if $\rho(\bar{x}) = 0$ and

$$\lim_{x^k \to \bar{x},\, x^k \neq \bar{x}} \frac{\rho(x^k)}{\|x^k - \bar{x}\|} = +\infty.$$

For $y \in \mathcal{R}^n$ and $\mu > 0$, define the shrinkage operator

$$S(y, \mu) := \operatorname{sgn}(y) \odot \max\{|y| - \mu, 0\},$$

where $\odot$ denotes componentwise multiplication. Denoting

$$\psi_\nu(x) = S(x - \nu g(x), \nu\mu) - x,$$

where ν is some positive constant, we shall consider the following identification function for problem (1.1),

$$\rho(x) = \min\{c_1, c_2\sqrt{\|\psi_\nu(x)\|}\},$$

where c_1 and c_2 are both positive constants. Based on this identification function, we give the following approximations $\mathcal{A}(x)$ to $\mathcal{G}(\bar{x})$ and $\mathcal{F}(x)$ to $\mathcal{T}(\bar{x})$, respectively,

$$\mathcal{A}(x) = \{i : |x_i| \leq \rho(x)\} \text{ and } \mathcal{F}(x) = \{i : |x_i| > \rho(x)\}.$$

The following theorem shows that $\rho(x)$ is an identification function for $\bar{x}$ indeed and the set $\mathcal{A}(x)$ can accurately identify the zero components in a certain neighborhood of an isolated stationary point of (1.1). Hence, the set $\mathcal{A}(x)$ is a good estimate of $\mathcal{G}(\bar{x})$. We make the following assumptions on the objective function.

Assumption 2.1.

(i) *The level set* $\Omega := \{x \in \mathcal{R}^n : \phi(x) \le \phi(x^0)\}$ *is bounded;*
(ii) f *is continuously differentiable on an open set* N *containing* Ω.

Theorem 2.1. *Suppose that Assumption* 2.1 *holds. If* $\bar{x}$ *is an isolated stationary point of problem* (1.1) *and satisfies the strong second-order sufficient optimality conditions, then there exists a neighborhood* $N(\bar{x})$ *of* $\bar{x}$ *such that*

$$\mathcal{A}(x) = \mathcal{G}(\bar{x}), \quad \forall\ x \in N(\bar{x}).$$

Proof. For any constant $\nu > 0$, we see by Lemma 2 in [36] that $\psi_\nu(x) = 0$ if and only if x is a stationary point of problem (1.1). Furthermore, by P4 and P5 of Lemma 2.1 in [39], we get

$$\|\psi_{\nu_1}(x)\| \ge \|\psi_{\nu_2}(x)\| \ \text{ and } \ \frac{\|\psi_{\nu_1}(x)\|}{\nu_1} \le \frac{\|\psi_{\nu_2}(x)\|}{\nu_2}$$

for any $x \in \mathcal{R}^n$ and for all $\nu_1 \ge \nu_2 > 0$. Hence we have $\|\psi_1(x)\| \le \|\psi_\nu(x)\|$ for $\nu > 1$ and $\|\psi_1(x)\| \le \frac{\|\psi_\nu(x)\|}{\nu}$ for $\nu \in (0, 1]$. It follows that $\|\psi_1(x)\| \le \|\psi_\nu(x)\| \max(1, \frac{1}{\nu})$ for all $\nu > 0$. By Lemma 4.7 in [39], we know that there exists some constant $\bar{\tau} > 0$ such that

$$\|x - \bar{x}\| \le \bar{\tau}\|\psi_1(x)\|.$$

Thus we have for all $x \in R^n$,

$$\|x - \bar{x}\| \le \bar{\tau}\|\psi_\nu(x)\| \max(1, \frac{1}{\nu}).$$

Therefore, we can obtain

$$\lim_{x\to\bar{x},\ x\neq\bar{x}} \frac{\rho(x)}{\|x - \bar{x}\|} \ge \frac{1}{\max(1, \frac{1}{\nu})} \frac{\min(c_1, c_2\sqrt{\|\psi_\nu(x)\|})}{\bar{\tau}\|\psi_\nu(x)\|} \to +\infty, \quad \text{as } x \to \bar{x},$$

because $\psi_\nu(x) \to 0$ as $x \to \bar{x}$. This implies that $\rho(x)$ is an identification function for $\bar{x}$. Then the statement follows from Lemma 4.15 in [39]. □

In what follows, we are going to develop an active set gradient-based method for solving (1.1). At first, we define the search direction. Let $x^k \in \Omega$ be the k-th iteration. For simplicity, we let $\mathcal{A}^k = \mathcal{A}(x^k)$ and $\mathcal{F}^k = \mathcal{F}(x^k)$. Define the direction $d^k = (d^k_{\mathcal{A}^k}, d^k_{\mathcal{F}^k})^T$ by

$$d^k_i = -\lambda^k \nabla\phi_i(x^k), \quad \forall i \in \mathcal{F}(x_k), \tag{2.1}$$

where $\alpha_{\min} \le \lambda^k \le \alpha_{\max}$, $\alpha_{\min}$ and $\alpha_{\max}$ are positive constants and $\nabla\phi_i(z) = \frac{\partial\phi(x)}{\partial x_i}|_{x=z}$. To define $d^k_{\mathcal{A}^k}$, we partition the active set $\mathcal{A}^k$ into two parts:

$$\begin{aligned} \mathcal{A}^k_1 &= \{i \in \mathcal{A}^k : |g^k_i| \le \mu\}, \\ \mathcal{A}^k_2 &= \{i \in \mathcal{A}^k : |g^k_i| > \mu\}, \end{aligned}$$

where $g^k_i = g_i(x^k)$. It is easy to see that $\mathcal{A}^k_1$ is the index set of variables that approximately satisfies the first-order necessary conditions. Thus, it is reasonable for us to fix those variables with indices in $\mathcal{A}^k_1$ to 0. On the other hand, noting that $\mathcal{A}^k_2$ is the index set of variables that violate the first-order necessary conditions, we further subdivided $\mathcal{A}^k_2$ into two subsets $\mathcal{A}^k_{21} = \{i \in \mathcal{A}^k_2 :\ x^k_i = 0\}$ and $\mathcal{A}^k_{22} = \{i \in \mathcal{A}^k_2 :\ x^k_i \neq 0\}$. We consider the direction of the form $-(g^k_i - \mu\,\mathrm{sgn}(g^k_i))$ for those

variables with indices in $\mathcal{A}_{21}^k$ and $-\nabla\phi_i(x^k)$ for those variables with indices in $\mathcal{A}_{22}^k$ to improve the corresponding components. Hence, we define the direction $d_{\mathcal{A}^k}^k$ by

$$d_i^k = -x_i^k, \quad i \in \mathcal{A}_1^k, \tag{2.2}$$

$$d_i^k = -(g_i^k - \mu\,\mathrm{sgn}(g_i^k)), \quad i \in \mathcal{A}_{21}^k, \tag{2.3}$$

$$d_i^k = -(g_i^k + \mu\,\mathrm{sgn}(x_i^k)), \quad i \in \mathcal{A}_{22}^k. \tag{2.4}$$

It is easy to see that

$$|d_i^k| = |g_i^k| - \mu > 0, \quad \forall i \in \mathcal{A}_{21}^k, \tag{2.5}$$

and

$$|d_i^k| \geq |g_i^k| - \mu > 0, \quad \forall i \in \mathcal{A}_{22}^k. \tag{2.6}$$

If the unit steplength is accepted, we will have that $x_i^{k+1} = 0$ for all $i \in \mathcal{A}_1^k$ (see Algorithm 2.1). We can prove that the d^k defined by (2.1)–(2.4) is a descent direction of ϕ at x^k (see Theorem 2.5). In addition, in our numerical experiments, we observed that the $d_{\mathcal{A}_{21}^k}^k$ performs better than taking the normal negative gradient for this part.

The following theorem shows that $d^k = 0$ if and only if the iteration point x^k is a stationary point of problem (1.1).

Theorem 2.2. *Let d^k be determined by* (2.1)–(2.4). *Then, $d^k = 0$ if and only if x^k is a stationary point of problem* (1.1).

Proof. Let $d^k = 0$. By (2.5) and (2.6), we obtain $\mathcal{A}_2^k = \emptyset$. If $x_i^k = 0$, then it must hold that

$$i \in \mathcal{A}_1^k,$$

which implies that $|g_i^k| \leq \mu$. If $x_i^k \neq 0$, by (2.2), then it must hold that

$$i \in \mathcal{F}^k.$$

This together with (2.1) implies that $\nabla\phi_i(x^k) = 0$. Therefore, $d^k = 0$ implies that x^k is a stationary point of ϕ.

Suppose that x^k is a stationary point of problem (1.1). Then we have $\rho(x^k) = 0$. Consequently, we have

$$\mathcal{A}^k = \{i : x_i^k = 0\} \text{ and } \mathcal{F}^k = \{i : x_i^k \neq 0\}.$$

Furthermore, we have $\mathcal{A}_2^k = \emptyset$. From (2.1)–(2.4), we immediately get that $d^k = 0$ since x^k is a stationary point of problem (1.1). The proof is completed. □

The following theorem shows that if $x^k \to x^*$ and $d^k \to 0$, then x^* is a stationary point of problem (1.1).

Theorem 2.3. *Assume that $x^k \to x^*$ and $d^k \to 0$, where d^k is determined by* (2.1)–(2.4). *Then x^* is a stationary point of problem* (1.1).

