Optimization Methods for Sparse Approximation

by

Yong Zhang

M.Sc., University of Regina, 2007

B.Sc. Hunan University, 2003

A Thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

in the
Department of Mathematics
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2014

APPROVAL

Name:	Yong Zhang
Degree:	Doctor of Philosophy
Title of Thesis:	Optimization Methods for Sparse Approximation
Examining Committee:	Dr. Alistair Lachlan, Professor Emeritus, Department of Mathematics, Chair
	Dr. Zhaosong Lu, Associate Professor, Department of Mathematics, Senior Supervisor
	Dr. Tamon Stephen, Associate Professor, Department of Mathematics, Supervisor
	Dr. Jiguo Cao, Associate Professor, Department of Statistics & Actuarial Science, Internal Examiner
	Dr. Shawn Xianfu Wang, Professor, Mathematics, University of British Columbia External Examiner
Date Approved:	August 26th, 2014

Partial Copyright Licence

Abstract

In the last two decades, there are numerous applications in which sparse solutions are concerned. Mathematically, all these applications can be formulated into the l_0 minimization problems. In this thesis, we first propose a novel *augmented Lagrangian* (AL) method for solving the l_1-norm relaxation problems of the original l_0 minimization problems and apply it to our proposed formulation of sparse *principal component analysis* (PCA). We next propose *penalty decomposition* (PD) methods for solving the original l_0 minimization problems in which a sequence of penalty subproblems are solved by a *block coordinate descent* (BCD) method.

For the AL method, we show that under some regularity assumptions, it converges to a stationary point. Additionally, we propose two nonmonotone gradient methods for solving the AL subproblems, and establish their global and local convergence. Moreover, we apply the AL method to our proposed formulation of sparse PCA and compare our approach with several existing methods on synthetic, Pitprops, and gene expression data, respectively. The computational results demonstrate that the sparse *principal components* (PCs) produced by our approach substantially outperform those by other methods in terms of total explained variance, correlation of PCs, and orthogonality of loading vectors.

For the PD methods, under some suitable assumptions, we establish some convergence results for both inner (the BCD method) and outer (the PD method) iterations, respectively. 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 when solutions of same cardinality are sought, our approach applied to the l_0-based models generally has better solution quality and/or speed than the existing approaches that are applied to the corresponding l_1-based models.

Finally, we adapt the PD method to solve our proposed wavelet frame based image

restoration problem. Some convergence analysis of the adapted PD method for this problem are provided. Numerical results show that the proposed model solved by the PD method can generate images with better quality than those obtained by either analysis based approach or balanced approach in terms of restoring sharp features as well as maintaining smoothness of the recovered images.

Keywords: Augmented Lagrangian method, Penalty decomposition method, Sparse PCA, Compressed sensing, Sparse logistic regression, Sparse inverse covariance selection

To my parents, my wife, and my children!

Acknowledgments

First, I would like to express my gratitude to my advisor, Zhaosong Lu. He was always supportive of my research work and contributed many insightful ideas for my thesis work. It was an honor being able to work with him.

Second, I would like to thank Tamon Stephen for being in my supervisory committee and spending the time to coordinate the operations research graduate program.

Third, I would like to thank Bin Dong and Ting Kei Pong for many fruitful discussions we had and their insightful comments.

Next, I would like to thank my examining committee for their time reading my thesis. I would also like to thank my friends at Simon Fraser University, especially, Xiaorui Li, Pooja Pandey, Xueyin Shen, Brad Woods, T.J. Yusun and Annie Zhang, for being supportive during my PhD study.

Special thanks are due to my family, especially my wife, for their encouragement and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the last two decades, numerous applications such as compressed sensing, sparse inverse covariance selection and sparse logistic regression have been identified to be optimization problems where sparse solutions are desired. There are many advantages working with sparse vectors. For example, calculations involving multiplying a vector by a matrix take less time to compute if the vector is sparse. Also sparse vectors require less space when being stored, as only the position and value of the entries need to be recorded. Nevertheless, the task of finding sparse approximations can be very difficult. For instance, in [92] it was shown that finding the sparsest solution to an underdetermined linear system is known to be NP hard in general.

This thesis proposes methods finding sparse approximations to optimization problems. The proposed methods are applied on solving some practical applications of sparse approximations.

1.1 Background

In this section, we introduce basic concepts and review some important applications of sparse approximation. Before proceeding, let us first consider the term *sparse*. The term *sparse* refers to a measurable property of a vector. It means that the vector is in a sense small, but not in the length of the vector. Instead sparsity concerns the number of non-zero entries in the vector. For example, when an under-determined linear system is considered, a solution with a smaller number of non-zero entries is more desirable in sparse approximation. To measure sparsity, we use the following definition.

