Sparse Approximation via Penalty Decomposition Methods *

Zhaosong Lu† Yong Zhang ‡

February 19, 2012

Abstract

In this paper we consider sparse approximation problems, that is, general l_0 minimization problems with the l_0-"norm" of a vector being a part of constraints or objective function. In particular, we first study the first-order optimality conditions for these problems. We then propose penalty decomposition (PD) methods for solving them in which a sequence of penalty subproblems are solved by a block coordinate descent (BCD) method. Under some suitable assumptions, we establish that any accumulation point of the sequence generated by the PD methods satisfies the first-order optimality conditions of the problems. Furthermore, for the problems in which the l_0 part is the only nonconvex part, we show that such an accumulation point is a local minimizer of the problems. In addition, we show that any accumulation point of the sequence generated by the BCD method is a saddle point of the penalty subproblem. Moreover, for the problems in which the l_0 part is the only nonconvex part, we establish that such an accumulation point is a local minimizer of the penalty subproblem. Finally, we test the performance of our PD methods by applying them to sparse logistic regression, sparse inverse covariance selection, and compressed sensing problems. The computational results demonstrate that our methods generally outperform the existing methods in terms of solution quality and/or speed.

Key words: l_0 minimization, penalty decomposition methods, block coordinate descent method, compressed sensing, sparse logistic regression, sparse inverse covariance selection

1 Introduction

Nowadays, there are numerous applications in which sparse solutions are concerned. For example, in compressed sensing, a large sparse signal is decoded by using a relatively small number of linear measurements, which can be formulated as finding a sparse solution to a system of linear equalities and/or inequalities. The similar ideas have also been widely used in linear regression. Recently, sparse inverse covariance selection becomes an important tool in discovering the conditional independence in graphical models. One popular approach for sparse inverse covariance selection is to find an approximate sparse inverse covariance while maximizing the log-likelihood (see, for example, [16]). Similarly, sparse logistic regression has been proposed as a promising method for feature selection in classification problems in which a sparse solution is sought to minimize the average logistic loss (see, for example, [41]).

*This work was supported in part by NSERC Discovery Grant.

†Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada. (email: zhaosong@sfu.ca).

‡Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada. (email: yza30@sfu.ca).

Mathematically, all these applications can be formulated into the following l_0 minimization problems:

$$\min_{x\in\mathcal{X}}\{f(x) : g(x) \le 0,\ h(x) = 0,\ \|x_J\|_0 \le r\}, \tag{1}$$

$$\min_{x\in\mathcal{X}}\{f(x) + \nu\|x_J\|_0 : g(x) \le 0,\ h(x) = 0\} \tag{2}$$

for some integer $r \ge 0$ and $\nu \ge 0$ controlling the sparsity of the solution, where $\mathcal{X}$ is a closed convex set in the n-dimensional Euclidean space $\Re^n$, $f : \Re^n \to \Re$, $g : \Re^n \to \Re^m$ and $h : \Re^n \to \Re^p$ are continuously differentiable functions, and $\|x_J\|_0$ denotes the cardinality of the subvector formed by the entries of x indexed by J. Some algorithms are proposed for solving special cases of these problems. For example, the iterative hard thresholding algorithms [26, 5, 6] and matching pursuit algorithms [38, 52] are developed for solving the l_0-regularized least squares problems arising in compressed sensing, but they cannot be applied to the general l_0 minimization problems (1) and (2). In the literature, one popular approach for dealing with (1) and (2) is to replace $\|\cdot\|_0$ by the l_1-norm $\|\cdot\|_1$ and solve the resulting relaxation problems instead (see, for example, [14, 41, 10, 51]). For some applications such as compressed sensing, it has been shown in [8] that under some suitable assumptions this approach is capable of solving (1) and (2). Recently, another relaxation approach has been proposed to solve problems (1) and (2) in which $\|\cdot\|_0$ is replaced by l_p-"norm" $\|\cdot\|_p$ for some $p \in (0, 1)$ (see, for example, [9, 11, 12]). In general, it is not clear about the solution quality of these approaches. Indeed, for the example given in the Appendix, the l_p relaxation approach for $p \in (0, 1]$ fails to recover the sparse solution.

In this paper we propose penalty decomposition (PD) methods for solving problems (1) and (2) in which a sequence of penalty subproblems are solved by a block coordinate descent (BCD) method. Under some suitable assumptions, we establish that any accumulation point of the sequence generated by the PD method satisfies the first-order optimality conditions of (1) and (2). Furthermore, when h's are affine, and f and g's are convex, we show that such an accumulation point is a local minimizer of the problems. In addition, we show that any accumulation point of the sequence generated by the BCD method is a saddle point of the penalty subproblem. Moreover, when h's are affine, and f and g's are convex, we establish that such an accumulation point is a local minimizer of the penalty subproblem. Finally, we test the performance of our PD methods by applying them to sparse logistic regression, sparse inverse covariance selection, and compressed sensing problems. The computational results demonstrate that our methods generally outperform the existing methods in terms of solution quality and/or speed.

The rest of this paper is organized as follows. In Subsection 1.1, we introduce the notation that is used throughout the paper. In Section 2, we establish the first-order optimality conditions for general l_0 minimization problems. In Section 3, we study a class of special l_0 minimization problems. We develop the PD methods for general l_0 minimization problems in Section 4 and establish some convergence results for them. In Section 5, we conduct numerical experiments to test the performance of our PD methods for solving sparse logistic regression, sparse inverse covariance selection, and compressed sensing problems. Finally, we present some concluding remarks in section 6.

1.1 Notation

In this paper, the symbols $\Re^n$ and $\Re^n_+$ denote the n-dimensional Euclidean space and the nonnegative orthant of $\Re^n$, respectively. Given a vector $v \in \Re^n$, the nonnegative part of v is denoted by $v^+ = \max(v, 0)$, where the maximization operates entry-wise. For any real vector, $\|\cdot\|_0$ and $\|\cdot\|$ denote the cardinality (i.e., the number of nonzero entries) and the Euclidean norm of the vector, respectively.

Given an index set $L \subseteq \{1,\ldots,n\}$, $|L|$ denotes the size of L, and the elements of L are denoted by $L(1),\ldots,L(|L|)$, which are always arranged in ascending order. x_L denotes the subvector formed by the entries of x indexed by L. Likewise, X_L denotes the submatrix formed by the columns of X indexed by L. In addition, For any two sets A and B, the subtraction of A and B is given by $A \setminus B = \{x \in A : x \notin B\}$. Given a closed set $C \subseteq \Re^n$, let $\mathcal{N}_C(x)$ and $\mathcal{T}_C(x)$ denote the normal and tangent cones of C at any $x \in C$, respectively. The space of all $m \times n$ matrices with real entries is denoted by $\Re^{m\times n}$, and the space of symmetric $n \times n$ matrices is be denoted by $\mathcal{S}^n$. We denote by I the identity matrix, whose dimension should be clear from the context. If $X \in \mathcal{S}^n$ is positive semidefinite (resp., definite), we write $X \succeq 0$ (resp., $X \succ 0$). The cone of positive semidefinite (resp., definite) matrices is denoted by $\mathcal{S}^n_+$ (resp., $\mathcal{S}^n_{++}$). $\mathscr{D}$ is an operator which maps a vector to a diagonal matrix whose diagonal consists of the vector. Given an $n \times n$ matrix X, $\widetilde{\mathscr{D}}(X)$ denotes a diagonal matrix whose ith diagonal element is X_{ii} for $i = 1,\ldots,n$.

2 First-order optimality conditions

In this section we study the first-order optimality conditions for problems (1) and (2). In particular, we first discuss the first-order necessary conditions for them. Then we study the first-order sufficient conditions for them when the l_0 part is the only nonconvex part.

We now establish the first-order necessary optimality conditions for problems (1) and (2).

Theorem 2.1 *Assume that x^* is a local minimizer of problem (1). Let $J^* \subseteq J$ be an index set with $|J^*| = r$ such that $x^*_j = 0$ for all $j \in \bar{J}^*$, where $\bar{J}^* = J \setminus J^*$. Suppose that the following Robinson condition*

$$\left\{ \begin{bmatrix} g'(x^*)d - v \\ h'(x^*)d \\ (I_{\bar{J}^*})^T d \end{bmatrix} : d \in \mathcal{T}_{\mathcal{X}}(x^*), v \in \Re^m, v_i \le 0, i \in \mathcal{A}(x^*) \right\} = \Re^m \times \Re^p \times \Re^{|J|-r} \tag{3}$$

holds, where $g'(x^*)$ and $h'(x^*)$ denote the Jacobian of the functions $g = (g_1,\ldots,g_m)$ and $h = (h_1,\ldots,h_p)$ at x^*, respectively, and

$$\mathcal{A}(x^*) = \{1 \le i \le m : g_i(x^*) = 0\}. \tag{4}$$

Then, there exists $(\lambda^*, \mu^*, z^*) \in \Re^m \times \Re^p \times \Re^n$ together with x^* satisfying

$$\begin{aligned} -\nabla f(x^*) - \nabla g(x^*)\lambda^* - \nabla h(x^*)\mu^* - z^* \in \mathcal{N}_{\mathcal{X}}(x^*), \\ \lambda^*_i \ge 0,\ \lambda^*_i g_i(x^*) = 0,\ i = 1,\ldots,m; \qquad z^*_j = 0,\ j \in \bar{J} \cup J^*. \end{aligned} \tag{5}$$

where $\bar{J}$ is the complement of J in $\{1,\ldots,n\}$.

Proof. By the assumption that x^* is a local minimizer of problem (1), one can observe that x^* is also a local minimizer of the following problem:

$$\min_{x\in\mathcal{X}} \{f(x) : g(x) \le 0,\ h(x) = 0,\ x_{\bar{J}^*} = 0\}. \tag{6}$$

Using this observation, (3) and Theorem 3.25 of [47], we see that the conclusion holds. ∎

Theorem 2.2 *Assume that x^* is a local minimizer of problem (2). Let $J^* = \{j \in J : x_j^* \neq 0\}$ and $\bar{J}^* = J \setminus J^*$. suppose that the following Robinson condition*

$$\left\{ \begin{bmatrix} g'(x^*)d - v \\ h'(x^*)d \\ (I_{\bar{J}^*})^T d \end{bmatrix} : d \in \mathcal{T}_{\mathcal{X}}(x^*), v \in \Re^m, v_i \leq 0, i \in \mathcal{A}(x^*) \right\} = \Re^m \times \Re^p \times \Re^{|\bar{J}^*|} \tag{7}$$

holds, where $\mathcal{A}(x^*)$ is defined in (4)*. Then, there exists $(\lambda^*, \mu^*, z^*) \in \Re^m \times \Re^p \times \Re^n$ together with x^* satisfying* (5)*.*

Proof. It is not hard to observe that x^* is a local minimizer of problem (2) if and only if x^* is a local minimizer of problem (6). Using this observation, (7) and Theorem 3.25 of [47], we see that the conclusion holds. ∎

We next establish the first-order sufficient optimality conditions for problems (1) and (2) when the l_0 part is the only nonconvex part.

Theorem 2.3 *Assume that h's are affine functions, and f and g's are convex functions. Let x^* be a feasible point of problem (1), and let $\mathcal{J}^* = \{J^* \subseteq J : |J^*| = r, x_j^* = 0, \forall j \in J \setminus J^*\}$. Suppose that for any $J^* \in \mathcal{J}^*$, there exists some $(\lambda^*, \mu^*, z^*) \in \Re^m \times \Re^p \times \Re^n$ such that* (5) *holds. Then, x^* is a local minimizer of problem (1).*

Proof. It follows from the above assumptions and Theorem 3.34 of [47] that x^* is a minimizer of problem (6) for all $\bar{J}^* \in \{J \setminus J^* : J^* \in \mathcal{J}^*\}$. Hence, there exists $\epsilon > 0$ such that $f(x) \geq f(x^*)$ for all $x \in \cup_{J^* \in \mathcal{J}^*} \mathcal{O}_{J^*}(x^*; \epsilon)$, where

$$\mathcal{O}_{J^*}(x^*; \epsilon) = \{x \in \mathcal{X} : g(x) \leq 0,\ h(x) = 0,\ x_{\bar{J}^*} = 0,\ \|x - x^*\| < \epsilon\}$$

with $\bar{J}^* = J \setminus J^*$. One can observe from (1) that for any $x \in \mathcal{O}(x^*; \epsilon)$, where

$$\mathcal{O}(x^*; \epsilon) = \{x \in \mathcal{X} : g(x) \leq 0,\ h(x) = 0,\ \|x_J\|_0 \leq r,\ \|x - x^*\| < \epsilon\},$$

there exists $J^* \in \mathcal{J}^*$ such that $x \in \mathcal{O}_{J^*}(x^*; \epsilon)$ and hence $f(x) \geq f(x^*)$. It implies that the conclusion holds. ∎

Theorem 2.4 *Assume that h's are affine functions, and f and g's are convex functions. Let x^* be a feasible point of problem (2), and let $J^* = \{j \in J : x_j^* \neq 0\}$. Suppose that for such J^*, there exists some $(\lambda^*, \mu^*, z^*) \in \Re^m \times \Re^p \times \Re^n$ such that* (5) *holds. Then, x^* is a local minimizer of problem (2).*

Proof. By virtue of the above assumptions and Theorem 3.34 of [47], we know that x^* is a minimizer of problem (6) with $\bar{J}^* = J \setminus J^*$. Also, we observe that any point is a local minimizer of problem (2) if and only if it is a local minimizer of problem (6). It then implies that x^* is a local minimizer of (2). ∎

Remark. The second-order necessary or sufficient optimality conditions for problems (1) and (2) can be similarly established as above. ∎

3 A class of special l_0 minimization

In this section we show that a class of special l_0 minimization problems have closed-form solutions, which can be used to develop penalty decomposition methods for solving general l_0 minimization problems.