Proof. Taking into account that the number of distinct sets $\mathcal{A}_1^k$, $\mathcal{A}_{21}^k$, $\mathcal{A}_{22}^k$, $\mathcal{F}^k$ is finite, there exists a subsequence that (without loss of generality, we label again $\{x^k\}$) such that index sets $\mathcal{A}_1^k$, $\mathcal{A}_{21}^k$, $\mathcal{A}_{22}^k$, $\mathcal{F}^k$ are constant and hence we can write

$$\mathcal{A}_1^k = \mathcal{A}_1, \ \ \mathcal{A}_{21}^k = \mathcal{A}_{21}, \ \ \mathcal{A}_{22}^k = \mathcal{A}_{22} \text{ and } \mathcal{F}^k = \mathcal{F}.$$

If $\rho(x^*) = 0$, then x^* is a stationary point of problem (1.1) and the conclusion is clear. Assume that $\rho(x^*) \neq 0$. By the continuity of $g(x)$ and the assumption that $\lim_{k\to\infty} d^k = 0$, we have

$$\lim_{k\to\infty} \mu \operatorname{sgn}(x_i^k) = -g_i(x^*) \neq 0, \ \ \forall i \in \mathcal{A}_{22}.$$

Thus, we get that

$$\lim_{k\to\infty} \operatorname{sgn}(x_i^k) = -\operatorname{sgn}(g_i(x^*)) \neq 0, \ \ \forall i \in \mathcal{A}_{22}. \tag{2.7}$$

We consider two cases for each $i \in \mathcal{A}_{22}$. In the first case, suppose that $x_i^* \neq 0$. By the assumption that $x_i^k \to x_i^* \neq 0$ and the preserving sign property of the limit, we get

$$\lim_{k\to\infty} \operatorname{sgn}(x_i^k) = \operatorname{sgn}(x_i^*).$$

By the continuity of $g(x)$, the assumption that $\lim_{k\to\infty} d^k = 0$ and the last equality, we have

$$0 = \lim_{k\to\infty} -d_i^k = \lim_{k\to\infty} g_i(x^k) + \mu \operatorname{sgn}(x_i^k) = g_i(x^*) + \mu \operatorname{sgn}(x_i^*) = \nabla\phi_i(x^*).$$

In the second case, suppose that $x_i^* = 0$. Again, by the continuity of $g(x)$, the assumption that $\lim_{k\to\infty} d^k = 0$ and (2.7), we have

$$0 = \lim_{k\to\infty} -d_i^k = \lim_{k\to\infty} g_i(x^k) + \mu \operatorname{sgn}(x_i^k) = g_i(x^*) - \mu \operatorname{sgn}(g_i(x^*)),$$

which implies that $|g_i(x^*)| = \mu$. Since $d^k \to 0$ and $x^k \to x^*$, by the definition of $\mathcal{F}$, $\mathcal{A}_1$, $\mathcal{A}_{21}$, $\mathcal{A}_{22}$ and the continuity of $\rho(x)$, we get the following system:

$$\begin{cases} x_i^* \neq 0 \ \text{ and } \ \nabla\phi_i(x^*) = 0, & \text{if } \ i \in \mathcal{F}, \\ x_i^* = 0 \ \text{ and } \ |g_i(x^*)| \leq \mu, & \text{if } \ i \in \mathcal{A}_1, \\ x_i^* = 0 \ \text{ and } \ |g_i(x^*)| = \mu, & \text{if } \ i \in \mathcal{A}_{21}, \\ x_i^* \neq 0, \ \ \nabla\phi_i(x^*) = 0 \ \text{ or } \ x_i^* = 0, \ \ |g_i(x^*)| = \mu, & \text{if } \ i \in \mathcal{A}_{22}, \end{cases}$$

which shows that x^* is a stationary point of problem (1.1). □

The next two theorems show that d^k is a descent direction of ϕ at x_k provided x_k is not a stationary point of problem (1.1).

Theorem 2.4. *Let d^k be determined by (2.1)–(2.4). Then, we have*

$$\sum_{i\in\mathcal{F}^k} -(\nabla\phi_i^k)^2 + \sum_{i\in\mathcal{A}_1^k} (g_i^k + \mu \operatorname{sgn}(x_i^k))\, d_i^k + \sum_{i\in\mathcal{A}_2^k} -(d_i^k)^2 \leq 0. \tag{2.8}$$

Furthermore, the equality holds if and only if x^k is a stationary point of problem (1.1).

Proof. To get (2.8), it suffices to prove

$$(g_i^k + \mu \operatorname{sgn}(x_i^k))\, d_i^k \leq 0, \ \ \forall i \in \mathcal{A}_1^k. \tag{2.9}$$

If $x_i^k = 0$, then $d_i^k = 0$ and the conclusion is clear. If $d_i^k = -x_i^k \neq 0$, by $|g_i^k| \leq \mu$, we have

$$g_i^k + \mu \operatorname{sgn}(x_i^k) = \begin{cases} g_i^k + \mu \geq 0 & \text{if } \ x_i^k > 0, \\ g_i^k - \mu \leq 0, & \text{if } \ x_i^k < 0. \end{cases}$$

Thus, we get (2.9).

Suppose that x^k is a stationary point of problem (1.1). Then we have $\rho(x^k) = 0$,

$$\mathcal{A}^k = \{i : x_i^k = 0\}, \ \ \mathcal{F}^k = \{i : x_i^k \neq 0\} \ \text{ and } \ \mathcal{A}_2^k = \emptyset.$$

Furthermore, we get

$$\nabla\phi_i^k = 0, \ \forall i \in \mathcal{F}^k \ \text{ and } \ d_i^k = 0, \ \ \forall i \in \mathcal{A}_1^k.$$

Thus, the left-hand side of the inequality (2.8) is equal to zero. On the other hand, suppose that the left-hand side of the inequality (2.8) is equal to zero. By (2.5), (2.6) and the assumption, we have $\mathcal{A}_2^k = \emptyset$ and

$$\nabla\phi_i^k = 0, \ \ \forall i \in \mathcal{F}^k \ \text{ and } \ (g_i^k + \mu\,\mathrm{sgn}(x_i^k))\, d_i^k = 0, \ \ \forall i \in \mathcal{A}_1^k. \tag{2.10}$$

If $x_i^k = 0$, then it must hold

$$i \in \mathcal{A}_1^k,$$

which implies that $|g_i^k| \leq \mu$. If $x_i^k \neq 0$, then it must hold that

$$i \in \mathcal{F}^k \cup \mathcal{A}_1^k,$$

which implies that $\nabla\phi_i(x^k) = 0$ by (2.2) and (2.10). This shows that x^k is a stationary point of problem (1.1). The proof is completed. □

The following theorem shows that d^k is a descent direction of ϕ at x_k provided x_k is not a stationary point of problem (1.1).

Theorem 2.5. *Suppose that x^k is not a stationary point of problem* (1.1). *Then the direction d^k determined by* (2.1)–(2.4) *is a descent direction of ϕ at x^k, i.e., there is a positive constant $\bar{\alpha}$ such that*

$$\phi(x^k + \alpha d^k) < \phi(x^k), \ \ \forall \alpha \in (0, \bar{\alpha}).$$

Proof. By the definition of d^k and (2.5), we have

$$\begin{aligned}
&\phi(x^k + \alpha d^k) - \phi(x^k) \\
&\quad = \ f(x^k + \alpha d^k) - f(x^k) + \mu(\|x^k + \alpha d^k\|_1 - \|x^k\|_1) \\
&\quad = \ \alpha g(x^k)^T d^k + \mu \sum_{i=1}^{n} (|x_i^k + \alpha d_i^k| - |x_i^k|) + o(\alpha) \\
&\quad \leq \ \sum_{i \in \mathcal{F}^k \cup \mathcal{A}_1^k \cup \mathcal{A}_{22}^k} (\alpha g_i^k d_i^k + \mu|x_i^k + \alpha d_i^k| - \mu|x_i^k|) + \alpha \sum_{i \in \mathcal{A}_{21}^k} (g_i^k d_i^k + \mu|d_i^k|) + o(\alpha) \\
&\quad = \ \alpha \sum_{i \in \mathcal{F}^k \cup \mathcal{A}_{22}^k} (g_i^k d_i^k + \mu d_i^k \mathrm{sgn}(x_i^k)) + \alpha \sum_{i \in \mathcal{A}_{21}^k} (g_i^k d_i^k + \mu|d_i^k|) \\
&\qquad + \sum_{i \in \mathcal{A}_1^k} (\alpha g_i^k d_i^k + \mu|(1-\alpha)x_i^k| - \mu|x_i^k|) + o(\alpha) \\
&\quad \leq \ \alpha \sum_{i \in \mathcal{F}^k} -\frac{(d_i^k)^2}{\alpha_{\max}} + \alpha \sum_{i \in \mathcal{A}_{22}^k} -(d_i^k)^2 + \alpha \sum_{i \in \mathcal{A}_{21}^k} \big(g_i^k(-g_i^k + \mu\,\mathrm{sgn}(g_i^k)) + \mu(|g_i^k| - \mu)\big) \\
&\qquad + \alpha \sum_{i \in \mathcal{A}_1^k} (g_i^k d_i^k - \mu|x_i^k|) + o(\alpha)
\end{aligned}$$

$$
\begin{aligned}
&= \alpha \sum_{i\in\mathcal{F}^k} -\frac{(d_i^k)^2}{\alpha_{\max}} + \alpha \sum_{i\in\mathcal{A}_{22}^k} -(d_i^k)^2 + \alpha \sum_{i\in\mathcal{A}_{21}^k} (-(g_i^k)^2 + 2\mu|g_i^k| - \mu^2) \\
&\quad +\alpha \sum_{i\in\mathcal{A}_1^k} (g_i^k d_i^k - \mu\,\mathrm{sgn}(x_i^k)x_i^k) + o(\alpha) \\
&= \alpha \sum_{i\in\mathcal{F}^k} -\frac{(d_i^k)^2}{\alpha_{\max}} + \alpha \sum_{i\in\mathcal{A}_{22}^k} -(d_i^k)^2 + \alpha \sum_{i\in\mathcal{A}_{21}^k} -(|g_i^k| - \mu)^2 \\
&\quad +\alpha \sum_{i\in\mathcal{A}_1^k} (g_i^k d_i^k + \mu\,\mathrm{sgn}(x_i^k)d_i^k) + o(\alpha) \\
&= \alpha \sum_{i\in\mathcal{F}^k} -\frac{(d_i^k)^2}{\alpha_{\max}} + \alpha \sum_{i\in\mathcal{A}_2^k} -(d_i^k)^2 + \alpha \sum_{i\in\mathcal{A}_1^k} (g_i^k + \mu\,\mathrm{sgn}(x_i^k))d_i^k + o(\alpha).
\end{aligned}
$$