Definition 1 *For a given vector $x \in \Re^n$, we denote l_0-"norm" of x as*

$$\|x\|_0 = \textit{the number of nonzero entries (cardinality) of the vector } x.$$

When considering a sparse approximation problem, we seek a solution that is as sparse as possible. We may do this by regularizing a l_0-"norm" penalty or imposing a l_0-"norm" constraint to optimization problems.

We now describe some key settings where sparse solutions are desired.

1.1.1 Compressed sensing

Compressed sensing is a signal processing technique for efficiently acquiring and reconstructing a signal, by finding solutions to underdetermined linear systems (see, for example, [39, 123, 83, 114, 35, 94, 126]). This takes advantage of the signal's sparseness or compressibility in some domain, allowing the entire signal to be determined from relatively few measurements.

For example, we consider a problem of signal reconstruction. We first observe that many types of signals can be well-approximated by a sparse expansion in terms of a suitable basis, that is, by only a small number of non-zero coefficients. This is the key to the efficiency of many lossy compression techniques such as JPEG, MP3 etc. The signal is compressed by simply storing only the largest basis coefficients. When reconstructing the signal, the non-stored coefficients are set to zero. This is certainly a reasonable strategy when full information of the signal is available. However, the full information of the signal has to be acquired by a somewhat costly, lengthy or otherwise difficult measurement (sensing) procedure, which seems to be a waste of resources: first, large efforts are spent in order to obtain full information on the signal, and afterwards most of the information is thrown away at the compression stage. One natural question is whether there is a clever way of obtaining the compressed version of the signal more efficiently, by taking only a small number of measurements of the signal. It is not obvious at all whether this is possible since directly measuring the large coefficients requires to know a priori their locations. Quite surprisingly, compressed sensing provides nevertheless a way of reconstructing a compressed version of the original signal by taking only a small number of linear measurements. The given underdetermined linear system approximating the coefficients has infinitely many solutions. A naive approach to a reconstruction algorithm consists in searching for the sparsest vector that is consistent with the linear measurements.

Based on the discussion above, we can reconstruct the sparsest vector by solving the following optimization problem

$$\min_{x\in\Re^p}\{\|x\|_0 :\ Ax = b\}, \tag{1.1}$$

where $A \in \Re^{n\times p}$ is a known matrix and $b \in \Re^n$ is an observation vector. Obviously, (1.1) is a combinatorial l_0-problem, which unfortunately is NP-hard in general [92].

Nevertheless, the formulation (1.1) only models the CS problem well when the measurements and the signal are noise free. In the noisy case, it is more reasonable to formulate the CS problem as

$$\min_{x\in\Re^p}\left\{\frac{1}{2}\|Ax-b\|_2^2 :\ \|x\|_0 \le r\right\}, \tag{1.2}$$

or

$$\min_{x\in\Re^p}\left\{\frac{1}{2}\|Ax-b\|_2^2 + \nu\|x\|_0\right\} \tag{1.3}$$

for some integer $r \ge 0$ or real number $\nu \ge 0$ controlling the sparsity (or cardinality) of the solution.

1.1.2 Sparse inverse covariance selection

The sparse inverse covariance selection problem has numerous real-world applications such as speech recognition and gene network analysis (see, for example, [8, 49]). Solving this problem discovers the conditional independence in graphical models. The basic model for continuous data assumes that the observations have a multivariate Gaussian distribution with mean μ and covariance matrix $\mathbf{\Sigma}$. If the ijth component of $\mathbf{\Sigma}^{-1}$ is zero, then variables i and j are conditionally independent, given the other variables. Thus $\mathbf{\Sigma}^{-1}$ would be sparse if there are a large number of pairs of conditionally independent variables.

One popular formulation for sparse inverse covariance selection is to find an approximate sparse inverse covariance matrix while maximizing the log-likelihood (see, for example, [62]). Given a sample covariance matrix $\mathbf{\Sigma^t} \in \mathcal{S}^p_{++}$ and a set Ω consisting of pairs of known conditionally independent nodes, the sparse inverse covariance selection problem can be formulated as

$$\begin{aligned} \max_{X\succeq 0}\quad & \log\det X - \mathbf{\Sigma^t}\bullet X \\ \text{s.t.}\quad & \sum_{(i,j)\in\bar{\Omega}}\|X_{ij}\|_0 \le r, \\ & X_{ij} = 0 \quad \forall (i,j)\in\Omega, \end{aligned} \tag{1.4}$$

where $\bar{\Omega} = \{(i,j) : (i,j) \notin \Omega,\ i \neq j\}$, and $r \in [1, |\bar{\Omega}|]$ is an integer that controls the sparsity (or cardinality) of the solution.