Proposition 3.1 *Let $\mathcal{X}_i \subseteq \Re$ and $\phi_i : \Re \to \Re$ for $i = 1, \ldots, n$ be given. Suppose that r is a positive integer and $0 \in \mathcal{X}_i$ for all i. Consider the following l_0 minimization problem:*

$$\min \left\{ \phi(x) = \sum_{i=1}^{n} \phi_i(x_i) : \|x\|_0 \leq r,\ x \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \right\}. \tag{8}$$

Let $\tilde{x}_i^* \in \operatorname{Arg\,min}\{\phi_i(x_i) : x_i \in \mathcal{X}_i\}$ and $I^* \subseteq \{1, \ldots, n\}$ be the index set corresponding to r largest values of $\{v_i^*\}_{i=1}^n$, where $v_i^* = \phi_i(0) - \phi_i(\tilde{x}_i^*)$ for $i = 1, \ldots, n$. Then, x^* is an optimal solution of problem (8), where x^* is defined as follows:

$$x_i^* = \begin{cases} \tilde{x}_i^* & \text{if } i \in I^*; \\ 0 & \text{otherwise}, \end{cases} \qquad i = 1, \ldots, n.$$

Proof. By the assumption that $0 \in \mathcal{X}_i$ for all i, and the definitions of x^*, $\tilde{x}^*$ and I^*, we see that $x^* \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and $\|x^*\|_0 \leq r$. Hence, x^* is a feasible solution of (8). It remains to show that $\phi(x) \geq \phi(x^*)$ for any feasible point x of (8). Indeed, let x be arbitrarily chosen such that $\|x\|_0 \leq r$ and $x \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, and let $L = \{i : x_i \neq 0\}$. Clearly, $|L| \leq r = |I^*|$. Let $\bar{I}^*$ and $\bar{L}$ denote the complement of I^* and L in $\{1, \ldots, n\}$, respectively. It then follows that

$$|\bar{L} \cap I^*| \ = \ |I^*| - |I^* \cap L| \ \geq \ |L| - |I^* \cap L| \ = \ |L \cap \bar{I}^*|.$$

In view of the definitions of x^*, $\tilde{x}^*$, I^*, $\bar{I}^*$, L and $\bar{L}$, we further have

$$\begin{aligned} \phi(x) - \phi(x^*) &= \textstyle\sum_{i \in L \cap I^*} (\phi_i(x_i) - \phi_i(x_i^*)) + \sum_{i \in \bar{L} \cap \bar{I}^*} (\phi_i(x_i) - \phi_i(x_i^*)) \\ &\quad + \textstyle\sum_{i \in \bar{L} \cap I^*} (\phi_i(x_i) - \phi_i(x_i^*)) + \sum_{i \in L \cap \bar{I}^*} (\phi_i(x_i) - \phi_i(x_i^*)), \\ &= \textstyle\sum_{i \in L \cap I^*} (\phi_i(x_i) - \phi_i(\tilde{x}_i^*)) + \sum_{i \in \bar{L} \cap \bar{I}^*} (\phi_i(0) - \phi_i(0)) \\ &\quad + \textstyle\sum_{i \in \bar{L} \cap I^*} (\phi_i(0) - \phi_i(\tilde{x}_i^*)) + \sum_{i \in L \cap \bar{I}^*} (\phi_i(x_i) - \phi_i(0)), \\ &\geq \textstyle\sum_{i \in \bar{L} \cap I^*} (\phi_i(0) - \phi_i(\tilde{x}_i^*)) + \sum_{i \in L \cap \bar{I}^*} (\phi_i(\tilde{x}_i^*) - \phi_i(0)), \\ &= \textstyle\sum_{i \in \bar{L} \cap I^*} (\phi_i(0) - \phi_i(\tilde{x}_i^*)) - \sum_{i \in L \cap \bar{I}^*} (\phi_i(0) - \phi_i(\tilde{x}_i^*)) \ \geq \ 0, \end{aligned}$$

where the last inequality follows from the definition of I^* and the relation $|\bar{L} \cap I^*| \geq |L \cap \bar{I}^*|$. Thus, we see that $\phi(x) \geq \phi(x^*)$ for any feasible point x of (8), which implies that the conclusion holds. ∎

It is straightforward to establish the following result.

Proposition 3.2 *Let $\mathcal{X}_i \subseteq \Re$ and $\phi_i : \Re \to \Re$ for $i = 1, \ldots, n$ be given. Suppose that $\nu \geq 0$ and $0 \in \mathcal{X}_i$ for all i. Consider the following l_0 minimization problem:*

$$\min \left\{ \nu \|x\|_0 + \sum_{i=1}^{n} \phi_i(x_i) : \quad x \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \right\}. \tag{9}$$

Let $\tilde{x}_i^* \in \operatorname{Arg\,min}\{\phi_i(x_i) : x_i \in \mathcal{X}_i\}$ and $v_i^* = \phi_i(0) - \nu - \phi_i(\tilde{x}_i^*)$ for $i = 1, \ldots, n$. Then, x^* is an optimal solution of problem (9), where x^* is defined as follows:

$$x_i^* = \begin{cases} \tilde{x}_i^* & \text{if } v_i^* \geq 0; \\ 0 & \text{otherwise}, \end{cases} \qquad i = 1, \ldots, n.$$

4 Penalty decomposition methods for general l_0 minimization

In this section we propose penalty decomposition (PD) methods for solving general l_0 minimization problems (1) and (2) and establish their convergence. Throughout this section, we make the following assumption for problems (1) and (2).

Assumption 1 *Problems (1) and (2) are feasible, and moreover, at least a feasible solution, denoted by $x^{\rm feas}$, is known.*

This assumption will be used to design the PD methods with nice convergence properties. It can be dropped, but the theoretical convergence of the corresponding PD methods may become weaker. We shall also mention that, for numerous real applications, $x^{\rm feas}$ is readily available or can be observed from the physical background of problems. For example, all application problems discussed in Section 5 have a trivial feasible solution. On the other hand, for some problems which do not have a trivial feasible solution, one can always approximate them by the problems which have a trivial feasible solution. For instance, problem (1) can be approximately solved as the following problem:

$$\min_{x\in\mathcal{X}}\{f(x)+\rho(\|u^+\|^2+\|v\|^2) : g(x)-u\le 0,\ h(x)-v=0,\ \|x_J\|_0\le r\}$$

for some large ρ. The latter problem has a trivial feasible solution when $\mathcal{X}$ is sufficiently simple.

4.1 Penalty decomposition method for problem (1)

In this subsection we propose a PD method for solving problem (1) and establish its convergence.

We observe that (1) can be equivalently reformulated as

$$\min_{x\in\mathcal{X},y\in\mathcal{Y}}\{f(x):\ g(x)\le 0,\ h(x)=0,\ x_J-y=0\}, \tag{10}$$

where

$$\mathcal{Y}=\{y\in\Re^{|J|} : \|y\|_0\le r\}.$$

The associated quadratic penalty function is defined as follows:

$$q_\varrho(x,y)=f(x)+\frac{\varrho}{2}(\|[g(x)]^+\|^2+\|h(x)\|^2+\|x_J-y\|^2)\quad \forall x\in\mathcal{X}, y\in\mathcal{Y} \tag{11}$$

for some penalty parameter $\varrho>0$.

We are now ready to propose a PD method for solving problem (10) (or equivalently, (1)) in which each penalty subproblem is approximately solved by a block coordinate descent (BCD) method.

Penalty decomposition method for (1):

Let $\{\epsilon_k\}$ be a positive decreasing sequence. Let $\varrho_0>0$, $\sigma>1$ be given. Choose an arbitrary $y_0^0\in\mathcal{Y}$ and a constant $\Upsilon\ge\max\{f(x^{\rm feas}),\min_{x\in\mathcal{X}} q_{\varrho_0}(x,y_0^0)\}$. Set $k=0$.

1) Set $l=0$ and apply the BCD method to find an approximate solution $(x^k,y^k)\in\mathcal{X}\times\mathcal{Y}$ for the penalty subproblem

$$\min\{q_{\varrho_k}(x,y):\ x\in\mathcal{X},\ y\in\mathcal{Y}\} \tag{12}$$

by performing steps 1a)-1d):

1a) Solve $x_{l+1}^k \in \operatorname{Arg}\min_{x\in\mathcal{X}} q_{\varrho_k}(x, y_l^k)$.

1b) Solve $y_{l+1}^k \in \operatorname{Arg}\min_{y\in\mathcal{Y}} q_{\varrho_k}(x_{l+1}^k, y)$.

1c) Set $(x^k, y^k) := (x_{l+1}^k, y_{l+1}^k)$. If (x^k, y^k) satisfies

$$\|\mathcal{P}_{\mathcal{X}}(x^k - \nabla_x q_{\varrho_k}(x^k, y^k)) - x^k\| \le \epsilon_k, \tag{13}$$

then go to step 2).

1d) Set $l \leftarrow l + 1$ and go to step 1a).

2) Set $\varrho_{k+1} := \sigma\varrho_k$.

3) If $\min_{x\in\mathcal{X}} q_{\varrho_{k+1}}(x, y^k) > \Upsilon$, set $y_0^{k+1} := x^{\text{feas}}$. Otherwise, set $y_0^{k+1} := y^k$.

4) Set $k \leftarrow k + 1$ and go to step 1).

end

Remark. The condition (13) will be used to establish the global convergence of the above method. It may not be easily verifiable unless $\mathcal{X}$ is simple. On the other hand, we observe that the sequence $\{q_{\varrho_k}(x_l^k, y_l^k)\}$ is non-increasing for any fixed k. In practice, it is thus reasonable to terminate the BCD method based on the progress of $\{q_{\varrho_k}(x_l^k, y_l^k)\}$. Another practical termination criterion for the BCD method is based on the relative change of the sequence $\{(x_l^k, y_l^k)\}$, that is,

$$\max\left\{\frac{\|x_l^k - x_{l-1}^k\|_\infty}{\max(\|x_l^k\|_\infty, 1)}, \frac{\|y_l^k - y_{l-1}^k\|_\infty}{\max(\|y_l^k\|_\infty, 1)}\right\} \le \epsilon_I \tag{14}$$

for some $\epsilon_I > 0$. In addition, we can terminate the outer iterations of the PD method once

$$\|x^k - y^k\|_\infty \le \epsilon_O \tag{15}$$

for some $\epsilon_O > 0$. Given that problem (12) is nonconvex, the BCD method may converge to a stationary point. To enhance the performance of the BCD method, one may execute it multiple times by restarting from a suitable perturbation of the current best approximate solution. For example, at the kth outer iteration, let (x^k, y^k) be the current best approximate solution of (12) found by the BCD method, and let $r_k = \|y^k\|_0$. Assume that $r_k > 1$. Before starting the $(k+1)$th outer iteration, one can re-apply the BCD method starting from $y_0^k \in \operatorname{Arg}\min\{\|y - y^k\| : \|y\|_0 \le r_k - 1\}$ and obtain a new approximate solution $(\tilde{x}^k, \tilde{y}^k)$ of (12). If $q_{\varrho_k}(\tilde{x}^k, \tilde{y}^k)$ is "sufficiently" smaller than $q_{\varrho_k}(x^k, y^k)$, one can set $(x^k, y^k) := (\tilde{x}^k, \tilde{y}^k)$ and repeat the above process. Otherwise, one can terminate the kth outer iteration and start the next outer iteration. Finally, it follows from Proposition 3.1 that the subproblem in step 1b) has a closed-form solution. ∎

We next establish a convergence result regarding the inner iterations of the above PD method. In particular, we will show that an approximate solution (x^k, y^k) of problem (12) satisfying (13) can be found by the BCD method described in steps 1a)-1d). For notational convenience, we omit the index k from (12) and consider the BCD method for solving the problem

$$\min\{q_\varrho(x, y) : x \in \mathcal{X},\ y \in \mathcal{Y}\} \tag{16}$$

instead. Accordingly, we rename the iterates of the above BCD method and present it as follows.

Block coordinate descent method for (16):

Choose an arbitrary initial point $y^0 \in \mathcal{Y}$. Set $l = 0$.

1) Solve $x^{l+1} \in \operatorname{Arg}\min_{x\in\mathcal{X}} q_\varrho(x, y^l)$.

2) Solve $y^{l+1} \in \operatorname{Arg}\min_{y\in\mathcal{Y}} q_\varrho(x^{l+1}, y)$.

3) Set $l \leftarrow l + 1$ and go to step 1).

end

Lemma 4.1 *Suppose that $(x^*, y^*) \in \Re^n \times \Re^{|J|}$ is a saddle point of problem* (16)*, that is,*

$$x^* \in \operatorname{Arg}\min_{x\in\mathcal{X}} q_\varrho(x, y^*), \quad y^* \in \operatorname{Arg}\min_{y\in\mathcal{Y}} q_\varrho(x^*, y). \tag{17}$$

Furthermore, assume that h's are affine functions, and f and g's are convex functions. Then, (x^*, y^*) is a local minimizer of problem (16)*.*

Proof. Let $K = \{i : y_i^* \neq 0\}$, and let h_x, h_y be any two vectors such that $x^* + h_x \in \mathcal{X}$, $y^* + h_y \in \mathcal{Y}$ and $|(h_y)_i| < |y_i^*|$ for all $i \in K$. Claim that

$$(y^* - x_J^*)^T h_y = 0. \tag{18}$$

If $\|x_J^*\|_0 > r$, we observe from the second relation of (17) and Proposition 3.1 that $\|y^*\|_0 = r$ and $y_i^* = x_{J(i)}^*$ for all $i \in K$, which, together with $y^* + h_y \in \mathcal{Y}$ and $|(h_y)_i| < |y_i^*|$ for all $i \in K$, implies that $(h_y)_i = 0$ for all $i \notin K$ and hence (18) holds. On the other hand, if $\|x_J^*\|_0 \leq r$, one can observe that $y^* = x_J^*$ and thus (18) also holds. In addition, by the assumption that h's are affine functions, and f and g's are convex functions, we know that q_ϱ is convex. It then follows from the first relation of (17) and the first-order optimality condition that $[\nabla_x q_\varrho(x^*, y^*)]^T h_x \geq 0$. Using this relation along with (18) and the convexity of q_ϱ, we have

$$\begin{aligned} q_\varrho(x^* + h_x, y^* + h_y) &\geq q_\varrho(x^*, y^*) + [\nabla_x q_\varrho(x^*, y^*)]^T h_x + [\nabla_y q_\varrho(x^*, y^*)]^T h_y \\ &= q_\varrho(x^*, y^*) + [\nabla_x q_\varrho(x^*, y^*)]^T h_x + \varrho(y^* - x_J^*)^T h_y \;\geq\; q_\varrho(x^*, y^*), \end{aligned}$$

which together with the above choice of h_x and h_y implies that (x^*, y^*) is a local minimizer of (16). ∎

Theorem 4.2 *Let $\{(x^l, y^l)\}$ be the sequence generated by the above BCD method, and let $\epsilon > 0$ be given. Suppose that (x^*, y^*) is an accumulation point of $\{(x^l, y^l)\}$. Then the following statements hold:*

(a) (x^*, y^*) is a saddle point of problem (16)*.*

(b) There exists some $l > 0$ such that

$$\|\mathcal{P}_\mathcal{X}(x^l - \nabla_x q_\varrho(x^l, y^l)) - x^l\| < \epsilon.$$

(c) *Furthermore, if h's are affine functions, and f and g's are convex functions, then (x^*, y^*) is a local minimizer of problem* (16).