In the above, the first inequality uses the triangle equality for each $i \in \mathcal{A}_{21}^k$, the third equality is due to (2.2) and the fact that for any $i \in \mathcal{F}^k \cup \mathcal{A}_{22}^k$,

$$
|x_i^k + \alpha d_i^k| - |x_i^k| = \begin{cases} \alpha d_i^k, & \text{if } x_i^k > 0, \\ -\alpha d_i^k, & \text{if } x_i^k < 0, \end{cases} = \alpha d_i^k \mathrm{sgn}(x_i^k) \tag{2.11}
$$

as $\alpha \to 0$. Since x^k is not a stationary point of problem (1.1), we immediately get the conclusion by Theorem 2.4. The proof is completed. □

Based on the above discussion, we propose an active set gradient-based method for solving (1.1) as follows.

Algorithm 2.1 (Active set gradient-based method)**.**

Step 0. *Given an initial point* $x^0 \in \mathcal{R}^n$ *and positive constants* M, $\eta \in (0,1)$ *and* $\delta \in (0,1)$. *Set* $k := 0$.
Step 1. *Perform the convergence test and terminate with an approximate solution* x^k *if the stopping criterion is satisfied.*
Step 2. *Compute* d^k *by* (2.1)–(2.4).
Step 3. *Determine* $\alpha^k := \max\{\eta^j, j = 0, 1, \ldots\}$ *satisfying*

$$
\phi(x^k + \alpha^k d^k) \le \phi_{\max}^k - \delta(\alpha^k \|d^k\|)^2, \tag{2.12}
$$

where $\phi_{\max}^k := \max\{\phi(x^{k-j}) : 0 \le j \le \min(k, M-1)\}$.
Step 4. *Let the next iterate be* $x^{k+1} := x^k + \alpha^k d^k$.
Step 5. *Set* $k := k + 1$ *and go to Step* 1.

Since $\phi_{\max}^k \ge \phi(x^k)$, Theorem 2.5 implies that the condition (2.12) must hold after a finite number of reductions of α^k. Consequently, Algorithm 2.1 is well-defined.

To accelerate the active set gradient-based method, we shall propose a subspace Barzilai-Borwein steplength and a subspace exact steplength, respectively. To this aim, we briefly recall the Barzilai-Borwein method (see, e.g., [4, 16]). Consider the unconstrained minimization problem

$$
\min_{x\in\mathcal{R}^n} h(x),
$$

where $h : \mathcal{R}^n \to \mathcal{R}$ is continuously differentiable. The Barzilai-Borwein method is defined by

$$
x^{k+1} = x^k - \alpha_{BB}^k \nabla h(x^k),
$$

where the scalar α_{BB}^k is given by

$$\alpha_{BB}^k = \frac{\|s^{k-1}\|^2}{(s^{k-1})^T y^{k-1}}, \tag{2.13}$$

where α_{BB}^k is called the Barzilai-Borwein steplength, $s^{k-1} = x^k - x^{k-1}$ and $y^{k-1} = \nabla h(x^k) - \nabla h(x^{k-1})$. The basic idea of Barzilai-Borwein (BB) [4] method is to use the diagonal matrix $D(\alpha) = \frac{1}{\alpha} I$, where I denotes the identity matrix, to approximate the Hessian $\nabla^2 h(x^k)$ by imposing a certain quasi-Newton condition on $D(\alpha)$:

$$\alpha_{BB}^k = \arg\min_{\alpha \in \mathcal{R}} \|D(\alpha) s_{k-1} - y_{k-1}\|^2. \tag{2.14}$$

By straightforward calculations and (2.14), we obtain (2.13). Due to its easy implementation, numerical efficiency and low storage requirement, BB-type methods have widely been used in many applications such as box constrained optimization [8, 15], nonlinear equations [13] and sparse reconstruction [25, 40]. Some recent analysis on the BB method can be found in [14, 15, 17]. Similar to the idea of the Barzilai-Borwein steplength, by solving the following optimization problem

$$\min_{\lambda \in \mathcal{R}} \|D(\lambda) s_{\mathcal{F}^k}^{k-1} - y_{\mathcal{F}^k}^{k-1}\|^2,$$

where $s_{\mathcal{F}^k}^{k-1} = x_{\mathcal{F}^k}^k - x_{\mathcal{F}^k}^{k-1}$ and $y_{\mathcal{F}^k}^{k-1} = g_{\mathcal{F}^k}^k + \mu \operatorname{sgn}(x_{\mathcal{F}^k}^k) - g_{\mathcal{F}^k}^{k-1} - \mu \operatorname{sgn}(x_{\mathcal{F}^k}^{k-1})$, we obtain the subspace Barzilai-Borwein steplength as follows

$$\lambda_{\mathrm{BB}}^k = \frac{\|s_{\mathcal{F}^k}^{k-1}\|^2}{(s_{\mathcal{F}^k}^{k-1})^T y_{\mathcal{F}^k}^{k-1}}.$$

To avoid small and large values of λ_{BB}^k, we project it in the interval $[\alpha_{\min}, \alpha_{\max}]$, where $\alpha_{\min} < \alpha_{\max}$ are given positive constants. That is, we let

$$\lambda_1^k = \max\{\alpha_{\min}, \min\{\alpha_{\max}, \lambda_{\mathrm{BB}}^k\}\}. \tag{2.15}$$

For simplicity, we call Algorithm 2.1 with the steplength (2.15) used in (2.1) as the active set Barzilai-Borwein algorithm and abbreviate it as **ABB**.

In what follows, we describe the subspace exact steplength for solving (1.2). To this aim, consider the unconstrained minimization problem

$$\min_{\alpha \in \mathcal{R}} F(x_{\mathcal{F}^k}^k + \alpha d_{\mathcal{F}^k}^k) := \frac{1}{2} \|A_{m \times \mathcal{F}^k}(x_{\mathcal{F}^k}^k + \alpha d_{\mathcal{F}^k}^k) - b\|^2 + \mu \|x_{\mathcal{F}^k}^k + \alpha d_{\mathcal{F}^k}^k\|_1.$$

The solution of the above problem is

$$\alpha = \frac{-(d_{\mathcal{F}^k}^k)^T (A_{m \times \mathcal{F}^k}^T r_{\mathcal{F}^k}^k + \mu \operatorname{sgn}(x_{\mathcal{F}^k}^k + \alpha d_{\mathcal{F}^k}^k))}{(A_{m \times \mathcal{F}^k} d_{\mathcal{F}^k})^T A_{m \times \mathcal{F}^k} d_{\mathcal{F}^k}^k},$$

where $d_{\mathcal{F}^k}^k = -\nabla \phi_{\mathcal{F}^k}$ and $r_{\mathcal{F}^k}^k = A_{m \times \mathcal{F}^k} x_{\mathcal{F}^k}^k - b$. The parameter α is contained in the right-hand side of the above equality as well and hence is difficult to calculate. To overcome this difficulty, we use the following $\bar{\alpha}$ to approximate α. That is,

$$\alpha \approx \bar{\alpha} = \frac{-(d_{\mathcal{F}^k}^k)^T (A_{m \times \mathcal{F}^k}^T r_{\mathcal{F}^k}^k + \mu \operatorname{sgn}(x_{\mathcal{F}^k}^k + d_{\mathcal{F}^k}^k))}{(A_{m \times \mathcal{F}^k} d_{\mathcal{F}^k})^T A_{m \times \mathcal{F}^k} d_{\mathcal{F}^k}^k}.$$

Since the matrix A is not column full rank, it follows that $A_{\mathcal{F}^k} d_{\mathcal{F}^k}$ may be equal to zero. Therefore, we confine it in the interval $[\alpha_{\min}, \alpha_{\max}]$. Namely,

$$\lambda_2^k := \max\{\alpha_{\min}, \min\{\alpha_{\max}, \bar{\alpha}\}\}.$$

In this case, we call Algorithm 2.1 with the steplength λ_2^k used in (2.1) as **AGE**. Notice that AGE cannot directly be used for solving (1.1) in general since the subspace exact steplength makes use of the special form of (1.2).

3. Convergence analysis

In this section, we analyze the convergence of Algorithm 2.1 under Assumption 2.1. By Assumption 2.1 and the definition of $\|x\|_1$, we get that $\phi(x)$ is uniformly continuous on the open set N containing Ω. Proceeding with a similar analysis as the one in [27], we get the following lemma. For completeness, we give the details of the proof.