1.1.3 Sparse logistic regression

Sparse logistic regression problem has many applications in machine learning, computer vision, data mining, bioinformatics, and neural signal processing (see, for example, [10, 129, 84, 103, 65, 104]). It 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, [99]).

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

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

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

$$\min_{v,w} \{l_{\text{avg}}(v, w) : \|w\|_0 \leq r\}, \tag{1.5}$$

where $r \in [1, p]$ is some integer for controlling the sparsity (or cardinality) of the solution.

1.1.4 More applications and summary

In addition, the similar ideas of compressed sensing have also been widely used in linear regression [124], for instance in the *lasso* method. It can effectively reduce the number of variables upon which the given solution is dependent.

During the last ten years, sparse *principal component analysis* (PCA) becomes an efficient tool in dimension reduction where a few sparse linear combinations of the random variables, so called *principal components* (PCs) are pursued so that their total explained variance is maximized (see, for example, [137, 117]). There are various sparse models for sparse PCA that have been proposed in literature [137, 42, 117, 77] (see Section 3.1 for more details).

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

$$\min_{x\in\mathcal{X}}\{f(x) : g(x) \le 0,\ h(x) = 0,\ \|x_J\|_0 \le r\}, \tag{1.6}$$

$$\min_{x\in\mathcal{X}}\{f(x) + \nu\|x_J\|_0 : g(x) \le 0,\ h(x) = 0\} \tag{1.7}$$

for some integer $r \ge 0$ or real number $\nu \ge 0$ controlling the sparsity (or cardinality) of the solution. Here $\mathcal{X}$ 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 is a subvector formed by the entries of x indexed by J. Given that $\|\cdot\|_0$ is an integer-valued, discontinuous and nonconvex function, it is generally hard to solve problems (1.6) and (1.7).

1.2 Existing approaches

In the literature, one popular approach for dealing with (1.6) and (1.7) is to replace $\|\cdot\|_0$ by the l_1-norm $\|\cdot\|_1$ and solve the resulting relaxation problems

$$\min_{x\in\mathcal{X}}\{f(x) : g(x) \le 0,\ h(x) = 0,\ \|x_J\|_1 \le \tau\}, \tag{1.8}$$

$$\min_{x\in\mathcal{X}}\{f(x) + \lambda\|x_J\|_1 : g(x) \le 0,\ h(x) = 0\} \tag{1.9}$$

for some positive real numbers τ and λ controlling the sparsity (or cardinality) of the solution (see, for example, [41, 99, 35, 124]).

For some applications such as compressed sensing, it has been shown in [23, 20] that under suitable conditions one can find the global optimal solutions of (1.6) and (1.7) by solving (1.8) and (1.9), respectively. For example, it has been shown in [23, 20] that the solution $\hat{x}$ to

$$\min_{x\in\Re^p}\{\|x\|_1 :\ Ax = b\}, \tag{1.10}$$

recovers the solution x^* to problem (1.1) exactly provided that 1) x^* is sufficiently sparse and 2) the matrix A obeys a condition known as the *restricted isometry property*.

To state their results, we first recall the concept of restricted isometry constants.

Definition 2 *For each integer $s = 1, 2, \ldots,$ define the isometry constant δ_s of a matrix A as the smallest number such that*

$$(1 - \delta_s)\|x\|_2^2 \le \|Ax\|_2^2 \le (1 + \delta_s)\|x\|_2^2$$

holds for all s-sparse vectors, where $\|\cdot\|_2$ denotes the Euclidean norm. A vector is said to be s-sparse if it has at most s nonzero entries.

By using this definition, they showed that

1. *If $\delta_{2s} < 1$, problem* (1.1) *has a unique s-sparse solution;*
2. *If $\delta_{2s} < \sqrt{2}-1$, the solution to problem* (1.10) *is that of problem* (1.1)*. In other words, the convex relaxation is exact.*

Besides l_1-norm relaxation, another relaxation approach has been recently proposed to solve problems (1.6) and (1.7) in which $\|\cdot\|_0$ is replaced by l_p-"norm" $\|\cdot\|_p$ for some $p \in (0,1)$ (see, for example, [34, 36, 37]). In general, the solution quality of these relaxation approaches may not be high. Indeed, for the example given below, the l_p relaxation approach for $p \in (0,1]$ fails to recover the sparse solution.