Proof. We first show that statement (a) holds. Indeed, one can observe that

$$q_\varrho(x^{l+1}, y^l) \leq q_\varrho(x, y^l) \quad \forall x \in \mathcal{X}, \tag{19}$$
$$q_\varrho(x^l, y^l) \leq q_\varrho(x^l, y) \quad \forall y \in \mathcal{Y}. \tag{20}$$

It follows that

$$q_\varrho(x^{l+1}, y^{l+1}) \leq q_\varrho(x^{l+1}, y^l) \leq q_\varrho(x^l, y^l) \quad \forall l \geq 1. \tag{21}$$

Hence, the sequence $\{q_\varrho(x^l, y^l)\}$ is non-increasing. Since (x^*, y^*) is an accumulation point of $\{(x^l, y^l)\}$, there exists a subsequence L such that $\lim_{l\in L\to\infty}(x^l, y^l) = (x^*, y^*)$. We then observe that $\{q_\varrho(x^l, y^l)\}_{l\in L}$ is bounded, which together with the monotonicity of $\{q_\varrho(x^l, y^l)\}$ implies that $\{q_\varrho(x^l, y^l)\}$ is bounded below and hence $\lim_{l\to\infty} q_\varrho(x^l, y^l)$ exists. This observation, (21) and the continuity of $q_\varrho(\cdot, \cdot)$ yield

$$\lim_{l\to\infty} q_\varrho(x^{l+1}, y^l) = \lim_{l\to\infty} q_\varrho(x^l, y^l) = \lim_{l\in L\to\infty} q_\varrho(x^l, y^l) = q_\varrho(x^*, y^*).$$

Using these relations, the continuity of $q_\varrho(\cdot, \cdot)$, and taking limits on both sides of (19) and (20) as $l \in L \to \infty$, we have

$$q_\varrho(x^*, y^*) \leq q_\varrho(x, y^*) \quad \forall x \in \mathcal{X}, \tag{22}$$
$$q_\varrho(x^*, y^*) \leq q_\varrho(x^*, y) \quad \forall y \in \mathcal{Y}. \tag{23}$$

In addition, from the definition of $\mathcal{Y}$, we know that $\|y^l\|_0 \leq r$, which immediately implies $\|y^*\|_0 \leq r$. Also, $x^* \in \mathcal{X}$ due to the closedness of $\mathcal{X}$. This together with (22) and (23) implies that (x^*, y^*) is a saddle point of (16) and hence statement (a) holds. Using (22) and the first-order optimality condition, we have

$$\|\mathcal{P}_\mathcal{X}(x^* - \nabla_x q_\varrho(x^*, y^*)) - x^*\| = 0.$$

By the continuity of $\mathcal{P}_\mathcal{X}(\cdot)$ and $\nabla_x q_\varrho(\cdot, \cdot)$, and the relation $\lim_{l\in L\to\infty}(x^l, y^l) = (x^*, y^*)$, one can see that

$$\lim_{l\in L\to\infty} \|\mathcal{P}_\mathcal{X}(x^l - \nabla_x q_\varrho(x^l, y^l)) - x^l\| = 0,$$

and hence, statement (b) immediately follows. In addition, statement (c) holds due to statement (a) and Lemma 4.1. ∎

The following theorem establishes the convergence of the outer iterations of the PD method for solving problem (1). In particular, we show that under some suitable assumption, any accumulation point of the sequence generated by the PD method satisfies the first-order optimality conditions of (1). Moreover, when the l_0 part is the only nonconvex part, we show that under some assumption, the accumulation point is a local minimizer of (1).

Theorem 4.3 *Assume that $\epsilon_k \to 0$. Let $\{(x^k, y^k)\}$ be the sequence generated by the above PD method, $I_k = \{i_1^k, \ldots, i_r^k\}$ be a set of r distinct indices in $\{1, \ldots, |J|\}$ such that $(y^k)_i = 0$ for any $i \notin I_k$, and let $J_k = \{J(i) : i \in I_k\}$. Suppose that the level set $\mathcal{X}_\Upsilon := \{x \in \mathcal{X} : f(x) \leq \Upsilon\}$ is compact. Then, the following statements hold:*

(a) *The sequence $\{(x^k, y^k)\}$ is bounded.*

(b) *Suppose (x^*, y^*) is an accumulation point of $\{(x^k, y^k)\}$. Then, $x^* = y^*$ and x^* is a feasible point of problem (1). Moreover, there exists a subsequence K such that $\{(x^k, y^k)\}_{k\in K} \to (x^*, y^*)$, $I_k = I^*$ and $J_k = J^* := \{J(i) : i \in I^*\}$ for some index set $I^* \subseteq \{1, \ldots, |J|\}$ when $k \in K$ is sufficiently large.*

(c) *Let x^*, K and J^* be defined above, and let $\bar{J}^* = J \setminus J^*$. Suppose that the Robinson condition (3) holds at x^* for such $\bar{J}^*$. Then, $\{(\lambda^k, \mu^k, \varpi^k)\}_{k\in K}$ is bounded, where*

$$\lambda^k = \varrho_k[g(x^k)]^+, \quad \mu^k = \varrho_k h(x^k), \quad \varpi^k = \varrho_k(x^k_J - y^k). \tag{24}$$

Moreover, each accumulation point $(\lambda^*, \mu^*, \varpi^*)$ of $\{(\lambda^k, \mu^k, \varpi^k)\}_{k\in K}$ together with x^* satisfies the first-order optimality conditions (5) with $z^*_j = \varpi^*_i$ for all $j = J(i) \in \bar{J}^*$. Further, if $\|x^*_J\|_0 = r$, h's are affine functions, and f and g's are convex functions, then x^* is a local minimizer of problem (1).

Proof. In view of (11) and our choice of y^k_0 that is specified in step 3), one can observe that

$$f(x^k) + \frac{\varrho_k}{2}(\|[g(x^k)]^+\|^2 + \|h(x^k)\|^2 + \|x^k_J - y^k\|^2) = q_{\varrho_k}(x^k, y^k) \le \min_{x\in\mathcal{X}} q_{\varrho_k}(x, y^k_0) \le \Upsilon \quad \forall k. \tag{25}$$

It immediately implies that $\{x^k\} \subseteq \mathcal{X}_\Upsilon$, and hence, $\{x^k\}$ is bounded. Moreover, we can obtain from (25) that

$$\|x^k_J - y^k\|^2 \le 2[\Upsilon - f(x^k)]/\varrho_k \le 2[\Upsilon - \min_{x\in\mathcal{X}_\Upsilon} f(x)]/\varrho_0,$$

which together with the boundedness of $\{x^k\}$ yields that $\{y^k\}$ is bounded. Therefore, statement (a) follows. We next show that statement (b) also holds. Since (x^*, y^*) is an accumulation point of $\{(x^k, y^k)\}$, there exists a subsequence $\{(x^k, y^k)\}_{k\in\bar{K}} \to (x^*, y^*)$. Recall that I_k is an index set. It follows that $\{(i^k_1, \ldots, i^k_r)\}_{k\in\bar{K}}$ is bounded for all k. Thus there exists a subsequence $K \subseteq \bar{K}$ such that $\{(i^k_1, \ldots, i^k_r)\}_{k\in K} \to (i^*_1, \ldots, i^*_r)$ for some r distinct indices $i^*_1, \ldots, i^*_r$. Since $i^k_1, \ldots, i^k_r$ are r distinct integers, one can easily conclude that $(i^k_1, \ldots, i^k_r) = (i^*_1, \ldots, i^*_r)$ for sufficiently large $k \in K$. Let $I^* = \{i^*_1, \ldots, i^*_r\}$. It then follows that $I_k = I^*$ and $J_k = J^*$ when $k \in K$ is sufficiently large, and moreover, $\{(x^k, y^k)\}_{k\in K} \to (x^*, y^*)$. Therefore, statement (b) holds. Finally, we show that statement (c) holds. Indeed, let s^k be the vector such that

$$\mathcal{P}_\mathcal{X}(x^k - \nabla_x q_{\varrho_k}(x^k, y^k)) = x^k + s^k.$$

It then follows from (13) that $\|s^k\| \le \epsilon_k$ for all k, which together with $\lim_{k\to\infty} \epsilon_k = 0$ implies $\lim_{k\to\infty} s^k = 0$. By a well-known property of the projection map $\mathcal{P}_\mathcal{X}$, we have

$$(x - x^k - s^k)^T[x^k - \nabla_x q_{\varrho_k}(x^k, y^k) - x^k - s^k] \ \le \ 0, \ \forall x \in \mathcal{X}.$$

Hence, we obtain that

$$-\nabla_x q_{\varrho_k}(x^k, y^k) - s^k \in \mathcal{N}_\mathcal{X}(x^k + s^k). \tag{26}$$

Using this relation, (26), (24) and the definition of q_ϱ, we have

$$-\nabla f(x^k) - \nabla g(x^k)\lambda^k - \nabla h(x^k)\mu^k - I_J\varpi^k - s^k \in \mathcal{N}_\mathcal{X}(x^k + s^k). \tag{27}$$

We now claim that $\{(\lambda^k, \mu^k, \varpi^k)\}_{k\in K}$ is bounded. Suppose for contradiction that it is unbounded. By passing to a subsequence if necessary, we can assume that $\{\|(\lambda^k, \mu^k, \varpi^k)\|\}_{k\in K} \to \infty$. Let $(\bar{\lambda}^k, \bar{\mu}^k, \bar{\varpi}^k) =$

$(\lambda^k, \mu^k, \varpi^k)/\|(\lambda^k, \mu^k, \varpi^k)\|$. Without loss of generality, we assume that $\{(\bar{\lambda}^k, \bar{\mu}^k, \bar{\varpi}^k)\}_{k\in K} \to (\bar{\lambda}, \bar{\mu}, \bar{\varpi})$ (otherwise, one can consider its convergent subsequence). Clearly, $\|(\bar{\lambda}, \bar{\mu}, \bar{\varpi})\| = 1$. Dividing both sides of (27) by $\|(\lambda^k, \mu^k, \varpi^k)\|$, taking limits as $k \in K \to \infty$, and using the relation $\lim_{k\in K\to\infty} s^k = 0$ and the semicontinuity of $\mathcal{N}_{\mathcal{X}}(\cdot)$, we obtain that

$$-\nabla g(x^*)\bar{\lambda} - \nabla h(x^*)\bar{\mu} - I_J\bar{\varpi} \in \mathcal{N}_{\mathcal{X}}(x^*). \tag{28}$$

We can see from (4) and (24) that $\bar{\lambda} \in \Re^m_+$, and $\bar{\lambda}_i = 0$ for $i \notin \mathcal{A}(x^*)$. Also, from Proposition 3.1 and the definitions of y_k, I_k and J_k, one can observe that $x^k_{J_k} = y^k_{I_k}$ and hence $\varpi^k_{I_k} = 0$. In addition, we know from statement (b) that $I_k = I^*$ when $k \in K$ is sufficiently large. Hence, $\bar{\varpi}_{I^*} = 0$. Since Robinson's condition (3) is satisfied at x^*, there exist $d \in \mathcal{T}_{\mathcal{X}}(x^*)$ and $v \in \Re^m$ such that $v_i \le 0$ for $i \in \mathcal{A}(x^*)$, and

$$g'(x^*)d - v = -\bar{\lambda}, \qquad h'(x^*)d = -\bar{\mu}, \qquad (I_{\bar{J}^*})^T d = -\bar{\varpi}_{\bar{I}^*},$$

where $\bar{I}^*$ is the complement of I^* in $\{1, \ldots, |J|\}$. Recall that $\bar{\lambda} \in \Re^m_+$, $\bar{\lambda}_i = 0$ for $i \notin \mathcal{A}(x^*)$, and $v_i \le 0$ for $i \in \mathcal{A}(x^*)$. Hence, $v^T\bar{\lambda} \le 0$. In addition, since $\bar{\varpi}_{I^*} = 0$, one has $I_J\bar{\varpi} = I_{\bar{J}^*}\bar{\varpi}_{\bar{I}^*}$. Using these relations, (28), and the facts that $d \in \mathcal{T}_{\mathcal{X}}(x^*)$ and $\bar{\varpi}_{I^*} = 0$, we have

$$\begin{aligned}
\|\bar{\lambda}\|^2 + \|\bar{\mu}\|^2 + \|\bar{\varpi}\|^2 &= -[(-\bar{\lambda})^T\bar{\lambda} + (-\bar{\mu})^T\bar{\mu} + (-\bar{\varpi}_{\bar{I}^*})^T\bar{\varpi}_{\bar{I}^*}] \\
&= -[(g'(x^*)d - v)^T\bar{\lambda} + (h'(x^*)d)^T\bar{\mu} + ((I_{\bar{J}^*})^T d)^T\bar{\varpi}_{\bar{I}^*}] \\
&= d^T(-\nabla g(x^*)\bar{\lambda} - \nabla h(x^*)\bar{\mu} - I_J\bar{\varpi}) + v^T\bar{\lambda} \ \le \ 0.
\end{aligned}$$