Lemma 3.1. *Suppose that f satisfies Assumption 2.1 and $\{x^k\}$ is generated by Algorithm 2.1. Then*

$$\lim_{k\to\infty} \alpha^k d^k = 0. \tag{3.1}$$

Proof. Let $m(k) = \min(k, M-1)$ and $l(k)$ be an integer such that

$$k - m(k) \le l(k) \le k$$

and

$$\phi(x^{l(k)}) = \max_{0\le j\le m(k)} \phi(x^{k-j}).$$

Note that

$$\phi(x^{l(k+1)}) = \phi_{\max}^{k+1} = \max_{0\le j\le m(k+1)} \phi(x^{k+1-j}) \le \max(\phi(x^{k+1}), \phi(x^{l(k)})) \le \phi(x^{l(k)}).$$

Thus, the sequence $\{\phi(x^{l(k)})\}$ is nonincreasing. Since $\phi(x^k) \le \phi(x^0)$ for all k, $x^k \in \Omega$, there exists a constant $\bar{\phi}$ such that

$$\lim_{k\to\infty} \phi(x^{l(k)}) = \bar{\phi}. \tag{3.2}$$

By the line search condition (2.12), we get for $k > M-1$,

$$\begin{aligned}
\phi(x^{l(k)}) &= \phi(x^{l(k)-1} + \alpha^{l(k)-1} d^{l(k)-1}) \\
&\le \phi_{\max}^{l(k)-1} - \delta(\alpha^{l(k)-1}\|d^{l(k)-1}\|)^2 \\
&= \max_{0\le j\le m(l(k)-1)} \phi(x^{l(k)-1-j}) - \delta(\alpha^{l(k)-1}\|d^{l(k)-1}\|)^2 \\
&= \phi(x^{l(l(k)-1)}) - \delta(\alpha^{l(k)-1}\|d^{l(k)-1}\|)^2.
\end{aligned}$$

This together with (3.2) implies that

$$\lim_{k\to\infty} \alpha^{l(k)-1}\|d^{l(k)-1}\| = 0. \tag{3.3}$$

Let

$$\bar{l}(k) = l(k+M+2).$$

In what follows, we shall show that

$$\lim_{k\to\infty} \alpha^{\bar{l}(k)-j}\|d^{\bar{l}(k)-j}\| = 0 \tag{3.4}$$

and

$$\lim_{k\to\infty} \phi(x^{\bar{l}(k)-j}) = \bar{\phi} \tag{3.5}$$

hold for any given $j \ge 1$. Without loss of generality, we assume that the iteration index k is large enough to avoid the occurrence of negative subscripts, that is,

$k \geq j-1$. If $j=1$, (3.4) follows from (3.3) since $\{\bar{l}(k)\} \subset \{l(k)\}$. This in turn implies $\lim_{k\to\infty} \|x^{\bar{l}(k)} - x^{\bar{l}(k)-1}\| = 0$. Thus, (3.5) holds for $j=1$ since $\phi(x)$ is uniformly continuous on Ω. Assume now that (3.4) and (3.5) hold for a given j. Again by the line search condition (2.12), we get

$$\begin{aligned} \phi(x^{\bar{l}(k)-j}) &\leq \phi_{\max}^{\bar{l}(k)-j-1} - \delta(\alpha^{\bar{l}(k)-j-1}\|d^{\bar{l}(k)-j-1}\|)^2 \\ &= \phi(x^{l(\bar{l}(k)-j-1)}) - \delta(\alpha^{\bar{l}(k)-j-1}\|d^{\bar{l}(k)-j-1}\|)^2. \end{aligned}$$

Letting $k \to \infty$ and using (3.5), we have

$$\lim_{k\to\infty} (\alpha^{\bar{l}(k)-j-1}\|d^{\bar{l}(k)-j-1}\|)^2 = 0,$$

which implies that (3.4) holds for $j+1$. The last equality implies $\|x^{\bar{l}(k)-j} - x^{\bar{l}(k)-j-1}\| \to 0$. By (3.5) and the uniform continuity of ϕ on Ω, we have

$$\lim_{k\to\infty} \phi(x^{\bar{l}(k)-j-1}) = \lim_{k\to\infty} \phi(x^{\bar{l}(k)-j}) = \bar{\phi}.$$

Thus, we conclude that (3.4) and (3.5) hold for any given $j \geq 1$.

Now for any k, we have

$$x^{k+1} = x^{\bar{l}(k)} - \sum_{j=1}^{\bar{l}(k)-k-1} \alpha^{\bar{l}(k)-j} d^{\bar{l}(k)-j}. \tag{3.6}$$

By the definition of $l(k)$, we have $\bar{l}(k) - k - 1 = l(k+M+2) - k - 1 \leq M+1$. Thus, by (3.6) and (3.4), we have

$$\lim_{k\to\infty} \|x^{k+1} - x^{\bar{l}(k)}\| = 0.$$

By (3.2) and the uniform continuity of ϕ on Ω, we have

$$\lim_{k\to\infty} \phi(x^k) = \bar{\phi}. \tag{3.7}$$

Again, by the line search condition (2.12), we have

$$\phi(x^{k+1}) \leq \phi(x^{l(k)}) - \delta(\alpha^k\|d^k\|)^2.$$

Taking limits for $k \to \infty$, by (3.7), we have

$$\lim_{k\to\infty} (\alpha^k\|d^k\|)^2 = 0,$$

which implies (3.1). The proof is completed. □

The following theorem together with Theorem 2.2 shows that every accumulation point of $\{x^k\}$ is a stationary point of problem (1.1).

Theorem 3.1. *Assume that f satisfies Assumption* 2.1. *Let $\{x^k\}$ be the sequence generated by Algorithm* 2.1. *If $d^k \neq 0$ for all k, every accumulation point x^* of $\{x^k\}$ is a stationary point of problem* (1.1).

Proof. Assume that $d^k \neq 0$ for all k. It is not difficult to see that $\{x^k\} \subset \Omega$. Let x^* be any accumulation point of $\{x^k\}$. Then there exists an infinite index set K such that

$$\lim_{k\in K} x^k = x^*.$$

Taking into account that the number of distinct sets $\mathcal{A}_1^k$, $\mathcal{A}_{21}^k$, $\mathcal{A}_{22}^k$, $\mathcal{F}^k$ is finite, there exists a subsequence of $\{x^k\}_{k\in K}$ that, without loss of generality, we label again $\{x^k\}_{k\in K}$ such that the index sets $\mathcal{A}_1^k$, $\mathcal{A}_{21}^k$, $\mathcal{A}_{22}^k$, $\mathcal{F}^k$ are constant and hence we can write

$$\mathcal{A}_1^k = \mathcal{A}_1, \;\; \mathcal{A}_{21}^k = \mathcal{A}_{21}, \;\; \mathcal{A}_{22}^k = \mathcal{A}_{22} \;\text{ and }\; \mathcal{F}^k = \mathcal{F}$$

for any $k \in K$. If the sequence $\{\alpha^k\}_{k\in K}$ has a nonzero limit, we get from (3.1) that

$$\lim_{k\in K} d^k = 0.$$

Then the conclusion follows from Theorem 2.3. Otherwise, we have that $\lim_{k\in K} \alpha^k = 0$. In this case, we know by the line search condition (2.12) that for all $k \geq 1$,

$$\phi(x^k + \frac{\alpha^k}{\eta}d^k) - \phi(x^k) \geq \phi(x^k + \frac{\alpha^k}{\eta}d^k) - \phi_{\max}^k > -\delta(\frac{\alpha^k}{\eta}\|d^k\|)^2. \tag{3.8}$$

On the other hand, for sufficiently large $k \in K$, by the mean-value theorem, we have

$$\begin{aligned}
\phi(x^k + \frac{\alpha^k}{\eta}d^k) - \phi(x^k) &= \frac{\alpha^k}{\eta} g(x^k + \theta^k \frac{\alpha^k}{\eta}d^k)^T d^k \\
&\quad + \mu \sum_{i\in\mathcal{F}\cup\mathcal{A}_{22}} \left(|x_i^k + \frac{\alpha^k}{\eta}d_i^k| - |x_i^k| \right) \\
&\quad + \mu \sum_{i\in\mathcal{A}_1\cup\mathcal{A}_{21}} \left(|x_i^k + \frac{\alpha^k}{\eta}d_i^k| - |x_i^k| \right) \\
&= \frac{\alpha^k}{\eta} g(x^k + \theta^k \frac{\alpha^k}{\eta}d^k)^T d^k + \mu \frac{\alpha^k}{\eta} \sum_{i\in\mathcal{F}\cup\mathcal{A}_{22}} d_i^k \mathrm{sgn}(x_i^k) \\
&\quad + \mu \sum_{i\in\mathcal{A}_1\cup\mathcal{A}_{21}} \left(|x_i^k + \frac{\alpha^k}{\eta}d_i^k| - |x_i^k| \right) \\
&\leq \frac{\alpha^k}{\eta} g(x^k + \theta^k \frac{\alpha^k}{\eta}d^k)^T d^k + \mu \frac{\alpha^k}{\eta} \sum_{i\in\mathcal{F}\cup\mathcal{A}_{22}} d_i^k \mathrm{sgn}(x_i^k) \\
&\quad + \mu \sum_{i\in\mathcal{A}_{21}} \frac{\alpha^k}{\eta}|d_i^k| - \mu \sum_{i\in\mathcal{A}_1} \frac{\alpha^k}{\eta}|x_i^k|,
\end{aligned}$$

where $\theta^k \in (0,1)$, the second equality uses (2.11) for $i \in \mathcal{F}\cup\mathcal{A}_{22}$ and the inequality uses the triangle inequality for $i \in \mathcal{A}_{21}$ and the definition (2.2) for $i \in \mathcal{A}_1$. Substituting (3.8) into the last inequality, we obtain

$$\begin{aligned}
&(g(x^k + \theta^k \frac{\alpha^k}{\eta}d^k) - g(x^k))^T d^k + g(x^k)^T d^k \\
&\quad + \mu \sum_{i\in\mathcal{F}\cup\mathcal{A}_{22}} d_i^k \mathrm{sgn}(x_i^k) + \mu \sum_{i\in\mathcal{A}_{21}} |d_i^k| - \mu \sum_{i\in\mathcal{A}_1} |x_i^k| > -\delta \frac{\alpha^k}{\eta}\|d^k\|^2.
\end{aligned}$$