Example. 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. ∎

Because of the possible failure of the l_p relaxation approach for some $p \in (0,1]$, some algorithms are proposed in the literature for solving special cases of (1.6) and (1.7) directly. For example, the iterative hard thresholding algorithms [72, 11, 12] and matching pursuit algorithms [92, 125] are developed for solving the l_0-regularized least squares problems arising in compressed sensing, but they cannot be applied to solve the general l_0 minimization problems (1.6) and (1.7).

1.3 Proposed methods and contributions

In this thesis, we propose two kinds of methods which are a novel augmented Lagrangian method and penalty decomposition methods, for solving the general sparse approximation problems. The augmented Lagrangian method can be applied to solve the general l_1 relaxation sparse approximation problems (i.e., (1.8) and (1.9)), while the penalty decomposition methods solve the original sparse approximation problems (i.e., (1.6) and (1.7)) directly. In addition, we study the convergence of the proposed methods.

1.3.1 The augmented Lagrangian method

We first propose a novel augmented Lagrangian method for solving a class of nonsmooth constrained optimization problems. These problems can be written as

$$\begin{array}{ll} \min & f(x) + P(x) \\ \text{s.t.} & g(x) \leq 0, \\ & h(x) = 0, \\ & x \in \mathcal{X}, \end{array} \tag{1.11}$$

where the function $P : \Re^n \to \Re$ is convex but not necessarily smooth. We can easily observe that both (1.8) and (1.9) are in the form of (1.11). Thus the proposed augmented Lagrangian method can be suitably applied to solve (1.8) and (1.9).

The proposed augmented Lagrangian method in this thesis differs from the classical augmented Lagrangian method in that: i) the values of the augmented Lagrangian functions at their approximate minimizers given by the method are bounded from above; and ii) the magnitude of penalty parameters outgrows that of Lagrangian multipliers (see Subsection 2.2 for details). In addition, we show that this method converges to a *feasible* point, and moreover it converges to a first-order stationary point under some regularity assumptions. We should mention that the aforementioned two novel properties of the proposed augmented Lagrangian method are crucial in ensuring convergence both theoretically and practically. In fact, we observed in our experiments that when either one of these properties are dropped, the resulting method almost always fails to converge to even a feasible point when applied to our formulation of sparse PCA, which is in the form of (1.11).

In addition, we also propose two nonmonotone gradient methods for minimizing a class of nonsmooth functions over a closed convex set, which can be suitably applied to the

subproblems arising in our augmented Lagrangian method. We further establish global convergence and, under a local Lipschitzian error bounds assumption [128], local linear rate of convergence for these gradient methods.

We next apply the proposed augmented Lagrangian method to solve our new formulation for sparse PCA. Sparse PCA has been an active research topic for more than a decade. The existing methods [76, 14, 78, 137, 42, 117, 95, 40, 77] can produce sparse PCs but the nice properties of the standard PCs are generally lost (see Subsection 3.1 for details). The proposed new formulation (see Subsection 3.2 for details) takes into account three nice properties of the standard PCA, that is, maximal total explained variance, uncorrelation of PCs, and orthogonality of loading vectors. Moreover, we also explore the connection of this formulation with the standard PCA and show that it can be viewed as a certain perturbation of the standard PCA. We then compare the sparse PCA approach proposed in this thesis with several existing methods [137, 42, 117, 77] on synthetic [137], Pitprops [74], and gene expression data [38], respectively. The computational results demonstrate that the sparse PCs obtained by our approach substantially outperform those by the other methods in terms of total explained variance, correlation of PCs, and orthogonality of loading vectors. In addition, the experiments on random data show that our method is capable of solving large-scale problems within very reasonable amount of time and it is also fairly stable.

1.3.2 The penalty decomposition methods

Because of the drawbacks for l_1 relaxation approaches mentioned in Section 1.2, in this thesis we also propose penalty decomposition (PD) methods for directly solving problems (1.6) and (1.7) 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 (1.6) and (1.7). 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 block coordinate minimizer 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 when solutions of same cardinality

are sought, our approach applied to the l_0-based models generally has better solution quality and/or speed than the existing approaches that are applied to the corresponding l_1-based models.

Finally, we adapt the PD method to solve our proposed wavelet frame based image restoration problem. The basic idea for wavelet frame based approaches is that images can be sparsely approximated by properly designed wavelet frames, and hence, the wavelet frame based image restoration problem can be formulated as a variant of (1.6), see Subsection 5.1 for details. Some convergence analysis of the adapted PD method for this problem are provided. Numerical results show that the proposed model solved by the PD method can generate images with better quality than those obtained by the existing l_1 relaxation approaches in terms of restoring sharp features as well as maintaining smoothness of the recovered images.