It yields $(\bar{\lambda}, \bar{\mu}, \bar{\varpi}) = (0, 0, 0)$, which contradicts the identity $\|(\bar{\lambda}, \bar{\mu}, \bar{\varpi})\| = 1$. Therefore, the subsequence $\{(\lambda^k, \mu^k, \varpi^k)\}_{k\in K}$ is bounded. Let $(\lambda^*, \mu^*, \varpi^*)$ be an accumulation point of $\{(\lambda^k, \mu^k, \varpi^k)\}_{k\in K}$. By passing to a subsequence if necessary, we can assume that $(\lambda^k, \mu^k, \varpi^k) \to (\lambda^*, \mu^*, \varpi^*)$ as $k \in K \to \infty$. Taking limits on both sides of (27) as $k \in K \to \infty$, and using the relations $\lim_{k\in K\to\infty} s^k = 0$ and the semicontinuity of $\mathcal{N}_{\mathcal{X}}(\cdot)$, we see that the first relation of (5) holds with $z^* = I_J\varpi^*$. By a similar argument as above, one can show that $\varpi^*_{I^*} = 0$. This together with the definitions of J^* and $\bar{J}^*$ implies that z^* satisfies

$$z^*_j = \begin{cases} 0 & \text{if } j \in \bar{J} \cup J^*, \\ \varpi^*_i & \text{if } j = J(i) \in \bar{J}^*, \end{cases}$$

where $\bar{J}$ is the complement of J in $\{1, \ldots, n\}$. In addition, we see from (24) that $\lambda^k_i \ge 0$ and $\lambda^k_i g_i(x^k) = 0$ for all i, which immediately lead to the second relation of (5). Hence, $(\lambda^*, \mu^*, \varpi^*)$ together with x^* satisfies (5). Suppose now that $\|x^*_J\|_0 = r$. Then, $\mathcal{J}^* = \{\tilde{J}^* \subseteq J : |\tilde{J}^*| = r, x^*_j = 0, \forall j \notin \tilde{J}^*\} = \{J^*\}$. Therefore, the assumptions of Theorem 2.3 hold. It then follows from Theorem 2.3 that x^* is a local minimizer of (1). ∎

4.2 Penalty decomposition method for problem (2)

In this subsection we propose a PD method for solving problem (2) and establish some convergence results for it.

We observe that problem (2) can be equivalently reformulated as

$$\min_{x\in\mathcal{X}, y\in\Re^{|J|}} \{f(x) + \nu\|y\|_0 : \ g(x) \le 0, \ h(x) = 0, \ \ x_J - y = 0\}. \tag{29}$$

The associated quadratic penalty function for (29) is defined as

$$p_\varrho(x, y) := f(x) + \nu\|y\|_0 + \frac{\varrho}{2}(\|[g(x)]^+\|^2 + \|h(x)\|^2 + \|x_J - y\|^2) \quad \forall x \in \mathcal{X}, y \in \Re^{|J|} \tag{30}$$

for some penalty parameter $\varrho > 0$.

We are now ready to present the PD method for solving (29) (or, equivalently, (2)) in which each penalty subproblem is approximately solved by a BCD method.

Penalty decomposition method for (2):

Let $\{\epsilon_k\}$ be a positive decreasing sequence. Let $\varrho_0 > 0$, $\sigma > 1$ be given, and let q_ϱ be defined in (11). Choose an arbitrary $y_0^0 \in \Re^{|J|}$ and a constant Υ such that $\Upsilon \geq \max\{f(x^{\rm feas})+\nu\|x^{\rm feas}\|_0, \min_{x\in\mathcal{X}} p_{\varrho_0}(x, y_0^0)\}$. Set $k = 0$.

1) Set $l = 0$ and apply the BCD method to find an approximate solution $(x^k, y^k) \in \mathcal{X} \times \Re^{|J|}$ for the penalty subproblem
$$\min\{p_{\varrho_k}(x, y) : \ x \in \mathcal{X}, \ y \in \Re^{|J|}\} \tag{31}$$
by performing steps 1a)-1d):

   1a) Solve $x_{l+1}^k \in \mathrm{Arg}\min_{x\in\mathcal{X}} p_{\varrho_k}(x, y_l^k)$.

   1b) Solve $y_{l+1}^k \in \mathrm{Arg}\min_{y\in\Re^{|J|}} p_{\varrho_k}(x_{l+1}^k, y)$.

   1c) Set $(x^k, y^k) := (x_{l+1}^k, y_{l+1}^k)$. If (x^k, y^k) satisfies
   $$\|\mathcal{P}_\mathcal{X}(x^k - \nabla_x q_{\varrho_k}(x^k, y^k)) - x^k\| \leq \epsilon_k, \tag{32}$$
   then go to step 2).

   1d) Set $l \leftarrow l + 1$ and go to step 1a).

2) Set $\varrho_{k+1} := \sigma\varrho_k$.

3) If $\min_{x\in\mathcal{X}} p_{\varrho_{k+1}}(x, y^k) > \Upsilon$, set $y_0^{k+1} := x^{\rm feas}$. Otherwise, set $y_0^{k+1} := y^k$.

4) Set $k \leftarrow k + 1$ and go to step 1).

end

Remark. The practical termination criteria proposed in Subsection 4.1 can also be applied to this PD method. In addition, one can apply a similar strategy as mentioned in Subsection 4.1 to enhance the performance of the BCD method for solving (31). Finally, in view of Proposition 3.2, the BCD subproblem in step 1b) has a closed-form solution. ■

We next establish a convergence result regarding the inner iterations of the above PD method. In particular, we will show that an approximate solution (x^k, y^k) of problem (31) satisfying (32) can be found by the BCD method described in steps 1a)-1d). For convenience of presentation, we omit the index k from (31) and consider the BCD method for solving the following problem:
$$\min\{p_\varrho(x, y) : \ x \in \mathcal{X}, \ y \in \Re^{|J|}\} \tag{33}$$
instead. Accordingly, we rename the iterates of the above BCD method. We can observe that the resulting BCD method is the same as the one presented in Subsection 4.1 except that p_ϱ and $\Re^{|J|}$ replace q_ϱ and $\mathcal{Y}$, respectively. For the sake of brevity, we omit the presentation of this BCD method.

Lemma 4.4 *Suppose that $(x^*, y^*) \in \Re^n \times \Re^{|J|}$ is a saddle point of problem* (33)*, that is,*

$$x^* \in \mathrm{Arg}\min_{x\in\mathcal{X}} p_\varrho(x, y^*), \quad y^* \in \mathrm{Arg}\min_{y\in\Re^{|J|}} p_\varrho(x^*, y). \tag{34}$$

Furthermore, assume that h's are affine functions, and f and g's are convex functions. Then, (x^*, y^*) is a local minimizer of problem (33)*.*

Proof. Let $K = \{i : y_i^* \neq 0\}$, and let h_x, h_y be any two vectors such that $x^* + h_x \in \mathcal{X}$, $|(h_y)_i| < \nu/(\rho|x^*_{J(i)}| + 1)$ for any $i \notin K$ and $|(h_y)_i| < |y_i^*|$ for all $i \in K$. We observe from the second relation of (34) and Proposition 3.2 that $y_i^* = x^*_{J(i)}$ for all $i \in K$. Also, for the above choice of h_y, one has $y_i^* + (h_y)_i \neq 0$ for all $i \in K$. Hence, $\|y_i^* + (h_y)_i\|_0 = \|y_i^*\|_0$ for every $i \in K$. Using these relations and the definition of h_y, we can see that

$$\rho(y^* - x_J^*)^T h_y + \nu\|y^* + h_y\|_0 - \nu\|y^*\|_0 \;=\; -\rho\sum_{i\notin K} x^*_{J(i)}(h_y)_i + \nu\sum_{i\notin K}\|(h_y)_i\|_0 \;\geq\; 0. \tag{35}$$

In addition, by the assumption that h's are affine functions, and f and g's are convex functions, we know that q_ϱ is convex, where q_ϱ is defined in (11). Also, notice that $p_\varrho(x, y) = q_\varrho(x, y) + \nu\|y\|_0$. It then follows from the first relation of (34) and the first-order optimality condition that $[\nabla_x q_\varrho(x^*, y^*)]^T h_x \geq 0$. Using this relation along with (35) and the convexity of q_ϱ, we have

$$\begin{aligned} p_\varrho(x^* + h_x, y^* + h_y) &= q_\varrho(x^* + h_x, y^* + h_y) + \nu\|y^* + h_y\|_0 \\ &\geq q_\varrho(x^*, y^*) + [\nabla_x q_\varrho(x^*, y^*)]^T h_x + [\nabla_y q_\varrho(x^*, y^*)]^T h_y + \nu\|y^* + h_y\|_0 \\ &\geq p_\varrho(x^*, y^*) + \varrho(y^* - x_J^*)^T h_y + \nu\|y^* + h_y\|_0 - \nu\|y^*\|_0 \;\geq\; p_\varrho(x^*, y^*), \end{aligned}$$

which together with the above choice of h_x and h_y implies that (x^*, y^*) is a local minimizer of (33). ∎

Theorem 4.5 *Let $\{(x^l, y^l)\}$ be the sequence generated by the above BCD method, and let $\epsilon > 0$ be given. Suppose that (x^*, y^*) is an accumulation point of $\{(x^l, y^l)\}$. Then the following statements hold:*

(a) (x^*, y^*) is a saddle point of problem (33)*.*

(b) There exists some $l > 0$ such that

$$\|\mathcal{P}_\mathcal{X}(x^l - \nabla_x q_\varrho(x^l, y^l)) - x^l\| < \epsilon,$$

where the function q_ϱ is defined in (11)*.*

(c) Furthermore, if h's are affine functions, and f and g's are convex functions, then (x^*, y^*) is a local minimizer of problem (33)*.*

Proof. We first show that statement (a) holds. Indeed, one can observe that

$$p_\varrho(x^{l+1}, y^l) \;\leq\; p_\varrho(x, y^l) \quad \forall x \in \mathcal{X}, \tag{36}$$

$$p_\varrho(x^l, y^l) \;\leq\; p_\varrho(x^l, y) \quad \forall y \in \Re^{|J|}. \tag{37}$$

It follows that

$$p_\varrho(x^{l+1}, y^{l+1}) \;\leq\; p_\varrho(x^{l+1}, y^l) \;\leq\; p_\varrho(x^l, y^l) \quad \forall l \geq 1. \tag{38}$$

Hence, the sequence $\{p_\varrho(x^l, y^l)\}$ is non-increasing. Since (x^*, y^*) is an accumulation point of $\{(x^l, y^l)\}$, there exists a subsequence L such that $\lim_{l\in L\to\infty}(x^l, y^l) = (x^*, y^*)$, and moreover, $x^* \in \mathcal{X}$ due to the closedness of $\mathcal{X}$. We then observe from (30) that $\{p_\varrho(x^l, y^l)\}_{l\in L}$ is bounded, which together with the monotonicity of $\{p_\varrho(x^l, y^l)\}$ implies that $\{p_\varrho(x^l, y^l)\}$ is bounded below and hence $\lim_{l\to\infty} p_\varrho(x^l, y^l)$ exists. This observation and (38) yield

$$\lim_{l\to\infty} p_\varrho(x^l, y^l) = \lim_{l\to\infty} p_\varrho(x^{l+1}, y^l). \tag{39}$$

For notational convenience, let

$$F(x) := f(x) + \frac{\varrho}{2}(\|[g(x)]^+\|^2 + \|h(x)\|^2).$$

It then follows from (30) that

$$p_\varrho(x, y) = F(x) + \nu\|y\|_0 + \frac{\varrho}{2}\|x_J - y\|^2, \quad \forall x \in \mathcal{X}, y \in \Re^{|J|}. \tag{40}$$

Since $\lim_{l\in L} y^l = y^*$, one has $\|y^l\|_0 \geq \|y^*\|_0$ for sufficiently large $l \in L$. Using this relation, (37) and (40), we obtain that, when $l \in L$ is sufficiently large,

$$p_\varrho(x^l, y) \;\geq\; p_\varrho(x^l, y^l) \;=\; F(x^l) + \nu\|y^l\|_0 + \frac{\varrho}{2}\|x_J^l - y^l\|^2 \;\geq\; F(x^l) + \nu\|y^*\|_0 + \frac{\varrho}{2}\|x_J^l - y^l\|^2.$$

Upon taking limits on both sides of the above inequality as $l \in L \to \infty$ and using the continuity of F, one has

$$p_\varrho(x^*, y) \;\geq\; F(x^*) + \nu\|y^*\|_0 + \frac{\varrho}{2}\|x_J^* - y^*\|^2 \;=\; p_\varrho(x^*, y^*), \quad \forall y \in \Re^{|J|}. \tag{41}$$

In addition, it follows from (36) and (40) that

$$\begin{array}{rcl} F(x) + \frac{1}{2}\|x_J - y^l\|^2 & = & p_\varrho(x, y^l) - \nu\|y^l\|_0 \;\geq\; p_\varrho(x^{l+1}, y^l) - \nu\|y^l\|_0 \\ & = & F(x^{l+1}) + \frac{1}{2}\|x_J^{l+1} - y^l\|^2, \quad \forall x \in \mathcal{X}. \end{array} \tag{42}$$