By direct calculations and (2.5), we get that the left-hand side of the above inequality

$$\begin{aligned}
&= (g(x^k+\theta^k\frac{\alpha^k}{\eta}d^k)-g(x^k))^Td^k-\sum_{i\in\mathcal{F}}\frac{(d_i^k)^2}{\lambda^k}\\
&\quad-\sum_{i\in\mathcal{A}_{22}}(d_i^k)^2+\sum_{i\in\mathcal{A}_{21}}(g_i^kd_i^k+\mu|d_i^k|)+\sum_{i\in\mathcal{A}_1}(g_i^kd_i^k-\mu|x_i^k|)\\
&= (g(x^k+\theta^k\frac{\alpha^k}{\eta}d^k)-g(x^k))^Td^k-\sum_{i\in\mathcal{F}}\frac{(d_i^k)^2}{\lambda^k}-\sum_{i\in\mathcal{A}_{22}}(d_i^k)^2\\
&\quad+\sum_{i\in\mathcal{A}_{21}}(g_i^k(-g_i^k+\mu\,\mathrm{sgn}(g_i^k))+\mu(|g_i^k|-\mu))+\sum_{i\in\mathcal{A}_1}(g_i^kd_i^k-\mu|x_i^k|)\\
&= (g(x^k+\theta^k\frac{\alpha^k}{\eta}d^k)-g(x^k))^Td^k-\sum_{i\in\mathcal{F}}\frac{(d_i^k)^2}{\lambda^k}-\sum_{i\in\mathcal{A}_{22}}(d_i^k)^2\\
&\quad-\sum_{i\in\mathcal{A}_{21}}(d_i^k)^2+\sum_{i\in\mathcal{A}_1}(g_i^kd_i^k-\mu|x_i^k|)\\
&> -\delta\frac{\alpha^k}{\eta}\|d^k\|^2. \qquad (3.9)
\end{aligned}$$

If $\rho(x^*)=0$, then the conclusion is clear. Assume that $\rho(x^*)\neq 0$. Since the sequence $\{\lambda^k\}_{k\in K}$ and $\{d^k\}_{k\in K}$ are both bounded, without loss of generality, there exist constants λ^* satisfying $\alpha_{\min}\le\lambda^*\le\alpha_{\max}$ and d^* such that

$$\lim_{k\in K}\lambda^k=\lambda^* \quad\text{and}\quad \lim_{k\in K}d^k=d^*.$$

Furthermore, using (3.1) and taking limits in K, we get from (3.9) that

$$\sum_{i\in\mathcal{F}}-\frac{(d_i^*)^2}{\lambda^*}+\sum_{i\in\mathcal{A}_2}-(d_i^*)^2+\sum_{i\in\mathcal{A}_1}(g_i(x^*)d_i^*-\mu|x_i^*|)\ge 0. \qquad (3.10)$$

Since $\lim_{k\in K}d^k=d^*$, we get that for each $i\in\mathcal{A}_1$,

$$\lim_{k\in K}d_i^k=\lim_{k\in K}-x_i^k=-x_i^*=d_i^*.$$

Proceeding with a similar analysis as those for Theorems 2.4 and 2.5, we get that for each $i\in\mathcal{A}_1$,

$$(g_i(x^*)+\mu\,\mathrm{sgn}(x_i^*))d_i^*\le 0. \qquad (3.11)$$

Notice that the left-hand side of (3.10) equals

$$\sum_{i\in\mathcal{F}}-\frac{(d_i^*)^2}{\lambda^*}+\sum_{i\in\mathcal{A}_2}-(d_i^*)^2+\sum_{i\in\mathcal{A}_1}(g_i(x^*)+\mu\,\mathrm{sgn}(x_i^*))d_i^*\le 0.$$

The last inequality together with (3.10) shows that

$$\sum_{i\in\mathcal{F}_1}-\frac{(d_i^*)^2}{\lambda^*}+\sum_{i\in\mathcal{A}_2}-(d_i^*)^2+\sum_{i\in\mathcal{A}_1}(g_i(x^*)+\mu\,\mathrm{sgn}(x_i^*))d_i^*=0. \qquad (3.12)$$

Then we have $\lim_{k\in K}d_i^k=d_i^*=0$, for all $i\in\mathcal{A}_{22}$. By the continuity of $g(x)$ and the definition of d^k, we have

$$\lim_{k\in K}\mu\,\mathrm{sgn}(x_i^k)=-g_i(x^*)\neq 0,\ \ \forall i\in\mathcal{A}_{22}.$$

Thus, we get that

$$\lim_{k\in K} \operatorname{sgn}(x_i^k) = -\operatorname{sgn}(g_i(x^*)) \neq 0, \ \ \forall i \in \mathcal{A}_{22}.$$

Proceeding with a similar analysis as those for Theorem 2.3, we get that for each $i \in \mathcal{A}_{22}$,

$$x_i^* \neq 0, \ \ \nabla\phi_i(x^*) = 0 \ \text{ or } \ x_i^* = 0, \ \ |g_i(x^*)| = \mu. \tag{3.13}$$

By the definition of d^k, $\mathcal{F}$, $\mathcal{A}_1$, $\mathcal{A}_{21}$ and the continuity of $g(x)$ and $\rho(x)$, taking limits in K, we have

$$\begin{cases} \lim\limits_{k\in K} d_i^k = \lambda^* \nabla\phi_i(x^*) \ \text{ and } \ |x_i^*| \geq \rho(x^*) > 0, & \forall i \in \mathcal{F}, \\ \lim\limits_{k\in K} d_i^k = -x_i^* \ \ |x_i^*| \leq \rho(x^*) \ \text{ and } \ |g_i(x^*)| \leq \mu, & \forall i \in \mathcal{A}_1, \\ \lim\limits_{k\in K} d_i^k = -(g_i(x^*) - \mu \operatorname{sgn}(g_i(x^*)), & \\ \qquad x_i^* = 0 \ \text{ and } \ |g_i(x^*)| \geq \mu, & \forall i \in \mathcal{A}_{21}. \end{cases} \tag{3.14}$$

By (3.11), (3.12), (3.13) and (3.14), we have the following system:

$$\begin{cases} x_i^* \neq 0 \ \text{ and } \ \nabla\phi_i(x^*) = 0, & \text{if } \ i \in \mathcal{F}, \\ x_i^* \neq 0, \ \ \nabla\phi_i(x^*) = 0 \ \text{ and } \ x_i^* = 0, \ \ |g_i(x^*)| \leq \mu, & \text{if } \ i \in \mathcal{A}_1, \\ x_i^* = 0 \ \text{ and } \ |g_i(x^*)| = \mu, & \text{if } \ i \in \mathcal{A}_{21}, \\ x_i^* \neq 0, \ \ \nabla\phi_i(x^*) = 0 \ \text{ or } \ x_i^* = 0, \ \ |g_i(x^*)| = \mu & \text{if } \ i \in \mathcal{A}_{22}, \end{cases}$$

which shows that x^* is a stationary point of problem (1.1). The proof is completed. □

The following theorem shows that $|x_i^k|$, $i \in \mathcal{G}_+(x^*)$, converges to zero monotonically decreasing for all k sufficiently large if $\{x^k\}$ converges to a stationary point x^* of problem (1.1).

Theorem 3.2. *Suppose that Assumption* 2.1 *holds and the sequence* $\{x^k\}$ *generated by Algorithm* 2.1 *converges to a stationary point* x^* *of problem* (1.1). *Then, for* $i \in \mathcal{G}_+(x^*)$, x_i^k *converges to zero after a finite number of steps or* $|x_i^k|$ *converges to zero monotonically decreasing for all* k *sufficiently large.*

Proof. Assume that $\mathcal{G}_+(x^*)$ is nonempty. Since $g(x)$ is continuous and the sequence $\{x^k\}$ converges to x^*, there exists an index $\bar{k}$ such that for all $k \geq \bar{k}$ and for all $i \in \mathcal{G}_+(x^*)$,

$$|g_i^k| < \mu. \tag{3.15}$$

Suppose that there exists $x_l^k = 0$ for $l \in \mathcal{G}_+(x^*)$ and $k \geq \bar{k}$. Then by (3.15), we have $l \in \mathcal{A}_1^k$. Furthermore, by the definition of d_l^k, we get $d_l^k = 0$. Consequently, when an index $l \in \mathcal{G}_+(x^*)$ becomes active, i.e., $x_l^k = 0$, at iterate x^k, $k > \bar{k}$, it remains active for all the subsequent iterations. On the other hand, we assume that there exists $l \in \mathcal{G}_+(x^*)$ and $x_l^k \neq 0$ for all $k \geq \bar{k}$. In this case, we have $l \in \mathcal{A}_1^k \cup \mathcal{F}^k$ for $k \geq \bar{k}$. If $l \in \mathcal{A}_1^k$ for $k \geq \bar{k}$, we have