1.4 Organization

The rest of this thesis is organized as follows. In Chapter 2, we propose the augmented Lagrangian method for solving a class of nonsmooth constrained optimization problems. We also propose two nonmonotone gradient methods for minimizing a class of nonsmooth functions over a closed convex set, which can be suitably applied to the subproblems arising in our augmented Lagrangian method. In Chapter 3, we first propose the new formulation for sparse PCA and then apply the proposed augmented Lagrangian method to solve this new formulation. In addition, we compare the sparse PCA approach proposed in this thesis with several existing methods on synthetic, Pitprops, and gene expression data, respectively. In Chapter 4, we first establish the first-order optimality conditions for general l_0 minimization problems and study a class of special l_0 minimization problems. We then develop the PD methods for general l_0 minimization problems. Finally, 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. We adapt the PD method to solve our proposed wavelet frame based image restoration problem in Chapter 5.

1.5 Notations

In this thesis, all vector spaces are assumed to be finite dimensional. The symbols $\Re^n$ and $\Re^n_+$ (resp., $\Re^n_-$) denote the n-dimensional Euclidean space and the nonnegative (resp., nonpositive) orthant of $\Re^n$, respectively, and $\Re_{++}$ denotes the set of positive real numbers. 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. 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. We define x_L 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 set difference 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 $\mathcal{S}^n$. Additionally, $\mathcal{D}^n$ denotes the space of $n \times n$ diagonal matrices. For a real matrix X, we denote by $|X|$ the absolute value of X, that is, $|X|_{ij} = |X_{ij}|$ for all ij, and by $\mathrm{sign}(X)$ the sign of X whose ijth entry equals the sign of X_{ij} for all ij. Also, the nonnegative part of X is denoted by $[X]^+$ whose ijth entry is given by $\max\{0, X_{ij}\}$ for all ij. The rank of X is denoted by $\mathrm{rank}(X)$. Further, the identity matrix and the all-ones matrix are denoted by I and E, respectively, 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., $\mathcal{S}^n_{++}$). For any X, $Y \in \mathcal{S}^n$, we write $X \succeq Y$ to mean $X - Y \succeq 0$. Given matrices X and Y in $\Re^{m\times n}$, the standard inner product is defined by $X \bullet Y := \mathrm{Tr}(XY^T)$, where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix, and the component-wise product is denoted by $X \odot Y$, whose ijth entry is $X_{ij}Y_{ij}$ for all ij. The Euclidean norm is defined by $\|\cdot\|$ as is its associated operator norm unless it is explicitly stated otherwise. The minimal (resp., maximal) eigenvalue of an $n\times n$ symmetric matrix X are denoted by $\lambda_{\min}(X)$ (resp., $\lambda_{\max}(X)$), respectively, and $\lambda_i(X)$ denotes its ith largest eigenvalue for $i = 1, \ldots, n$. The operator is defined by $\mathscr{D}$ 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 $i = 1, \ldots, n$. Let $f : \Re^n \to \Re$ be a proper convex function and x be a point in the domain of f, the subdifferential of f at x is denoted by $\partial f(x)$. Let $\mathcal{U}$ be a real vector space. Given a closed convex set $C \subseteq \mathcal{U}$, let $\mathrm{dist}(\cdot, C) : \mathcal{U} \to \Re_+$ denote

the distance function to C measured in terms of $\|\cdot\|$, that is,

$$\operatorname{dist}(u, C) := \inf_{\tilde{u} \in C} \|u - \tilde{u}\| \quad \forall u \in \mathcal{U}. \tag{1.12}$$

Chapter 2

An augmented Lagrangian approach

In this chapter we propose a novel augmented Lagrangian method for a class of nonsmooth constrained nonlinear programming problems. In particular, we study first-order optimality conditions and then we develop an augmented Lagrangian method and establish its global convergence. In addition, we propose two nonmonotone gradient methods for minimizing a class of nonsmooth functions over a closed convex set, which can be suitably applied to the subproblems arising in our augmented Lagrangian method. We also establish global and local convergence for these gradient methods.

This chapter is based on the paper [90] co-authored with Zhaosong Lu.

2.1 First-order optimality conditions

In this subsection we introduce a class of nonsmooth constrained nonlinear programming problems and study first-order optimality conditions for them.

Consider the nonlinear programming problem

$$\begin{array}{ll} \min & f(x) + P(x) \\ \text{s.t.} & g_i(x) \le 0, \quad i = 1, \ldots, m, \\ & h_i(x) = 0, \quad i = 1, \ldots, p, \\ & x \in X. \end{array} \tag{2.1}$$

We assume that the functions $f : \Re^n \to \Re$, $g_i : \Re^n \to \Re$, $i = 1, \ldots, m$, and $h_i : \Re^n \to \Re$,

$i = 1, \ldots, p$, are continuously differentiable, and that the function $P : \Re^n \to \Re$ is convex but not necessarily smooth, and that the set $X \subseteq \Re^n$ is closed and convex. For convenience of the subsequent presentation, we denote by Ω the feasible region of problem (2.1).