Since $\{\|y^l\|_0\}_{l\in L}$ is bounded, there exists a subsequence $\bar{L} \subseteq L$ such that $\lim_{l\in\bar{L}\to\infty}\|y^l\|_0$ exists. Then we have

$$\begin{array}{l} \lim\limits_{l\in\bar{L}\to\infty} F(x^{l+1}) + \frac{1}{2}\|x_J^{l+1} - y^l\|^2 \;=\; \lim\limits_{l\in\bar{L}\to\infty} p_\varrho(x^{l+1}, y^l) - \nu\|y^l\|_0 \\ = \lim\limits_{l\in\bar{L}\to\infty} p_\varrho(x^{l+1}, y^l) - \nu \lim\limits_{l\in\bar{L}\to\infty}\|y^l\|_0 \;=\; \lim\limits_{l\in\bar{L}\to\infty} p_\varrho(x^l, y^l) - \nu \lim\limits_{l\in\bar{L}\to\infty}\|y^l\|_0 \\ = \lim\limits_{l\in\bar{L}\to\infty} p_\varrho(x^l, y^l) - \nu\|y^l\|_0 = \lim\limits_{l\in\bar{L}\to\infty} F(x^l) + \frac{1}{2}\|x_J^l - y^l\|^2 \;=\; F(x^*) + \frac{1}{2}\|x_J^* - y^*\|^2, \end{array}$$

where the third equality is due to (39). Using this relation and taking limits on both sides of (42) as $l \in \bar{L} \to \infty$, we further have

$$F(x) + \frac{1}{2}\|x_J - y^*\|^2 \;\geq\; F(x^*) + \frac{1}{2}\|x_J^* - y^*\|^2, \quad \forall x \in \mathcal{X},$$

which together with (30) yields

$$p_\varrho(x, y^*) \;\geq\; p_\varrho(x^*, y^*), \quad \forall x \in \mathcal{X}.$$

This relation along with (41) implies that (x^*, y^*) is a saddle point of (33) and hence statement (a) holds. Statement (b) can be similarly proved as that of Theorem 4.2. In addition, statement (c) holds due to statement (a) and Lemma 4.4. ∎

Remark. A similar result as in statement (c) is recently established in [58] for the BCD method when applied to solve the *unconstrained* problem:

$$\min_{x,y} \frac{1}{2}\|Ax - b\|^2 + \frac{\varrho}{2}\|Wx - y\|^2 + \sum_i \nu_i \|y_i\|_0 \tag{43}$$

under the assumption that $A^T A \succ 0$, $W^T W = I$, $\varrho > 0$, and $\nu_i \geq 0$ for all i. The proof of [58] strongly relies on this assumption and the fact that the BCD subproblems have closed-form solutions. We believe that it cannot be extended to problem (33). In addition, it is not hard to observe that problem (43) can be equivalently reformulated into a problem in the form of (33) and thus the convergence of the BCD method for (43) directly follows from Theorem 4.5 above. ∎

We next establish the convergence of the outer iterations of the PD method for solving problem (2). In particular, we show that under some suitable assumption, any accumulation point of the sequence generated by the PD method satisfies the first-order optimality conditions of (2). Moreover, when the l_0 part is the only nonconvex part, we show that the accumulation point is a local minimizer of (2).

Theorem 4.6 *Assume that $\epsilon_k \to 0$. Let $\{(x^k, y^k)\}$ be the sequence generated by the above PD method. Suppose that the level set $\mathcal{X}_\Upsilon := \{x \in \mathcal{X} : f(x) \leq \Upsilon\}$ is compact. Then, the following statements hold:*

(a) The sequence $\{(x^k, y^k)\}$ is bounded;

(b) Suppose (x^*, y^*) is an accumulation point of $\{(x^k, y^k)\}$. Then, $x^* = y^*$ and x^* is a feasible point of problem (2).

(c) Let (x^*, y^*) be defined above. Suppose that $\{(x^k, y^k)\}_{k\in K} \to (x^*, y^*)$ for some subsequence K. Let $J^* = \{j \in J : x_j^* \neq 0\}$, $\bar{J}^* = J \setminus J^*$. Assume that the Robinson condition (7) holds at x^* for such $\bar{J}^*$. Then, $\{(\lambda^k, \mu^k, \varpi^k)\}_{k\in K}$ is bounded, where

$$\lambda^k = \varrho_k[g(x^k)]^+, \quad \mu^k = \varrho_k h(x^k), \quad \varpi^k = \varrho_k(x_J^k - y^k).$$

Moreover, each accumulation point $(\lambda^*, \mu^*, \varpi^*)$ of $\{(\lambda^k, \mu^k, \varpi^k)\}_{k\in K}$ together with x^* satisfies the first-order optimality condition (5) with $z_j^* = \varpi_i^*$ for all $j = J(i) \in \bar{J}^*$. Further, if h's are affine functions, and f and g's are convex functions, then x^* is a local minimizer of problem (2).

Proof. Statement (a) and (b) can be similarly proved as those of Theorem 4.3. We now show that statement (c) holds. Let $I^* = \{i : J(i) \in J^*\}$. From Proposition 3.2 and the definitions of y^k and J^*, we can observe that $y_{I^*}^k = x_{J^*}^k$ when $k \in K$ is sufficiently large. Hence, $\varpi_{I^*}^k = 0$ for sufficiently large $k \in K$. The rest of the proof for the first two conclusions of this statement is similar to that of statement (c) of Theorem 4.3. The last conclusion of this statement holds due to its second conclusion and Theorem 2.4. ∎

5 Numerical results

In this section, we conduct numerical experiments to test the performance of our PD methods proposed in Section 4 by applying them to sparse logistic regression, sparse inverse covariance selection, and compressed sensing problems. The codes of all the methods implemented in this section are written in Matlab, which are available online at www.math.sfu.ca/∼zhaosong. All experiments are performed in Matlab 7.11.0 (2010b) on a workstation with an Intel Xeon E5410 CPU (2.33 GHz) and 8GB RAM running Red Hat Enterprise Linux (kernel 2.6.18).

5.1 Sparse logistic regression problem

In this subsection, we apply the PD method studied in Subsection 4.1 to solve sparse logistic regression problem, which has numerous applications in machine learning, computer vision, data mining, bioinformatics, and neural signal processing (see, for example, [3, 54, 32, 43, 22, 44]).

Given n samples $\{z^1, \ldots, z^n\}$ with p features, and n binary outcomes $b_1, \ldots, b_n$, let $a^i = b_i z^i$ for $i = 1, \ldots, n$. The *average logistic loss* function is defined as

$$l_{\rm avg}(v, w) := \sum_{i=1}^{n} \theta(w^T a^i + v b_i)/n$$

for some model variables $v \in \Re$ and $w \in \Re^p$, where θ is the *logistic loss* function

$$\theta(t) := \log(1 + \exp(-t)).$$

Then the *sparse logistic regression* problem can be formulated as

$$\min_{v,w} \{l_{\rm avg}(v, w) : \|w\|_0 \le r\}, \tag{44}$$

where $r \in [1, p]$ is some integer for controlling the sparsity of the solution. In the literature, one common approach for finding an approximate solution to (44) is by solving the following l_1 regularization problem:

$$\min_{v,w} l_{\rm avg}(v, w) + \lambda \|w\|_1 \tag{45}$$

for some regularization parameter $\lambda \ge 0$ (see, for example, [28, 19, 42, 30, 33, 49]). Our aim below is to apply the PD method studied in Subsection 4.1 to solve (44) directly.

Letting $x = (v, w)$, $J = \{2, \ldots, p + 1\}$ and $f(x) = l_{\rm avg}(x_1, x_J)$, we can see that problem (44) is in the form of (1). Therefore, the PD method proposed in Subsection 4.1 can be suitably applied to solve (44). Also, we observe that the main computation effort of the PD method when applied to (44) lies in solving the subproblem arising in step 1a), which is in the form of

$$\min_x \left\{ l_{\rm avg}(x_1, x_J) + \frac{\varrho}{2}\|x - c\|^2 : \ x \in \Re^{p+1} \right\} \tag{46}$$

for some $\varrho > 0$ and $c \in \Re^{p+1}$. To efficiently solve (46), we apply the nonmonotone projected gradient method proposed in [4, Algorithm 2.2]; in particular, we set its parameter $M = 2$ and terminate the method when $\|\nabla F(x)\|/\max\{|F(x)|, 1\} \le 10^{-4}$, where $F(x)$ denotes the objective function of (46).

We now address the initialization and the termination criteria for our PD method when applied to (44). In particular, we randomly generate $z \in \Re^{p+1}$ such that $\|z_J\|_0 \le r$ and set the initial point $y_0^0 = z$. We choose the initial penalty parameter ϱ_0 to be 0.1, and set the parameter $\sigma = \sqrt{10}$. In addition, we

use (14) and (15) as the inner and outer termination criteria for the PD method and set their accuracy parameters ϵ_I and ϵ_O to be 5×10^{-4} and 10^{-3}, respectively.

We next conduct numerical experiments to test the performance of our PD method for solving (44) on some real and random data. We also compare the quality of the approximate solutions of (44) obtained by our method with that of (45) found by a first-order solver SLEP [33]. For the latter method, we set opts.mFlag=1, opts.lFlag=1 and opts.tFlag=2. And the rest of its parameters are set by default.

In the first experiment, we compare the solution quality of our PD method with SLEP on three small- or medium-sized benchmark data sets which are from the UCI machine learning bench market repository [40] and other sources [23]. The first data set is the colon tumor gene expression data [23] with more features than samples; the second one is the ionosphere data [40] with less features than samples; and the third one is the Internet advertisements data [40] with roughly same magnitude of features as samples. We discard the samples with missing data and standardize each data set so that the sample mean is zero and the sample variance is one. For each data set, we first apply SLEP to solve problem (45) with four different values of λ, which are the same ones as used in [28], namely, $0.5\lambda_{\max}$, $0.1\lambda_{\max}$, $0.05\lambda_{\max}$, and $0.01\lambda_{\max}$, where $\lambda_{\max}$ is the upper bound on the useful range of λ that is defined in [28]. For each such λ, let w^*_λ be the approximate optimal w obtained by SLEP. We then apply our PD method to solve problem (44) with $r = \|w^*_\lambda\|_0$ so that the resulting approximate optimal w is at least as sparse as w^*_λ.

To compare the solution quality of the above two methods, we introduce a criterion, that is, *error rate*. Given any model variables (v, w) and a sample vector $z \in \Re^p$, the outcome predicted by (v, w) for z is given by

$$\phi(z) = \mathrm{sgn}(w^T z + v),$$

where

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

Recall that z^i and b_i are the given samples and outcomes for $i = 1, \ldots, n$. The *error rate* of (v, w) for predicting the outcomes $b_1, \ldots, b_n$ is defined as

$$\text{Error} := \left\{ \sum_{i=1}^{n} \|\phi(z^i) - b_i\|_0 / n \right\} \times 100\%.$$

The computational results are presented in Table 1. In detail, the name and dimensions of each data set are given in the first three columns. The fourth column gives the ratio between λ and its upper bound $\lambda_{\max}$. The fifth column lists the value of r, that is, the cardinality of w^*_λ which is defined above. In addition, the average logistic loss, the error rate and the CPU time (in seconds) for both SLEP and PD are reported in columns six to eleven. We can observe that, although SLEP is faster than the PD method in most cases, the PD method substantially outperforms SLEP in terms of the solution quality since it generally achieves lower average logistic loss and error rate while the sparsity of both solutions is the same.

In the second experiment, we test our PD method on the random data sets of three different sizes. For each size, we randomly generate the data set consisting of 100 instances. In particular, the first data set has more features than samples; the second data set has more samples than features; and the last data set has equal number of features as samples. The samples $\{z^1, \ldots, z^n\}$ and the corresponding outcomes $b_1, \ldots, b_n$ are generated in the same manner as described in [28]. In detail, for each instance we choose equal number of positive and negative samples, that is, $m_+ = m_- = m/2$, where m_+ (resp., m_-) is the number of samples with outcome $+1$ (resp., -1). The features of positive (resp., negative)

Table 1: Computational results on three real data sets

Data	Features	Samples			SLEP			PD		
	p	n	$\lambda/\lambda_{\max}$	r	l_{avg}	Error (%)	Time	l_{avg}	Error (%)	Time
Colon	2000	62	0.5	7	0.4398	17.74	0.2	0.4126	12.9	9.1
			0.1	22	0.1326	1.61	0.5	0.0150	0	6.0
			0.05	25	0.0664	0	0.6	0.0108	0	5.0
			0.01	28	0.0134	0	1.3	0.0057	0	5.4
Ionosphere	34	351	0.5	3	0.4804	17.38	0.1	0.3466	13.39	0.7
			0.1	11	0.3062	11.40	0.1	0.2490	9.12	1.0
			0.05	14	0.2505	9.12	0.1	0.2002	8.26	1.1
			0.01	24	0.1846	6.55	0.4	0.1710	5.98	1.7
Advertisements	1430	2359	0.5	3	0.2915	12.04	2.3	0.2578	7.21	31.9
			0.1	36	0.1399	4.11	14.2	0.1110	4.11	56.0
			0.05	67	0.1042	2.92	21.6	0.0681	2.92	74.1
			0.01	197	0.0475	1.10	153.0	0.0249	1.10	77.4

Table 2: Computational results on random data sets

Size			SLEP			PD		
$n \times p$	$\lambda/\lambda_{\max}$	r	l_{avg}	Error (%)	Time	l_{avg}	Error (%)	Time
1000×2000	0.9	17.0	0.6411	9.76	0.4	0.2145	8.49	9.9
	0.7	52.9	0.5090	3.96	1.0	0.0588	2.66	20.0
	0.5	96.6	0.3838	2.23	1.7	0.0060	0.02	34.9
	0.3	138.7	0.2611	1.22	2.1	0.0022	0	25.5
	0.1	192.0	0.1228	0.31	2.0	0.0013	0	16.0
2000×1000	0.9	11.0	0.6441	11.46	0.4	0.2763	10.67	15.2
	0.7	42.8	0.5083	3.63	1.1	0.0376	1.49	38.9
	0.5	78.0	0.3776	1.65	2.0	0.0032	0	34.4
	0.3	115.5	0.2490	0.6	2.6	0.0015	0	25.3
	0.1	160.8	0.1056	0.03	3.1	0.0010	0	15.8
1000×1000	0.9	11.7	0.6417	11.00	0.1	0.2444	9.67	2.3
	0.7	37.2	0.5086	3.95	0.2	0.0572	2.46	5.8
	0.5	67.6	0.3805	2.15	0.3	0.0060	0.01	6.2
	0.3	100.1	0.2544	0.81	0.4	0.0016	0	4.6
	0.1	137.9	0.1124	0.12	0.5	0.0011	0	3.3

samples are independent and identically distributed, drawn from a normal distribution $N(\mu, 1)$, where μ is in turn drawn from a uniform distribution on $[0, 1]$ (resp., $[-1, 0]$). For each such instance, similar to the previous experiment, we first apply SLEP to solve problem (45) with five different values of λ, which are $0.9\lambda_{\max}$, $0.7\lambda_{\max}$, $0.5\lambda_{\max}$, $0.3\lambda_{\max}$ and $0.1\lambda_{\max}$. For each such λ, let w^*_λ be the approximate optimal w obtained by SLEP. We then apply our PD method to solve problem (44) with $r = \|w^*_\lambda\|_0$ so that the resulting approximate optimal w is at least as sparse as w^*_λ. The average results of each data set over 100 instances are reported in Table 2. We also observe that the PD method is slower than SLEP, but it has better solution quality than SLEP in terms of average logistic loss and error rate.