$$|x_l^{k+1}| = |x_l^k + \alpha^k d_l^k| = (1 - \alpha^k)|x_l^k| < |x_l^k|.$$

If $l \in \mathcal{F}^k$ for $k \geq \bar{k}$, we have

$$\begin{aligned} |x_l^{k+1}| &= |x_l^k + \alpha^k d_l^k| \\ &= |x_l^k - \alpha^k \lambda^k (g_l^k + \mu \operatorname{sgn}(x_l^k))| \\ &= \begin{cases} |x_l^k - \alpha^k \lambda^k (g_l^k + \mu)|, & \text{if } x_l^k > 0; \\ |x_l^k - \alpha^k \lambda^k (g_l^k - \mu)|, & \text{if } x_l^k < 0 \end{cases} \\ &< |x_l^k|, \end{aligned}$$

where the last inequality follows from (3.15). The proof is completed. □

4. Numerical experiments

In this section, we first present some numerical experiments to illustrate how the identification technique performs. Second, we partition the rest of the experiments into two classes based on different types of ϕ. In the first class, we use our algorithm to solve ℓ_1-regularized nonconvex problem and compare it with the Matlab function **fminsearch**. In the second class, we use our algorithm to solve ℓ_1-regularized least squares, which mainly appear in compressive sensing. Specifically, we provide some numerical experiments to test the performance of the proposed method and compare it with the following three existing solvers: FPC_AS [38], GPSR_BB [25] and SpaRSA [40]. All codes are written in MATLAB R2013a and all tests described in this section were performed on a PC with Intel I5-3230 2.6GHZ CPU processor and 6G RAM memory with a Windows operating system. We implemented Algorithm 2.1 with the following parameters: $M = 5$, $\alpha_{\min} = 10^{-10}$, $\alpha_{\max} = 10^{10}$, $\delta = 10^{-2}$ and $\eta = 0.5$. The parameters in $\rho(x)$ are set to $c_1 = 0.05$, $c_2 = 1$, $\nu = 0.01$ for subsection 4.1 and $\nu = \max((\frac{1}{2})^k, 0.01)$ for other subsections. The initial point of all tested algorithms is the zero vector. The other three algorithms were run with default parameters.

Experiments in [25, 28, 38, 40] have confirmed the effectiveness of continuation. The basic idea of continuation is that, instead of solving problem (1.1) directly from scratch, one solves a sequence of problems ($x^*_{\mu_k} := \arg\min_{x \in R^n} \{\phi_{\mu_k}(x) = f(x) + \mu_k \|x\|_1\}$), where $\mu_0 > \mu_1 \cdots > \mu$, using the solution (or an approximate solution) $x^*_{\mu_{k-1}}$ as the initial estimate of the solution to the next problem. Therefore, we embedded our method in a continuation procedure. Now, we describe our continuation strategy in details. The sequence of μ_k is defined by

$$\mu_k = \begin{cases} \max(\gamma_1 \|g(0)\|_\infty, \frac{\mu}{\gamma_1}), & \text{if } k = 0, \\ \max(\gamma_1 \min(\|g_{\mathcal{A}^k}(x^k)\|_\infty, \mu_k), \mu), & \text{if } k > 0, \end{cases}$$

where $\gamma_1 \in (0, 1)$ is a constant. The main difference between our continuation strategy and the continuation strategy in [38] consists in the definition of the active set $\mathcal{A}^k$. In our implementation, at the end of iteration k, the next parameter μ_k is set to a value smaller than μ_{k-1} if the point x^k satisfies the condition

$$|\phi_{\mu_{k-1}}(x^k) - \phi_{\mu_{k-1}}(x^{k-1})| \leq \epsilon_f \max(|\phi_{\mu_{k-1}}(x^k)|, |\phi_{\mu_{k-1}}(x^{k-1})|, 1),$$

where $\epsilon_f = 0.01$.

In Table 1, we summarize a list of symbols used in the subsequent tables and figures.

TABLE 1. Summary of symbols used in all subsequent tables and figures

m, n	numbers of rows and columns of A, respectively
cpu	cpu time
nnzx	number of the nonzeros in the reovered solution
nMat	total number of matrix-vector products involving A and A^T
sgn	$\|\{i : x_i(x_s)_i < 0\}\|$, number of corresponding entries of x and x_s that are nonzero but have opposite signs
miss	$\|\{i : x_i = 0, (x_s)_i \neq 0\}\|$, number of zero entries in x with a corresponding nonzero entry in x_s
over	$\|\{i : x_i \neq 0, (x_s)_i = 0\}\|$, number of nonzero entries in x with a corresponding zero entry in x_s

4.1. **Identification ability of the active set.** In this subsection, we illustrate the identification possibility of our method with respect to sparse signals in compressed sensing. Namely, we consider the problem (1.2), where the goal is to reconstruct a length-n sparse signal from m observations, where $m < n$. The $m \times n$ measure matrix A is obtained by first filling it with independent samples of a standard Gaussian distribution and then orthonormalizing the rows. These random matrices are generated by using MATLAB command **randn**. In this experiment, we set $n = 1024$ and $m = \text{round}(0.1 * n)$. The observation b is generated by $b = Ax_s$. To generate the signal x_s, we first generated the support by randomly selecting T indices between 1 and n and then assigned a value to x_i for each i in the support by one of the following four methods:

Type 1: one (zero-one signal);
Type 2: the sign of a normally distributed random variable;
Type 3: a normally distributed random variable (Gaussian signal);
Type 4: a uniformly distributed random variable $(-1, 1)$.

Let $\mathcal{S}$ denote the support of x_s. Then for all $i \in \mathcal{S}$, for type 1, $x_i = 1$; for type 2, x_i is a Bernoulli random variable with success probability 1/2; for type 3, $x_i \sim \mathcal{N}(0, 1)$; for type 4, $x_i \sim \mathcal{U}(-1, 1)$. The regularization parameter μ is taken as $\mu = 0.01\|A^T b\|_\infty$. Notice that $\mu \geq \|A^T b\|_\infty$ and the unique minimum of (1.2) is the zero vector [30].

Random points x at different fixed distances were generated from the exact sparsest solution x_s More precisely, for each pair (ϵ, T), $\epsilon \in \{10, 1, 10^{-1}, 10^{-2}, 10^{-3}\}$ and $T \in \{1, 2, \ldots, 100\}$, we generated 100 random points x on the boundary of the set

$$\mathcal{K} = \{x : \|x - x_s\|_\infty < \epsilon\}.$$

For each component of signal x_s and the generated random points, we consider it as a nonzero component when its absolute value is great than $0.001\|x_s\|_\infty$. For each random point, we compare our approximate active sets $\mathcal{A}(x)$ with the exact active set $\mathcal{G}(x_s)$. In Table 2, we report the total number of the correctly identified $\mathcal{G}(x_s)$, the values of "sgn", "miss" and "over". Noting that, if x matches x_s in terms of support and sign, the values of "sgn", "miss" and "over" should all be zero. For types 1-2, $\epsilon = 10^{-2}, 10^{-3}$ and $T = 1 : 100$, we can see from Table 2 that the active set identification technique can accurately identify the zero components of x_s. For types 1-4, we also see that the closer the distance between the random points and the exact sparse solution x_s is, the stronger the identifiable ability of

TABLE 2. Numerical results for $n = 1024$ and $m = \text{round}(0.1 * n)$

Type 1	$T = 10$				$T = 1 : 100$			
ϵ	total	sgn	miss	over	total	sgn	miss	over
$\epsilon = 10$	0	1	0	88526	0	185	378	8498493
$\epsilon = 1$	0	0	0	56360	0	0	0	5305683
$\epsilon = 10^{-1}$	0	0	0	2597	0	0	0	0
$\epsilon = 10^{-2}$	100	0	0	0	10000	0	0	0
$\epsilon = 10^{-3}$	100	0	0	0	10000	0	0	0
Type 2	$T = 10$				$T = 1 : 100$			
ϵ	total	sgn	miss	over	total	sgn	miss	over
$\epsilon = 10$	0	1	2	88614	0	206	360	8499047
$\epsilon = 1$	0	0	0	55634	0	0	0	5306896
$\epsilon = 10^{-1}$	0	0	0	2567	0	0	0	149750
$\epsilon = 10^{-2}$	100	0	0	0	10000	0	0	0
$\epsilon = 10^{-3}$	100	0	0	0	10000	0	0	0
Type 3	$T = 10$				$T = 1 : 100$			
ϵ	total	sgn	miss	over	total	sgn	miss	over
$\epsilon = 10$	0	135	50	88648	0	51023	25326	8499649
$\epsilon = 1$	0	5	31	55891	0	2346	8617	5368878
$\epsilon = 10^{-1}$	0	0	0	3411	0	4	3173	248648
$\epsilon = 10^{-2}$	100	0	0	0	8292	0	1808	0
$\epsilon = 10^{-3}$	100	0	0	0	8400	0	1700	0
Type 4	$T = 10$				$T = 1 : 100$			
ϵ	total	sgn	miss	over	total	sgn	miss	over
$\epsilon = 10$	0	155	74	88588	0	49734	24901	8499210
$\epsilon = 1$	0	15	61	56662	0	2165	8639	5377848
$\epsilon = 10^{-1}$	0	0	0	3888	0	1	2670	276123
$\epsilon = 10^{-2}$	100	0	0	0	8205	0	1937	0
$\epsilon = 10^{-3}$	100	0	0	0	8500	0	1600	0

the the proposed active set identification technique is. In a summary, from Table 2, we can see that the proposed active set identification technique has a strong ability to accurately identify the zero components in a neighborhood of the solution.

4.2. **Test on the ℓ_1-regularized nonconvex problem.** In this subsection, we demonstrate the viability of our approach and focus on the specific problem of binary classification using the ℓ_1-regularized logistic regression. For these problems, the function $f(x)$ in (1.1) is set to

$$f(x) = \sum_{i=1}^{m} \log(1 + e^{-y_i x^T w_i}),$$

where $w_i \in \mathcal{R}^n$, $i = 1, 2, \ldots, n$, are the training samples and $y_i \in \{-1, +1\}$ are the corresponding labels. Such problems are used for training a linear classifier $x \in \mathcal{R}^n$. For more details about support vector machine problems, we refer the reader to [42]. All nine data sets used were downloaded from the web site http://www.csie.ntu..tw/∼cjlin/liblinear and have been featurewise normalized to $[-1, 1]$.