For the case where P is a smooth function, the first-order optimality conditions for problem (2.1) have been well studied in literature (see, for example, Theorem 3.25 of [113], but there is little study when P is a nonsmooth convex function. We next aim to establish first-order optimality conditions for problem (2.1). Before proceeding, we describe a general constraint qualification condition for (2.1), that is, Robinson's condition that was proposed in [107].

Let $x \in \Re^n$ be a feasible point of problem (2.1). We denote the set of active inequality constraints at x as

$$\mathcal{A}(x) = \{1 \leq i \leq m : g_i(x) = 0\}.$$

In addition, x is said to satisfy *Robinson's condition* if

$$\left\{ \begin{bmatrix} g'(x)d - v \\ h'(x)d \end{bmatrix} : d \in T_X(x), v \in \Re^m, v_i \leq 0, i \in \mathcal{A}(x) \right\} = \Re^m \times \Re^p, \tag{2.2}$$

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. Other equivalent expressions of Robinson's condition can be found, for example, in [107, 108, 113].

The following proposition demonstrates that Robinson's condition is indeed a constraint qualification condition for problem (2.1). For the sake of completeness, we include a brief proof for it.

Proposition 2.1.1 *Given a feasible point $x \in \Re^n$ of problem (2.1), let $T_\Omega(x)$ be the tangent cone to Ω at x, and $(T_\Omega(x))^\circ$ be its polar cone. If Robinson's condition (2.2) holds at x, then*

$$\begin{aligned} T_\Omega(x) &= \left\{ d \in T_X(x) : \begin{array}{ll} d^T \nabla g_i(x) \leq 0, & i \in \mathcal{A}(x), \\ d^T \nabla h_i(x) = 0, & i = 1, \ldots, p \end{array} \right\}, \\ (T_\Omega(x))^\circ &= \left\{ \sum_{i \in \mathcal{A}(x)} \lambda_i \nabla g_i(x) + \sum_{i=1}^{p} \mu_i \nabla h_i(x) + N_X(x) : \lambda \in \Re^m_+, \ \mu \in \Re^p \right\}, \end{aligned} \tag{2.3}$$

where $T_X(x)$ and $N_X(x)$ are the tangent and normal cones to X at x, respectively.

Proof. By Theorem A.10 of [113], we see that Robinson's condition (2.2) implies that the assumption of Theorem 3.15 of [113] is satisfied with

$$x_0 = x, \quad X_0 = X, \quad Y_0 = \Re^m_- \times \Re^p, \quad g(\cdot) = (g_1(\cdot); \ldots; g_m(\cdot); h_1(\cdot); \ldots; h_p(\cdot)).$$

The first statement then follows from Theorem 3.15 of [113] with the above x_0, X_0, Y_0 and $g(\cdot)$. Further, let $A(x)$ denote the matrix whose rows are the gradients of all active constraints at x in the same order as they appear in (2.1). Then, Robinson's condition (2.2) implies that the assumptions of Theorem 2.36 of [113] are satisfied with

$$A = A(x), \quad K_1 = T_X(x), \quad K_2 = \Re^{|\mathcal{A}(x)|}_- \times \Re^p.$$

Let $K = \{d \in K_1 : Ad \in K_2\}$. Then, it follows from Theorem 2.36 of [113] that

$$(T_\Omega(x))^\circ = K^\circ = K_1^\circ + \{A^T\xi : \xi \in K_2^\circ\},$$

which together with the identity $(T_X(x))^\circ = N_X(x)$ and the definitions of A, K_1 and K_2, implies that the second statement holds. ∎

We are now ready to establish first-order optimality conditions for problem (2.1).

Theorem 2.1.2 *Let $x^* \in \Re^n$ be a local minimizer of problem (2.1). Assume that Robinson's condition (2.2) is satisfied at x^*. Then there exist Lagrange multipliers $\lambda \in \Re^m_+$ and $\mu \in \Re^p$ such that*

$$0 \in \nabla f(x^*) + \partial P(x^*) + \sum_{i=1}^m \lambda_i \nabla g_i(x^*) + \sum_{i=1}^p \mu_i \nabla h_i(x^*) + N_X(x^*), \tag{2.4}$$

and

$$\lambda_i g_i(x^*) = 0, \quad i = 1, \ldots, m. \tag{2.5}$$

Moreover, the set of Lagrange multipliers $(\lambda, \mu) \in \Re^m_+ \times \Re^p$ satisfying the above conditions, denoted by $\Lambda(x^*)$, is convex and compact.