In summary, the above experiments demonstrate that the quality of the approximate solution of (44) obtained by our PD method is generally better than that of (45) found by SLEP when the same sparsity is considered. This observation is actually not surprising as (45) is a relaxation of (44).

5.2 Sparse inverse covariance selection problem

In this subsection, we apply the PD method proposed in Subsection 4.1 to solve the sparse inverse covariance selection problem, which has numerous real-world applications such as speech recognition

and gene network analysis (see, for example, [2, 18]).

Given a sample covariance matrix $\mathbf{\Sigma} \in \mathcal{S}^p_{++}$ and a set Ω consisting of pairs of known conditionally independent nodes, the sparse inverse covariance selection problem can be formulated as

$$\begin{array}{ll} \max\limits_{X \succeq 0} & \log\det X - \langle \mathbf{\Sigma}, X\rangle \\ \text{s.t.} & \sum\limits_{(i,j)\in\bar{\Omega}} \|X_{ij}\|_0 \le r, \\ & X_{ij} = 0 \quad \forall (i,j) \in \Omega, \end{array} \tag{47}$$

where $\bar{\Omega} = \{(i,j) : (i,j) \notin \Omega,\ i \neq j\}$, and $r \in [1, |\bar{\Omega}|]$ is some integer for controlling the sparsity of the solution. In the literature, one common approach for finding an approximate solution to (47) is by solving the following l_1 regularization problem:

$$\begin{array}{ll} \max\limits_{X \succeq 0} & \log\det X - \langle \mathbf{\Sigma}, X\rangle - \sum\limits_{(i,j)\in\bar{\Omega}} \rho_{ij}|X_{ij}| \\ \text{s.t.} & X_{ij} = 0 \quad \forall (i,j) \in \Omega, \end{array} \tag{48}$$

where $\{\rho_{ij}\}_{(i,j)\in\bar{\Omega}}$ is a set of regularization parameters (see, for example, [14, 15, 1, 35, 36, 21, 56, 34]). Our goal below is to apply the PD method studied in Subsection 4.1 to solve (47) directly.

Letting $\mathcal{X} = \left\{X \in \mathcal{S}^p_+ : X_{ij} = 0,\ (i,j) \in \Omega\right\}$ and $J = \bar{\Omega}$, we clearly see that problem (47) is in the form of (1) and thus it can be suitably solved by the PD method proposed in Subsection 4.1 with

$$\mathcal{Y} = \left\{ Y \in \mathcal{S}^p : \sum_{(i,j)\in\bar{\Omega}} \|Y_{ij}\|_0 \le r \right\}.$$

Notice that the main computation effort of the PD method when applied to (47) lies in solving the subproblem arising in step 1a), which is in the form of

$$\min_{X \succeq 0} \left\{ -\log\det X + \frac{\varrho}{2}\|X - C\|_F^2 : X_{ij} = 0\ \forall (i,j) \in \Omega \right\} \tag{49}$$

for some $\varrho > 0$ and $C \in \mathcal{S}^p$. Given that problem (49) generally does not have a closed-form solution, we now slightly modify the above sets $\mathcal{X}$ and $\mathcal{Y}$ by replacing them by

$$\mathcal{X} = \mathcal{S}^p_+, \qquad \mathcal{Y} = \left\{ Y \in \mathcal{S}^p : \sum_{(i,j)\in\bar{\Omega}} \|Y_{ij}\|_0 \le r,\ Y_{ij} = 0,\ (i,j) \in \Omega \right\},$$

respectively, and then apply the PD method presented in Subsection 4.1 to solve (47). For this PD method, the subproblem arising in step 1a) is now in the form of

$$\min_{X} \left\{ -\log\det X + \frac{\varrho}{2}\|X - C\|_F^2 : X \succeq 0 \right\} \tag{50}$$

for some $\varrho > 0$ and $C \in \mathcal{S}^p$. It can be shown that problem (50) has a closed-form solution, which is given by $V\mathscr{D}(x^*)V^T$, where $x_i^* = (\lambda_i + \sqrt{\lambda_i^2 + 4/\varrho})/2$ for all i and $V\mathscr{D}(\lambda)V^T$ is the eigenvalue decomposition of C for some $\lambda \in \Re^p$ (see, for example, Proposition 2.7 of [37]). Also, it follows from Proposition 3.1 that the subproblem arising in step 1b) for the above $\mathcal{Y}$ has a closed-form solution.

We now address the initialization and the termination criteria for the above PD method. In particular, we set the initial point $Y_0^0 = (\widetilde{\mathscr{D}}(\mathbf{\Sigma}))^{-1}$, the initial penalty parameter $\varrho_0 = 1$, and the parameter $\sigma = \sqrt{10}$. In addition, we use (15) and

$$\frac{|q_{\varrho_k}(x_{l+1}^k, y_{l+1}^k) - q_{\varrho_k}(x_l^k, y_l^k)|}{\max\{|q_{\varrho_k}(x_l^k, y_l^k)|, 1\}} \leq \epsilon_I$$

as the outer and inner termination criteria for the PD method, and set the associated accuracy parameters $\epsilon_O = 10^{-4}$ and $\epsilon_I = 10^{-4},\ 10^{-3}$ for the random and real data below, respectively.

We next conduct numerical experiments to test the performance of our PD method for solving (47) on some random and real data. We also compare the quality of the approximate solutions of (47) obtained by our method with that of (48) found by the proximal point algorithm (PPA) [56]. Both methods call the LAPACK routine **dsyevd.f** [29] for computing the full eigenvalue decomposition of a symmetric matrix, which is usually faster than the Matlab's **eig** routine when p is larger than 500. For PPA, we set $\mathrm{Tol} = 10^{-6}$ and use the default values for all other parameters.

In the first experiment, we compare the solution quality of our PD method with PPA on a set of random instances which are generated in a similar manner as described in [14, 35, 36, 56, 34]. In particular, we first generate a true covariance matrix $\mathbf{\Sigma^t} \in \mathcal{S}_{++}^p$ such that its inverse $(\mathbf{\Sigma^t})^{-1}$ is with the prescribed density δ, and set

$$\Omega = \left\{(i,j) : (\mathbf{\Sigma^t})_{ij}^{-1} = 0,\ |i-j| \geq \lfloor p/2 \rfloor\right\}.$$

We then generate a matrix $B \in \mathcal{S}^p$ by letting

$$B = \mathbf{\Sigma^t} + \tau V,$$

where $V \in \mathcal{S}^p$ contains pseudo-random values drawn from a uniform distribution on the interval $[-1, 1]$, and τ is a small positive number. Finally, we obtain the following sample covariance matrix:

$$\mathbf{\Sigma} = B - \min\{\lambda_{\min}(B) - \vartheta, 0\}I,$$

where ϑ is a small positive number. Specifically, we choose $\tau = 0.15$, $\vartheta = 1.0e-4$, $\delta = 10\%$, 50% and 100%, respectively. It is clear that for $\delta = 100\%$ and the set Ω is an empty set. In addition, for all $(i,j) \in \bar{\Omega}$, we set $\rho_{ij} = \rho_{\bar{\Omega}}$ for some $\rho_{\bar{\Omega}} > 0$. For each instance, we first apply PPA to solve (48) for four values of $\rho_{\bar{\Omega}}$, which are 0.01, 0.1, 1, and 10. For each $\rho_{\bar{\Omega}}$, let $\tilde{X}^*$ be the solution obtained by PPA. We then apply our PD method to solve problem (47) with $r = \sum_{(i,j)\in\bar{\Omega}} \|\tilde{X}_{ij}^*\|_0$ so that the resulting solution is at least as sparse as $\tilde{X}^*$.

As mentioned in [34], to evaluate how well the true inverse covariance matrix $(\mathbf{\Sigma^t})^{-1}$ is recovered by a matrix $X \in \mathcal{S}_{++}^p$, one can compute the *normalized entropy loss* which is defined as follows:

$$\mathrm{Loss} := \frac{1}{p}(\langle \mathbf{\Sigma^t}, X\rangle - \log\det(\mathbf{\Sigma^t} X) - p).$$

The results of PPA and the PD method on these instances are presented in Tables 3-5, respectively. In each table, the order p of $\mathbf{\Sigma}$ is given in column one. The size of Ω is given in column two. The values of $\rho_{\bar{\Omega}}$ and r are given in columns three and four. The log-likelihood (i.e., the objective value of (47)), the normalized entropy loss and the CPU time (in seconds) of PPA and the PD method are given in the last six columns, respectively. We observe that our PD method is substantially faster than PPA

Table 3: Computational results for $\delta = 10\%$

Problem				PPA			PD		
p	$\lvert\Omega\rvert$	$\rho_{\bar{\Omega}}$	r	Likelihood	Loss	Time	Likelihood	Loss	Time
500	56724	0.01	183876	−950.88	2.4594	34.1	−936.45	2.3920	2.5
		0.10	45018	−999.89	2.5749	44.8	−978.61	2.4498	5.3
		1.00	5540	−1046.44	2.9190	66.2	−1032.79	2.6380	24.8
		10.0	2608	−1471.67	4.2442	75.1	−1129.50	2.8845	55.5
1000	226702	0.01	745470	−2247.14	3.1240	150.2	−2220.47	3.0486	13.1
		0.10	186602	−2344.03	3.2291	158.7	−2301.12	3.1224	19.8
		1.00	29110	−2405.88	3.5034	349.8	−2371.68	3.2743	59.1
		10.0	9604	−3094.57	4.6834	395.9	−2515.80	3.4243	129.5
1500	509978	0.01	1686128	−3647.71	3.4894	373.7	−3607.23	3.4083	35.7
		0.10	438146	−3799.02	3.5933	303.6	−3731.17	3.5059	44.9
		1.00	61222	−3873.93	3.8319	907.4	−3832.88	3.6226	155.3
		10.0	17360	−4780.33	4.9264	698.8	−3924.94	3.7146	328.0
2000	905240	0.01	3012206	−5177.80	3.7803	780.0	−5126.09	3.7046	65.5
		0.10	822714	−5375.21	3.8797	657.5	−5282.37	3.7901	94.3
		1.00	126604	−5457.90	4.0919	907.4	−5424.66	3.9713	200.2
		10.0	29954	−6535.54	5.1130	1397.4	−5532.03	4.0019	588.0

Table 4: Computational results for $\delta = 50\%$

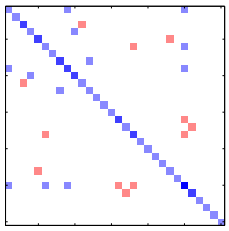
Problem				PPA			PD		
p	$\lvert\Omega\rvert$	$\rho_{\bar{\Omega}}$	r	Likelihood	Loss	Time	Likelihood	Loss	Time
500	37738	0.01	202226	−947.33	3.1774	37.2	−935.11	3.1134	2.2
		0.10	50118	−1001.23	3.3040	41.8	−978.03	3.1662	4.7
		1.00	11810	−1052.09	3.6779	81.1	−101.80	3.2889	14.5
		10.0	5032	−1500.00	5.0486	71.1	−1041.64	3.3966	28.1
1000	152512	0.01	816070	−2225.875	3.8864	149.7	−2201.98	3.8126	12.1
		0.10	203686	−2335.81	4.0029	131.0	−2288.11	3.8913	17.2
		1.00	46928	−2400.81	4.2945	372.7	−2349.02	4.0085	44.1
		10.0	17370	−3128.63	5.5159	265.2	−2390.09	4.1138	84.3
1500	340656	0.01	1851266	−3649.78	4.2553	361.2	−3616.72	4.1787	32.0
		0.10	475146	−3815.09	4.3668	303.4	−3743.19	4.2725	42.3
		1.00	42902	−3895.09	4.6025	1341.0	−3874.68	4.4823	155.8
		10.0	7430	−4759.67	5.6739	881.2	−4253.34	4.6876	468.6
2000	605990	0.01	3301648	−5149.12	4.5763	801.3	−5104.27	4.5006	61.7
		0.10	893410	−5371.26	4.6851	620.0	−5269.06	4.5969	82.4
		1.00	153984	−5456.54	4.9033	1426.0	−5406.89	4.7614	175.9
		10.0	33456	−6560.54	5.9405	1552.3	−5512.48	4.7982	565.5

for these instances. Moreover, it outperforms PPA in terms of solution quality since it achieves larger log-likelihood and smaller normalized entropy loss.