For each test problem, we first ran **fminsearch** and stored the final value of the objective function and then ran the ABB algorithm until it reaches at least the same

TABLE 3. Numerical results for the ℓ_1-regularized nonconvex problem with $\mu = 0.1$

Problem			ABB			fminsearch	
name	m	n	iter	time	nnz	time	nnz
heart	270	13	16	1.09	13	15.72	13
australian	690	14	2	5.67	14	79.25	14
fourclass	862	2	17	0.85	2	1.03	2
diabetes	768	8	13	2.45	8	29.42	8
ionosphere	351	34	50	17.37	32	218.60	34
sonar	208	60	27	6.63	60	150.46	60
svmguide3	41	22	17	0.28	2	2.06	19
liver-disorders	345	6	47	2.10	6	4.01	6
breast	683	9	21	2.24	9	28.82	9

TABLE 4. Numerical results for ℓ_1-regularized nonconvex problem with $\mu = 1$

Problem			ABB			fminsearch	
name	m	n	iter	time	nnz	time	nnz
heart	270	13	9	0.9204	13	20.47	13
australian	690	14	13	2.87	14	47.31	14
fourclass	862	2	17	0.78	2	1.45	2
diabetes	768	8	10	1.76	8	19.95	6
ionosphere	351	34	43	9.84	29	132.57	34
sonar	208	60	14	3.82	54	119.57	60
svmguide3	41	22	11	0.20	2	2.27	18
liver-disorders	345	6	44	2.04	6	3.47	5
breast	683	9	20	2.54	9	15.83	9

objective function value. Each component of the final solution $\bar{x}$ obtained by the two tested methods is considered as a nonzero component when its absolute value is greater than $0.001\|\bar{x}\|_\infty$. The detailed numerical results are presented in Tables 3 and 4, which contain the name of the tested data set, the number of training samples m and features n, the number of iterations (iter), the CPU time required in seconds (time) and the number of nonzero components of solutions (nnz).
As shown in Tables 3 and 4, ABB yields much better performance for the test data sets and is a promising method for solving ℓ_1-regularized nonconvex problems.

4.3. **Test on the ℓ_2-ℓ_1 problem.** In this experiment, we tested the matrix A with size $n = 2^{12}$ with $m = \text{round}(0.1 * n)$ or $m = \text{round}(0.2 * n)$, and $n = 2^{14}$ with $m = \text{round}(0.1 * n)$ and considered a range of degrees of sparsity: the number T of nonzero spikes in x_s ranges from 1 to 30 for each type of elements in the support. We generated random matrices A using the same strategy as that of subsection 4.1. The observation b is generated by $b = Ax_s$ and the regularization parameter μ is taken as $\mu = 0.01\,\|A^T b\|_\infty$. The above procedure yields a total of 360 problems.

For each test problem, we stop all tested algorithms if the condition

$$\frac{|\phi(x^k) - \phi(x^{k-1})|}{\max(1, |\phi(x^{k-1}), |\phi(x^k)|)} \leq 10^{-12}$$

is satisfied or the number of iterations exceeds 10^4.

Again, each component of signal x_s and the final solution obtained by each tested method is considered as a nonzero component when its absolute value is greater than $0.001\|x_s\|_\infty$. We adopt the performance profiles by Dolan and Moré [22] to evaluate the CPU time, the final function value, the relative error (rel.err= $\frac{\|x_s - x\|}{\|x_s\|}$), and the numbers of nMat and nnz. Figures 1–10 show the performance profiles of the five methods relative to the CPU time (Figures 1 and 6), the final function value (Figures 4 and 9), the relative errors (Figures 5 and 10), and the numbers of nMat and nnz (Figures 2, 3, 7, and 8). It shows that the AGE method performs best for the 360 test problems; it generally requires least CPU time, fewest numbers of nMat and obtains almost the same numbers of nnz, the final function value and the relative error.

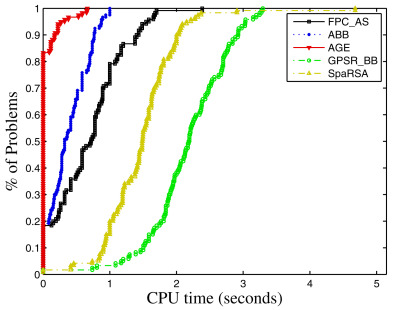

FIGURE 1. Performance profiles based on CPU time in log2 scale for $n = 2^{12}$ and $m = \text{round}(0.1n)$

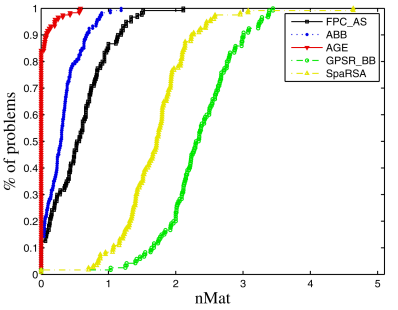

FIGURE 2. Performance profiles based on nMat in log2 scale for $n = 2^{12}$ and $m = \text{round}(0.1n)$

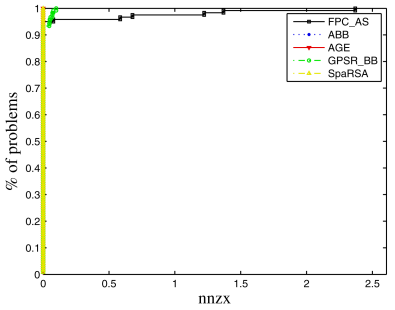

FIGURE 3. Performance profiles based on nnz in log2 scale for $n = 2^{12}$ and $m = \text{round}(0.1n)$

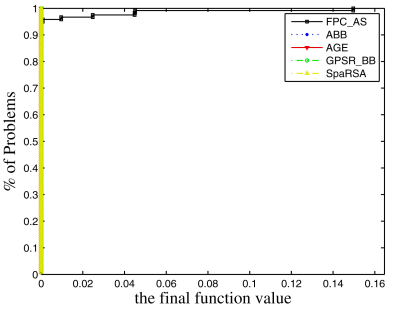

FIGURE 4. Performance profiles based on the final function value in log2 scale $n = 2^{12}$ and $m = \text{round}(0.1n)$

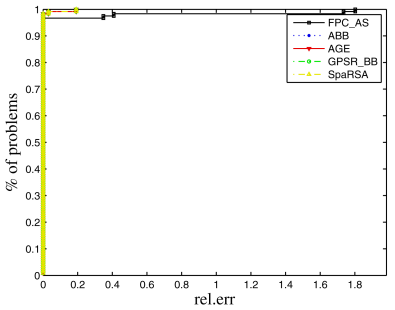

FIGURE 5. Performance profiles based on the relative error in log2 scale for $n = 2^{12}$ and $m = \text{round}(0.1n)$

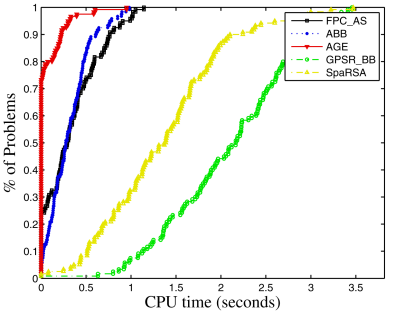

FIGURE 6. Performance profiles based on CPU time in log2 scale $n = 2^{14}$ and $m = \text{round}(0.1n)$

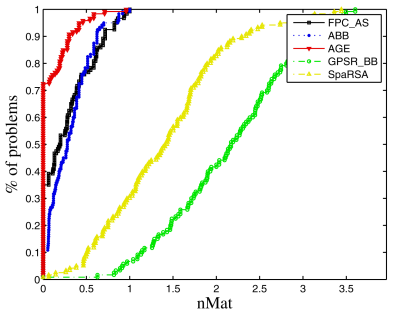

FIGURE 7. Performance profiles based on nMat in log2 scale for $n = 2^{14}$ and $m = \text{round}(0.1n)$

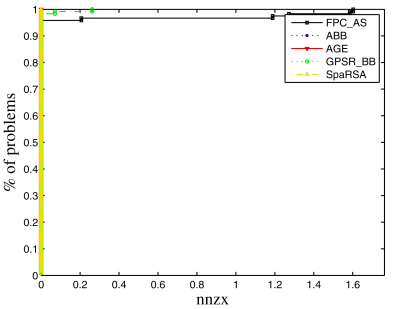

FIGURE 8. Performance profiles based on nnz in log2 scale for $n = 2^{14}$ and $m = \text{round}(0.1n)$

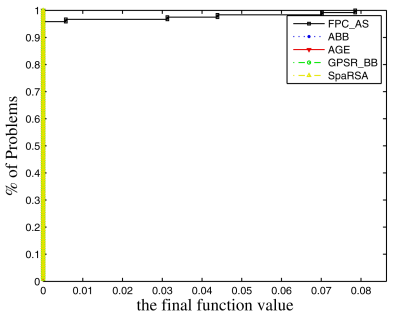

FIGURE 9. Performance profiles based on the final function value in log2 scale for $n = 2^{14}$ and $m = \text{round}(0.1n)$

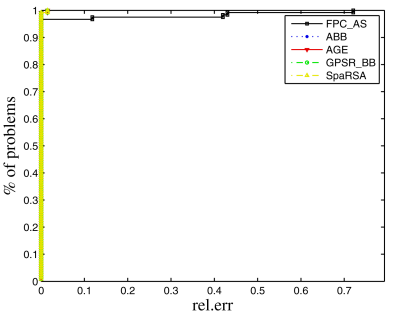

FIGURE 10. Performance profiles based on the relative error in log2 scale $n = 2^{14}$ and $m = \text{round}(0.2n)$

5. Conclusions

In the paper, we have proposed an identification function and developed an active set identification technique. The active set identification technique has a strong ability to accurately identify the zero components in a neighborhood of an isolated stationary point of (1.1) without strict complementarity conditions for the ℓ_1 optimization problem. Based on the active set identification technique, we proposed a gradient-based method for solving (1.1). To accelerate the algorithm, a subspace Barzilai-Borwein steplength and a subspace exact steplength have been developed, respectively. Under appropriate conditions, we showed that the method with the nonmonotone line search technique is globally convergent. The numerical results in Section 4 demonstrated the effectiveness of the algorithm for solving some standard ℓ_2-ℓ_1 problems.