Proof. We first show that

$$d^T \nabla f(x^*) + P'(x^*; d) \geq 0 \quad \forall d \in T_\Omega(x^*). \tag{2.6}$$

Let $d \in T_\Omega(x^*)$ be arbitrarily chosen. Then, there exist sequences $\{x^k\}_{k=1}^\infty \subseteq \Omega$ and $\{t_k\}_{k=1}^\infty \subseteq \Re_{++}$ such that $t_k \downarrow 0$ and

$$d = \lim_{k\to\infty} \frac{x^k - x^*}{t_k}.$$

Thus, we have $x^k = x^* + t_k d + o(t_k)$. Using this relation along with the fact that the function f is differentiable and P is convex in $\Re^n$, we can have

$$f(x^* + t_k d) - f(x^k) = o(t_k), \quad P(x^* + t_k d) - P(x^k) = o(t_k), \tag{2.7}$$

where the first equality follows from the Mean Value Theorem while the second one comes from Theorem 10.4 of [109]. Clearly, $x^k \to x^*$. This together with the assumption that x^* is a local minimizer of (2.1), implies that

$$f(x^k) + P(x^k) \geq f(x^*) + P(x^*) \tag{2.8}$$

when k is sufficiently large. In view of (2.7) and (2.8), we obtain that

$$\begin{aligned} d^T \nabla f(x^*) + P'(x^*; d) &= \lim_{k\to\infty} \frac{f(x^* + t_k d) - f(x^*)}{t_k} + \lim_{k\to\infty} \frac{P(x^* + t_k d) - P(x^*)}{t_k}, \\ &= \lim_{k\to\infty} \left[\frac{f(x^k) + P(x^k) - f(x^*) - P(x^*)}{t_k} + \frac{o(t_k)}{t_k} \right], \\ &= \lim_{k\to\infty} \frac{f(x^k) + P(x^k) - f(x^*) - P(x^*)}{t_k} \ \geq \ 0, \end{aligned}$$

and hence (2.6) holds.

For simplicity of notations, let $T_\Omega^\circ = (T_\Omega(x^*))^\circ$ and $S = -\nabla f(x^*) - \partial P(x^*)$. We next show that $S \cap T_\Omega^\circ \neq \emptyset$. Suppose for contradiction that $S \cap T_\Omega^\circ = \emptyset$. This together with the fact that S and T_Ω° are nonempty closed convex sets and S is bounded, implies that there exists some $d \in \Re^n$ such that $d^T y \leq 0$ for any $y \in T_\Omega^\circ$, and $d^T y \geq 1$ for any $y \in S$. Clearly, we see that $d \in (T_\Omega^\circ)^\circ = T_\Omega(x^*)$, and

$$\begin{aligned} 1 \leq \inf_{y\in S} d^T y &= \inf_{z\in\partial P(x^*)} d^T(-\nabla f(x^*) - z) = -d^T \nabla f(x^*) - \sup_{z\in\partial P(x^*)} d^T z \\ &= -d^T \nabla f(x^*) - P'(x^*; d), \end{aligned}$$

which contradicts (2.6). Hence, we have $S \cap T_\Omega^\circ \neq \emptyset$. Using this relation, (2.3), the definitions of S and $\mathcal{A}(x^*)$, and letting $\lambda_i = 0$ for $i \notin \mathcal{A}(x^*)$, we easily see that (2.4) and (2.5) hold.

In view of the fact that $\partial P(x^*)$ and $N_X(x^*)$ are closed and convex, and moreover $\partial P(x^*)$ is bounded, we know that $\partial P(x^*) + N_X(x^*)$ is closed and convex. Using this result, it is straightforward to see that $\Lambda(x^*)$ is closed and convex. We next show that $\Lambda(x^*)$ is bounded. Suppose for contradiction that $\Lambda(x^*)$ is unbounded. Then, there exists a sequence $\{(\lambda^k, \mu^k)\}_{k=1}^\infty \subseteq \Lambda(x^*)$ such that $\|(\lambda^k, \mu^k)\| \to \infty$, and

$$0 = \nabla f(x^*) + z^k + \sum_{i=1}^m \lambda_i^k \nabla g_i(x^*) + \sum_{i=1}^p \mu_i^k \nabla h_i(x^*) + v^k \tag{2.9}$$

for some $\{z^k\}_{k=1}^{\infty} \subseteq \partial P(x^*)$ and $\{v^k\}_{k=1}^{\infty} \subseteq N_X(x^*)$. Let $(\bar{\lambda}^k, \bar{\mu}^k) = (\lambda^k, \mu^k)/\|(\lambda^k, \mu^k)\|$.