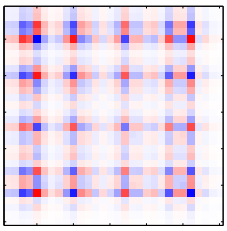
Our second experiment is similar to the one conducted in [14, 36]. We intend to compare sparse recoverability of our PD method with PPA. To this aim, we specialize $p = 30$ and $(\mathbf{\Sigma^t})^{-1} \in S^p_{++}$ to be the matrix with diagonal entries around one and a few randomly chosen, nonzero off-diagonal entries equal to $+1$ or -1. And the sample covariance matrix $\mathbf{\Sigma}$ is then similarly generated as above. In addition, we set $\Omega = \{(i,j) : (\mathbf{\Sigma^t})^{-1}_{ij} = 0,\ |i-j| \geq 15\}$ and $\rho_{ij} = \rho_{\bar{\Omega}}$ for all $(i,j) \in \bar{\Omega}$, where $\rho_{\bar{\Omega}}$ is the smallest number such that the approximate solution obtained by PPA shares the same number of nonzero off-diagonal entries as $(\mathbf{\Sigma^t})^{-1}$. For problem (47), we choose $r = \sum_{(i,j)\in\bar{\Omega}} \|(\mathbf{\Sigma^t})^{-1}_{ij}\|_0$ (i.e., the number of nonzero off-diagonal entries of $(\mathbf{\Sigma^t})^{-1}$). PPA and the PD method are then applied to solve (48) and (47) with the aforementioned ρ_{ij} and r, respectively. In Figure 1, we plot the sparsity patterns

Table 5: Computational results for $\delta = 100\%$

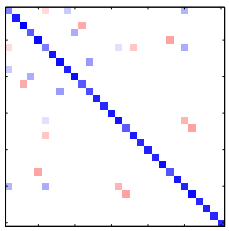
Problem				PPA			PD		
p	$\lvert\Omega\rvert$	$\rho_{\bar{\Omega}}$	r	Likelihood	Loss	Time	Likelihood	Loss	Time
500	0	0.01	238232	−930.00	3.5345	36.0	−918.52	3.4838	1.3
		0.10	57064	−1000.78	3.6826	43.6	−973.06	3.5313	4.0
		1.00	15474	−1053.04	4.0675	76.1	−1006.95	3.6500	10.6
		10.0	7448	−1511.88	5.4613	51.4	−1023.82	3.7319	18.1
1000	0	0.01	963400	−2188.06	4.1983	156.3	−2161.58	4.1383	5.3
		0.10	231424	−2335.09	4.3387	122.4	−2277.90	4.2045	16.8
		1.00	47528	−2401.69	4.6304	329.6	−2349.74	4.3449	42.6
		10.0	18156	−3127.94	5.8521	244.1	−2388.22	4.4466	79.0
1500	0	0.01	2181060	−3585.21	4.5878	364.1	−3545.43	4.5260	12.3
		0.10	551150	−3806.07	4.7234	288.2	−3717.25	4.6059	41.3
		1.00	102512	−3883.94	4.9709	912.8	−3826.26	4.7537	93.5
		10.0	31526	−4821.26	6.0886	848.7	−3898.50	4.8824	185.4
2000	0	0.01	3892592	−5075.44	4.8867	734.1	−5021.95	4.8222	23.8
		0.10	1027584	−5367.86	5.0183	590.6	−5246.45	4.9138	76.1
		1.00	122394	−5456.64	5.2330	1705.8	−5422.48	5.1168	197.8
		10.0	25298	−6531.08	6.2571	1803.4	−5636.74	5.3492	417.1



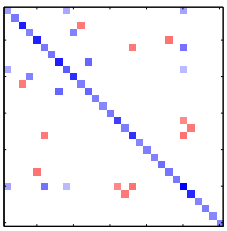
(a) True inverse $(\mathbf{\Sigma^t})^{-1}$



(b) Noisy inverse $\mathbf{\Sigma}^{-1}$



(c) Approximate solution of (48)



(d) Approximate solution of (47)

Figure 1: Sparse recovery.

of the original inverse covariance matrix $(\mathbf{\Sigma^t})^{-1}$, the noisy inverse sample covariance matrix $\mathbf{\Sigma}^{-1}$, and the approximate solutions to (48) and (47) obtained by PPA and our PD method, respectively. We first observe that the sparsity of both solutions is the same as $(\mathbf{\Sigma^t})^{-1}$. Moreover, the solution of our PD method completely recovers the sparsity patterns of $(\mathbf{\Sigma^t})^{-1}$, but the solution of PPA misrecovers a few patterns. In addition, we present the log-likelihood and the normalized entropy loss of these solutions in Table 6. One can see that the solution of our PD method achieves much larger log-likelihood and smaller normalized entropy loss.

Table 6: Numerical results for sparse recovery

	nnz	Likelihood	Loss
PPA	24	−35.45	0.178
PD	24	−29.56	0.008

Table 7: Computational results on two real data sets

Data	Genes	Samples			PPA			PD		
	p	n	$\rho_{\bar{\Omega}}$	r	Likelihood	Loss	Time	Likelihood	Loss	Time
Lymph	587	148	0.01	144294	790.12	23.24	101.5	1035.24	22.79	38.0
			0.05	67474	174.86	24.35	85.2	716.97	23.27	31.5
			0.10	38504	−47.03	24.73	66.7	389.65	23.85	26.1
			0.50	4440	−561.38	25.52	33.2	−260.32	24.91	24.8
			0.70	940	−642.05	25.63	26.9	−511.70	25.30	22.0
			0.90	146	−684.59	25.70	22.0	−598.05	25.51	14.9
Leukemia	1255	72	0.01	249216	3229.75	28.25	705.7	3555.38	28.12	177.1
			0.05	169144	1308.38	29.85	491.1	2996.95	28.45	189.2
			0.10	107180	505.02	30.53	501.4	2531.62	28.82	202.8
			0.50	37914	−931.59	31.65	345.9	797.23	30.16	256.6
			0.70	4764	−1367.22	31.84	125.7	−1012.48	31.48	271.6
			0.90	24	−1465.70	31.90	110.6	−1301.99	31.68	187.8

In the third experiment, we aim to compare the performance of our PD method with the PPA on two gene expression data sets that have been widely used in the literature (see, for example, [24, 45, 57, 17, 34]). We first pre-process the data by the same procedure as described in [34] to obtain a sample covariance matrix $\mathbf{\Sigma}$, and set $\Omega = \emptyset$ and $\rho_{ij} = \rho_{\bar{\Omega}}$ for some $\rho_{\bar{\Omega}} > 0$. We apply PPA to solve problem (48) with $\rho_{\bar{\Omega}} = 0.01$, 0.05, 0.1, 0.5, 0.7 and 0.9, respectively. For each $\rho_{\bar{\Omega}}$, we choose r to be the number of nonzero off-diagonal entries of the solution of PPA, which implies that the solution of the PD method when applied to (47) is at least as sparse as that of PPA. As the true covariance matrix $\mathbf{\Sigma^t}$ is unknown for these data sets, we now modify the normalized entropy loss defined above by replacing $\mathbf{\Sigma^t}$ by $\mathbf{\Sigma}$. The results of PPA and our PD method on these two data sets are presented in Table 7. In detail, the name and dimension of each data set are given in the first three columns. The values of $\rho_{\bar{\Omega}}$ and r are listed in the fourth and fifth columns. The log-likelihood, the normalized entropy loss and the CPU time (in seconds) of PPA and the PD method are given in the last six columns, respectively. We can observe that our PD method is generally faster than PPA. Moreover, our PD method outperforms PPA in terms of log-likelihood and normalized entropy loss.

As a summary, the above experiments show that the quality of the approximate solution of (47) obtained by our PD method is generally better than that of (48) found by PPA when the same sparsity is considered.

5.3 Compressed sensing

In this subsection, we apply the PD methods proposed in Section 4 to solve the compressed sensing (CS) problem, which has important applications in signal processing (see, for example, [13, 50, 31, 48, 10, 39, 53]).

When the observation is noise free, the CS problem can be formulated as

$$\min_{x \in \Re^p} \{\|x\|_0 : \ Ax = b\}, \tag{51}$$

where $A \in \Re^{n \times p}$ is a data matrix and $b \in \Re^n$ is an observation vector. One popular approach for finding

an approximate solution to (51) is to solve the following l_1 regularization problem:

$$\min_{x\in\Re^p}\{\|x\|_1 :\ Ax = b\}, \tag{52}$$

(see, for example, [55, 10]). Our aim below is to apply the PD method studied in Subsection 4.2 to solve problem (51) directly.

Clearly, problem (51) is in the form of (2) and thus the PD method proposed in Subsection 4.2 can be suitably applied to solve (51). Also, one can observe that the main computation effort of the PD method when applied to (51) lies in solving the subproblem arising in step 1a), which is in the form of

$$\min_x\{\|x - c\|^2 : Ax = b\} \tag{53}$$

for some $c \in \Re^p$. It is well known that problem (53) has a closed-form solution given by

$$x^* = c - A^T(AA^T)^{-1}(Ac - b).$$

We now address the initialization and the termination criteria for the PD method. In particular, we choose y_0^0 to be a feasible point of (51) with at most n nonzero entries which can be obtained by executing the Matlab command $A \setminus b$. Also, we set the initial penalty parameter $\varrho_0 = 0.1$ and the parameter $\sigma = 10$. In addition, we use (14) and

$$\frac{\|x^k - y^k\|_\infty}{\max\{|p_{\varrho_k}(x^k, y^k)|, 1\}} \le \epsilon_O$$

as the inner and outer termination criteria, and set the associated accuracy parameters $\epsilon_I = 10^{-5}$ and $\epsilon_O = 10^{-6}$, respectively.

We next conduct experiments to test the performance of our PD method for solving problem (51) on random data. We also compare the quality of the approximate solutions of (51) obtained by our PD method with that of (52) found by a first-order solver SPGL1 [55]. For the latter method, we use the default values for all parameters.

In the first experiment, given an integer $r \in [1, p]$, we randomly generate 100 instances according to the standard Gaussian distribution. Each one consists of a sparse signal u with cardinality r and a data matrix $A \in \Re^{n\times p}$. Then we generate the corresponding observation vector b by letting $b = Au$. In particular, we choose $n = 1024$, $p = 4096$. The values of r range from 30 to 300 (see Table 8). We now try to recover u by applying the PD method and SPGL1 to solve (51) and (52), respectively. To evaluate the solution quality of these methods, we adopt a similar criterion as described in [46, 7]. Given an approximate recovery x^* for u, we define the mean squared error as

$$\text{MSE} := \|x^* - u\|/p.$$

We say u is *successfully recovered* by x^* if the cardinality of x^* is the same as u and moreover the corresponding MSE is less than 10^{-4}. The computational results of both methods are presented in Table 8. In detail, the values of r are given in the first column. The number of successfully recovered signals (NS) and the CPU time for both methods are reported in columns two to five, respectively. We observe that the recoverability of two methods is similar for the instances with relatively small r, but the PD method outperforms SPGL1 when r becomes larger. We also see that the speed of both methods is comparable.

Table 8: Computational results for A with non-orthonormal rows

	SPGL1		PD	
Cardinality	NS	Time	NS	Time
30	100	0.6	100	2.4
60	100	1.0	100	2.8
90	100	1.4	100	3.2
120	100	2.0	100	3.5
150	94	3.0	98	3.9
180	93	4.8	97	4.7
210	81	10.4	86	6.1
240	22	23.6	68	12.4
270	0	25.8	18	20.0
300	0	28.2	0	22.6

Table 9: Computational results for A with orthonormal rows

	SPGL1		PD	
Cardinality	NS	Time	NS	Time
30	100	0.4	100	1.8
60	100	0.6	100	2.1
90	100	0.7	100	2.4
120	100	1.0	100	2.8
150	92	1.4	95	3.2
180	91	2.1	95	4.0
210	73	3.9	92	5.5
240	29	9.0	61	12.1
270	1	10.9	11	18.9
300	0	11.3	1	19.1

The second experiment is similar to the first one except that A is randomly generated with orthonormal rows. The computational results of both methods are presented in Table 9. We also observe that the PD method outperforms SPGL1 in terms of recoverability.

In the remainder of this subsection we consider the CS problem with noisy observation. In this case, the CS problem can be formulated as

$$\min_{x\in\Re^p}\left\{\frac{1}{2}\|Ax-b\|^2:\ \|x\|_0\le r\right\}, \tag{54}$$

where $A\in\Re^{n\times p}$ is a data matrix, $b\in\Re^n$ is an observation vector, and $r\in[1,p]$ is some integer for controlling the sparsity of the solution. One popular approach for finding an approximate solution to (54) is to solve the following l_1 regularization problem:

$$\min_{x\in\Re^p}\frac{1}{2}\|Ax-b\|^2+\lambda\|x\|_1, \tag{55}$$

where $\lambda\ge 0$ is a regularization parameter (see, for example, [20, 25, 27]). Our goal below is to apply the PD method studied in Subsection 4.1 to solve (54) directly.

Clearly, problem (54) is in the form of (1) and thus the PD method proposed in Subsection 4 can be suitably applied to solve (54). The main computation effort of the PD method when applied to (54) lies in solving the subproblem arising in step 1a), which is an unconstrained quadratic programming problem that can be solved by the conjugate gradient method. We now address the initialization and the termination criteria for the PD method. In particular, we randomly choose an initial point $y_0^0\in\Re^p$

such that $\|y_0^0\|_0 \le r$. Also, we set the initial penalty parameter $\varrho_0 = 1$ and the parameter $\sigma = \sqrt{10}$. In addition, we use

$$\frac{|q_{\varrho_k}(x_{l+1}^k, y_{l+1}^k) - q_{\varrho_k}(x_l^k, y_l^k)|}{\max\{|q_{\varrho_k}(x_l^k, y_l^k)|, 1\}} \le \epsilon_I$$

and

$$\frac{\|x^k - y^k\|_\infty}{\max\{|q_{\varrho_k}(x^k, y^k)|, 1\}} \le \epsilon_O$$

as the inner and outer termination criteria for the PD method, and set their associated accuracy parameters $\epsilon_I = 10^{-2}$ and $\epsilon_O = 10^{-3}$.

We next conduct numerical experiments to test the performance of our PD method for solving problem (54) on random data. We also compare the quality of the approximate solutions of (54) obtained by our PD method and the iterative hard-thresholding algorithm (IHT) [5, 6] with that of (55) found by a first-order solver GPSR [20]. For IHT, we set $stopTol = 10^{-6}$ and use the default values for all other parameters. And for GPSR, all the parameters are set as their default values.