Acknowledgments

The authors are very grateful to the two anonymous referees for their valuable suggestions and comments, which greatly helped to improve the quality of this paper.

References

[1] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, *Fast image recovery using variable splitting and constrained optimization*, IEEE Trans. Image Process. **19** (2010), no. 9, 2345–2356, DOI 10.1109/TIP.2010.2047910. MR2798930

[2] N. S. Aybat and G. Iyengar, *A first-order smoothed penalty method for compressed sensing*, SIAM J. Optim. **21** (2011), no. 1, 287–313, DOI 10.1137/090762294. MR2783217

[3] N. S. Aybat and G. Iyengar, *A first-order augmented Lagrangian method for compressed sensing*, SIAM J. Optim. **22** (2012), no. 2, 429–459, DOI 10.1137/100786721. MR2968861

[4] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal. **8** (1988), no. 1, 141–148, DOI 10.1093/imanum/8.1.141. MR967848

[5] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci. **2** (2009), no. 1, 183–202, DOI 10.1137/080716542. MR2486527

[6] A. Beck and Y. C. Eldar, *Sparsity constrained nonlinear optimization: optimality conditions and algorithms*, SIAM J. Optim. **23** (2013), no. 3, 1480–1509, DOI 10.1137/120869778. MR3080197

[7] J. M. Bioucas-Dias and M. A. T. Figueiredo, *A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration*, IEEE Trans. Image Process. **16** (2007), no. 12, 2992–3004, DOI 10.1109/TIP.2007.909319. MR2472806

[8] E. G. Birgin, J. M. Martínez, and M. Raydan, *Nonmonotone spectral projected gradient methods on convex sets*, SIAM J. Optim. **10** (2000), no. 4, 1196–1211 (electronic), DOI 10.1137/S1052623497330963. MR1777088

[9] D. Boley, *Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs*, SIAM J. Optim. **23** (2013), no. 4, 2183–2207, DOI 10.1137/120878951. MR3123832

[10] A. M. Bruckstein, D. L. Donoho, and M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, SIAM Rev. **51** (2009), no. 1, 34–81, DOI 10.1137/060657704. MR2481111

[11] J. V. Burke and J. J. Moré, *On the identification of active constraints*, SIAM J. Numer. Anal. **25** (1988), no. 5, 1197–1211, DOI 10.1137/0725068. MR960873

[12] E. Candes and J. Romberg, ℓ_1-magic: A collection of MATLAB Routines for Solving the Convex Optimization Programs Central to Compressive Sampling 2006 [Online]. Available: www.acm.caltech. edu/l1magic/

[13] W. Cheng and Z. Chen, *Nonmonotone spectral method for large-scale symmetric nonlinear equations*, Numer. Algorithms **62** (2013), no. 1, 149–162, DOI 10.1007/s11075-012-9572-z. MR3009560

[14] Y. H. Dai, *A new analysis on the Barzilai-Borwein gradient method*, J. Oper. Res. Soc. China. **1** (2013), pp. 187–198.

[15] Y.-H. Dai and R. Fletcher, *Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming*, Numer. Math. **100** (2005), no. 1, 21–47, DOI 10.1007/s00211-004-0569-y. MR2129700

[16] Y.-H. Dai, W. W. Hager, K. Schittkowski, and H. Zhang, *The cyclic Barzilai-Borwein method for unconstrained optimization*, IMA J. Numer. Anal. **26** (2006), no. 3, 604–627, DOI 10.1093/imanum/drl006. MR2241317

[17] Y.-H. Dai and L.-Z. Liao, **R**-*linear convergence of the Barzilai and Borwein gradient method*, IMA J. Numer. Anal. **22** (2002), no. 1, 1–10, DOI 10.1093/imanum/22.1.1. MR1880051

[18] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math. **57** (2004), no. 11, 1413–1457, DOI 10.1002/cpa.20042. MR2077704

[19] G. Davis, S. Mallat, and M. Avellaneda, *Adaptive greedy approximations*, Constr. Approx. **13** (1997), no. 1, 57–98, DOI 10.1007/s003659900033. MR1424364

[20] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory **52** (2006), no. 4, 1289–1306, DOI 10.1109/TIT.2006.871582. MR2241189

[21] D. L. Donoho, M. Elad, and V. N. Temlyakov, *Stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Trans. Inform. Theory **52** (2006), no. 1, 6–18, DOI 10.1109/TIT.2005.860430. MR2237332

[22] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Math. Program. **91** (2002), no. 2, Ser. A, 201–213, DOI 10.1007/s101070100263. MR1875515

[23] M. Elad, B. Matalon, and M. Zibulevsky, *Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization*, Appl. Comput. Harmon. Anal. **23** (2007), no. 3, 346–367, DOI 10.1016/j.acha.2007.02.002. MR2362407

[24] M. A. T. Figueiredo and R. D. Nowak, *An EM algorithm for wavelet-based image restoration*, IEEE Trans. Image Process. **12** (2003), no. 8, 906–916, DOI 10.1109/TIP.2003.814255. MR2008658

[25] 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 J. Sel. Top. Signal Process. **1** (2007), pp. 586–597.

[26] K. Fountoulakis and J. Gondzio, *A second-order method for strongly convex ℓ_1-regularization problems*, Math. Program. **156** (2016), no. 1-2, Ser. A, 189–219, DOI 10.1007/s10107-015-0875-4. MR3459199

[27] L. Grippo, F. Lampariello, and S. Lucidi, *A nonmonotone line search technique for Newton's method*, SIAM J. Numer. Anal. **23** (1986), no. 4, 707–716, DOI 10.1137/0723046. MR849278

[28] E. T. Hale, W. Yin, and Y. Zhang, *Fixed-point continuation for l_1-minimization: methodology and convergence*, SIAM J. Optim. **19** (2008), no. 3, 1107–1130, DOI 10.1137/070698920. MR2460734

[29] W. W. Hager, D. T. Phan, and H. Zhang, *Gradient-based methods for sparse recovery*, SIAM J. Imaging Sci. **4** (2011), no. 1, 146–165, DOI 10.1137/090775063. MR2792408

[30] S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, *An interior-point method for large-scale ℓ_1-regularized least squares*, IEEE J. Sel. Top. Signal Process. **1** (2007), pp. 606–617.

[31] N. S. Keskar, J. Nocedal, F. Öztoprak, and A. Wächter, *A second-order method for convex ℓ_1-regularized optimization with active set prediction*, http://arxiv.org/abs/1505.04315

[32] Y. Nesterov, *Gradient methods for minimizing composite objective function*, 2007, CORE Discussion Paper 2007/76 [Online]. Available: http://www.optimization-online.org/DB_HTML/2007/09/1784.html

[33] R. T. Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, NJ, 1970. MR0274683

[34] S. M. Robinson, *Linear convergence of epsilon-subgradient descent methods for a class of convex functions*, Math. Program. **86** (1999), no. 1, Ser. A, 41–50, DOI 10.1007/s101070050078. MR1712472

[35] M. Saunders, PDCO: Primal-dual interior method for convex objectives 2002 [Online]. Available: http://www.stanford.edu/group/SOL/software/pdco.html

[36] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program. **117** (2009), no. 1-2, Ser. B, 387–423, DOI 10.1007/s10107-007-0170-0. MR2421312

[37] J. A. Tropp, *Greed is good: algorithmic results for sparse approximation*, IEEE Trans. Inform. Theory **50** (2004), no. 10, 2231–2242, DOI 10.1109/TIT.2004.834793. MR2097044

[38] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang, *A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation*, SIAM J. Sci. Comput. **32** (2010), no. 4, 1832–1857, DOI 10.1137/090747695. MR2678081

[39] Z. Wen, W. Yin, H. Zhang, and D. Goldfarb, *On the convergence of an active-set method for ℓ_1 minimization*, Optim. Methods Softw. **27** (2012), no. 6, 1127–1146, DOI 10.1080/10556788.2011.591398. MR2955298

[40] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Process. **57** (2009), no. 7, 2479–2493, DOI 10.1109/TSP.2009.2016892. MR2650165

[41] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, *Bregman iterative algorithms for l_1-minimization with applications to compressed sensing*, SIAM J. Imaging Sci. **1** (2008), no. 1, 143–168, DOI 10.1137/070703983. MR2475828

[42] G. X. Yuan, C. H. Ho, and C. J. Lin, *Recent advances of large-scale linear classification*, Proceedings of the IEEE **100** (2012), 2584–2603.

College of Computer, Dongguan University of Technology, Dongguan 523000, People's Republic of China

E-mail address: chengwanyou@sina.com

LSEC, ICMSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, People's Republic of China

E-mail address: dyh@lsec.cc.ac.cn