We first randomly generate a data matrix $A \in \Re^{n\times p}$ and an observation vector $b \in \Re^n$ according to a standard Gaussian distribution. Then we apply GPSR to problem (55) with a set of p distinct λ's so that the cardinality of the resulting approximate solution gradually increases from 1 to p. Accordingly, we apply our PD method and IHT to problem (54) with $r = 1, \ldots, p$. It shall be mentioned that a warm-start strategy is applied to all three methods. That is, an approximate solution of problem (54) (resp., (52)) for current r (resp., λ) is used as the initial point for the PD method and IHT (resp., GPSR) when applied to the problem for next r (resp., λ). The average computational results of both methods over 100 random instances with $(n, p) = (1024, 4096)$ are plotted in Figure 2. In detail, we plot the average residual $\|Ax - b\|$ against the cardinality in the left graph and the average accumulated CPU time [1] (in seconds) against the cardinality in the right graph. We observe that the residuals of the approximate solutions of (55) obtained by our PD method and IHT are almost equal and substantially smaller than that of (54) found by GPSR when the same sparsity is considered. In addition, we can see that GPSR is faster than the other two methods.

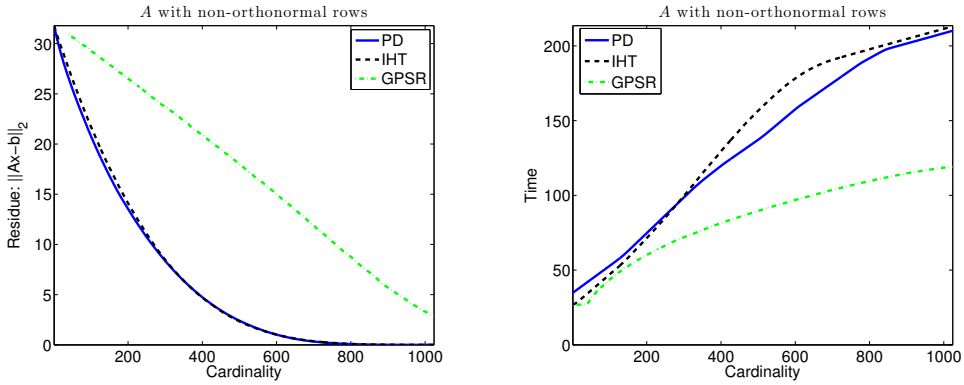
We also conduct a similar experiment as above except that A is randomly generated with orthonormal rows. The results are plotted in Figure 3. We observe that the PD method and IHT are generally slower than GPSR, but they have better solution quality than GPSR in terms of residuals.

6 Concluding remarks

In this paper we propose penalty decomposition methods for general l_0 minimization problems in which each subproblem is solved by a block coordinate descend method. Under some suitable assumptions, we establish that any accumulation point of the sequence generated by the PD methods satisfies the first-order optimality conditions of the problems. Furthermore, for the problems in which the l_0 part is the only nonconvex part, we show that such an accumulation point is a local minimizer of the problems. The computational results on compressed sensing, sparse logistic regression and sparse inverse covariance selection problems demonstrate that our methods generally outperform the existing methods in terms of solution quality and/or speed.

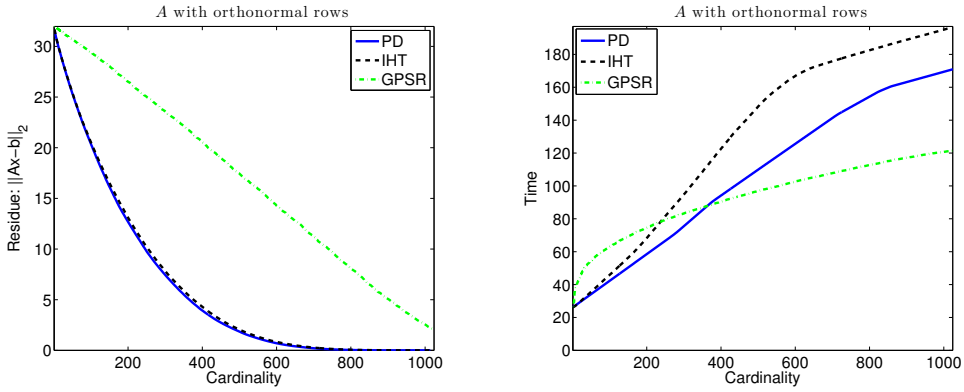
We shall remark that the augmented Lagrangian decomposition methods can be developed for solving l_0 minimization problems (1) and (2) simply by replacing the quadratic penalty functions in the

[1] For a cardinality r, the corresponding accumulated CPU time is the total CPU time used to compute approximate solutions of problem (54) or (52) with cardinality from 1 to r.



(a) Residual vs. Cardinality (b) Time vs. Cardinality

Figure 2: Trade-off curves.



(a) Residual vs. Cardinality (b) Time vs. Cardinality

Figure 3: Trade-off curves.

PD methods by augmented Lagrangian functions. Nevertheless, as observed in our experiments, their practical performance is generally worse than the PD methods.

Appendix

In this appendix we provide an example to demonstrate that the l_p-norm relaxation approaches for $p \in (0, 1]$ may fail to recover the sparse solution.

Let $p \in (0, 1]$ be arbitrarily chosen. Given any b^1, $b^2 \in \Re^n$, let $b = b^1 + b^2$, $\alpha = \|(b^1; b^2)\|_p$ and $A = [b^1,\ b^2,\ \alpha I_n,\ \alpha I_n]$, where I_n denotes the $n \times n$ identity matrix and $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ for all $x \in \Re^n$. Consider the linear system $Ax = b$. It is easy to observe that this system has the sparse solution $x^s = (1, 1, 0, \ldots, 0)^T$. However, x^s cannot be recovered by solving the l_p-"norm" regularization problem:

$$f^* = \min_x \left\{ f(x) := \frac{1}{2}\|Ax - b\|^2 + \nu\|x\|_p \right\}$$

for any $\nu > 0$. Indeed, let $\bar{x} = (0, 0, b^1/\alpha, b^2/\alpha)^T$. Then, we have $f(x^s) = 2^{1/p}\nu$ and $f(\bar{x}) = \nu$, which

implies that $f(x^s) > f(\bar{x}) \geq f^*$. Thus, x^s cannot be an optimal solution of the above problem for any $\nu > 0$. Moreover, the relative error between $f(x^s)$ and f^* is fairly large since

$$(f(x^s) - f^*)/f^* \ \geq \ (f(x^s) - f(\bar{x}))/f(\bar{x}) \ = \ 2^{1/p} - 1 \ \geq \ 1.$$

Therefore, the true sparse solution x^s may not even be a "good" approximate solution to the l_p-"norm" regularization problem.

References

[1] O. Banerjee, L. E. Ghaoui, and A. D'Aspremont. Model selection through sparse maximum likelihood estimation. *J. Mach. Learn. Res.*, 9:485-516, 2008.

[2] J. A. Bilmes. Factored sparse inverse covariance matrices. *International Conference on Acoustics, Speech and Signal processing*, Washington, D.C., 1009-1012, 2000.

[3] C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2007.

[4] E. G. Birgin, J. M. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM J. Optimiz*, 4:1196–1211, 2000.

[5] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *J. FOURIER ANAL. APPL.*, 14:629–654, 2008.

[6] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, 27(3):265–274, 2009.

[7] E. J. Candés and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 2009.

[8] E. J. Candés, J. Romberg and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE T. Inform. Theory*, 52:489-509, 2006.

[9] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Proc. Let.*, 14:707-710, 2007.

[10] S. Chen, D. Donoho and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20:33-61, 1998.

[11] X. Chen, F. Xu and Y. Ye. Lower bound theory of nonzero entries in solutions of l_2-l_p Minimization. *SIAM J. Sci. Comput.*, 32:2832-2852, 2010.

[12] X. Chen and W. Zhou. Convergence of reweighted l_1 minimization algorithms and unique solution of truncated l_p minimization. Technical report, 2010.

[13] J. Claerbout and F. Muir. Robust modelling of erratic data. *Geophysics*, 38:826-844, 1973.

[14] A. D'Aspremont, O. Banerjee and L. E. Ghaoui. First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. A.*, 30(1):56-66, 2008.

[15] J. Dahl, L. Vandenberghe and V. Roychowdhury. Covariance selection for nonchordal graphs via chordal embedding. *Optim. Method. Softw.*, 23(4):501-520, 2008.

[16] A. Dempster. Covariance selection. *Biometrics*, 28:157-175, 1978.

[17] A. Dobra. Dependency networks for genome-wide data. Biostatistics, 8(1):1-28, 2007.

[18] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao and M. West. Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.*, 90:196-212, 2004.

[19] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. Least angle regression. *Ann. Stat.*, 32(2):407-499, 2004.

[20] 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. Signa.: Special Issue on Convex Optimization Methods for Signal Processing*, 1(4):586-598, 2007.

[21] J. Friedman, T. Hastie and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostat.*, 9(3):432-441, 2008.

[22] A. D. Gerson, L. C. Parra and P. Sajda. Cortical origins of response time variability during rapid discrimination of visual objects. *Neuroimage*, 28(2):342-353, 2005.

[23] G. Golub and C. Van Loan. Matrix Computations, volume 13 of Studies in Applied Mathematics. John Hopkins University Press, third edition, 1996.

[24] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by expression monitoring. *Science*, 286:531-537, 1999.

[25] E. T. Hale, W. Yin and Y. Zhang. Fixed-point continuation applied to compressed sensing: Implementation and numerical experiments. *J. Comput. Math*, 28(2):170-194, 2010.

[26] K. K. Herrity, A. C. Gilbert and J. A. Tropp. Sparse approximation via iterative thresholding. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[27] 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. Signa.*, 1(4):606-617, December 2007.

[28] K. Koh, S. J. Kim and S. Boyd. An interior-point method for large-scale l_1-regularized logistic regression. *J. Mach. Learn. Res.*, 8:1519-1555, 2007.

[29] Linear Algebra PACKage. Available at http://www.netlib.org/lapack/index.html.

[30] S. Lee, H. Lee, P. Abbeel and A. Ng. Efficient l_1-regularized logistic regression. In *21th National Conference on Artificial Intelligence (AAAI)*, 2006.

[31] S. Levy and P. Fullagar. Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution. *Geophysics*, 46:1235-1243, 1981.

[32] J. G. Liao and K. V. Chin. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics*, 23(15):1945-1951, 2007.

[33] J. Liu, S. Ji and J. Ye. *SLEP: Sparse learning with efficient projections.* Arizona State University, 2009. Available at http://www.public.asu.edu/∼jye02/Software/SLEP.

[34] L. Li and K. C. Toh. An inexact interior point method for l_1-regularized sparse covariance selection. *Math. Program. Comput.*, 2:291-315, 2010.

[35] Z. Lu. Smooth optimization approach for sparse covariance selection. *SIAM J. Optimiz.*, 19(4):1807-1827, 2009.

[36] Z. Lu. Adaptive first-order methods for general sparse inverse covariance selection. *SIAM J. Matrix Anal. A.*, 31(4):2000-2016, 2010.

[37] Z. Lu and Y. Zhang. Penalty decomposition methods for rank minimization. Technical report, Department of Mathematics, Simon Fraser University, Canada, 2010.

[38] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE T. Image Process.*, 41(12):3397–3415, 1993.

[39] A. Miller. Subset selection in regression. Chapman and Hall, London, 2002.

[40] D. Newman, S. Hettich, C. Blake and C. Merz. UCI repository of machine learning databases, 1998. Available at www.ics.uci.edu/∼mlearn/MLRepository.html.

[41] A. Y. Ng. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In *Proceedings of the Twenty-First International Conference on Machine learning (ICML)*, 72-85, 2004.

[42] M. Y. Park and T. Hastie. *Regularization path algorithms for detecting gene interactions.* Department of Statistics, Stanford University, 2006.

[43] L. C. Parra, C. D. Spence, A. D. Gerson and P. Sajda. Recipes for the linear analysis of EEG. *Neuroimage*, 28(2):326-341, 2005.

[44] M. G. Philiastides and P. Sajda. Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cereb. Cortex*, 16(4):509-518, 2006.

[45] J. Pittman, E. Huang, H. Dressman, C. F. Horng, S. H. Cheng, M. H. Tsou, C. M. Chen, A. Bild, E. S. Iversen, A. T. Huang, J. R. Nevins and M. West. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *P. Natl. Acad. Sci. USA*, 101(22):8431-8436, 2004.

[46] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 2007.

[47] A. Ruszczyński. *Nonlinear Optimization.* Princeton University Press, 2006.

[48] F. Santosa and W. Symes. Linear inversion of band-limited reflection histograms. *SIAM J. Sci. Stat. Comp.*, 7:1307-1330, 1986.

[49] J. Shi, W. Yin, S. Osher and P. Sajda. A fast hybrid algorithm for large-scale l_1-regularized logistic regression. *J. Mach. Learn. Res.*, 11:713-741, 2010.

[50] H. Taylor, S. Bank and J. McCoy. Deconvolution with the l_1-norm. *Geophysics*, 44:39-52, 1979.

[51] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B*, 58(1):267-288, 1996.

[52] J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE T. Inform. Theory*, 50(10):2231–2242, 2004.

[53] J. Tropp. Just relax: Convex programming methods for identifying sparse signals. *IEEE T. Inform. Theory*, 51:1030-1051, 2006.

[54] Y. Tsuruoka, J. McNaught, J. Tsujii and S. Ananiadou. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics*, 23(20):2768-2774, 2007.

[55] E. Van Den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comp.*, 31(2)890-912, 2008.

[56] C. Wang, D. Sun and K. C. Toh. Solving log-determinant optimization problems by a Newton-CG proximal point algorithm. *SIAM J. Optimiz.*, 20(6):2994-3013, 2010.

[57] K. Y. Yeung, R. E. Bumgarner and A. E. Raftery. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, 21(10):2394-2402, 2005.

[58] Y. Zhang, B. Dong and Z. Lu. l_0 minimization for wavelet frame based image restoration. To appear in *Math. Comput.*, 2011.