By passing to a subsequence if necessary, we can assume that $(\bar{\lambda}^k, \bar{\mu}^k) \to (\bar{\lambda}, \bar{\mu})$. We clearly see that $\|(\bar{\lambda}, \bar{\mu})\| = 1$, $\bar{\lambda} \in \Re^m_+$, and $\bar{\lambda}_i = 0$ for $i \notin \mathcal{A}(x^*)$. Note that $\partial P(x^*)$ is bounded and $N_X(x^*)$ is a closed cone. In view of this fact, and upon dividing both sides of (2.9) by $\|(\lambda^k, \mu^k)\|$ and taking limits on a subsequence if necessary, we obtain that

$$0 = \sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(x^*) + \sum_{i=1}^{p} \bar{\mu}_i \nabla h_i(x^*) + \bar{v} \tag{2.10}$$

for some $\bar{v} \in N_X(x^*)$. Since Robinson's condition (2.2) is satisfied at x^*, there exist $d \in T_X(x^*)$ and $v \in \Re^m$ such that $v_i \leq 0$ for $i \in \mathcal{A}(x^*)$, and

$$\begin{aligned} d^T \nabla g_i(x^*) - v_i &= -\bar{\lambda}_i \quad \forall i \in \mathcal{A}(x^*), \\ d^T \nabla h_i(x^*) &= -\bar{\mu}_i, \quad i = 1, \ldots, p. \end{aligned}$$

Using these relations, (2.10) and the fact that $d \in T_X(x^*)$, $\bar{v} \in N_X(x^*)$, $\bar{\lambda} \in \Re^m_+$, and $\bar{\lambda}_i = 0$ for $i \notin \mathcal{A}(x^*)$, we have

$$\begin{aligned} \sum_{i=1}^{m} \bar{\lambda}_i^2 + \sum_{i=1}^{p} \bar{\mu}_i^2 &\leq -\sum_{i=1}^{m} \bar{\lambda}_i d^T \nabla g_i(x^*) - \sum_{i=1}^{p} \bar{\mu}_i d^T \nabla h_i(x^*), \\ &= -d^T \left(\sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(x^*) + \sum_{i=1}^{p} \bar{\mu}_i \nabla h_i(x^*) \right) = d^T \bar{v} \leq 0. \end{aligned}$$

It yields $(\bar{\lambda}, \bar{\mu}) = (0, 0)$, which contradicts the identity $\|(\bar{\lambda}, \bar{\mu})\| = 1$. Thus, $\Lambda(x^*)$ is bounded. ■

2.2 Augmented Lagrangian method for (2.1)

For a convex program, it is known that under some mild assumptions, any accumulation point of the sequence generated by the classical augmented Lagrangian method is an optimal solution (e.g., see Section 6.4.3 of [113]). Nevertheless, when problem (2.1) is a nonconvex program, especially when the function h_i is not affine or g_i is nonconvex, the classical augmented Lagrangian method may not even converge to a feasible point, that is, any accumulation point of the sequence generated by the method may violate some constraints of (2.1). We actually observed in our experiments that this ill phenomenon almost always happens when the classical augmented Lagrangian method is applied to our proposed formulation of sparse PCA. To alleviate this drawback, we propose a novel augmented Lagrangian method for problem (2.1) and establish its global convergence in this subsection.

Throughout this subsection, we make the following assumption for problem (2.1).

Assumption 1 *Problem (2.1) is feasible, and moreover at least a feasible solution, denoted by x^{feas}, is known.*

It is well-known that for problem (2.1) the associated augmented Lagrangian function $L_\varrho(x,\lambda,\mu):\Re^n\times\Re^m\times\Re^p\to\Re$ is given by

$$L_\varrho(x,\lambda,\mu) := w(x) + P(x), \tag{2.11}$$

where

$$w(x) := f(x) + \frac{1}{2\varrho}(\|[\lambda+\varrho g(x)]^+\|^2 - \|\lambda\|^2) + \mu^T h(x) + \frac{\varrho}{2}\|h(x)\|^2, \tag{2.12}$$

and $\varrho>0$ is a penalty parameter (e.g., see [7, 113]). Roughly speaking, an augmented Lagrangian method, when applied to problem (2.1), solves a sequence of subproblems in the form of

$$\min_{x\in X} L_\varrho(x,\lambda,\mu)$$

while updating the Lagrangian multipliers (λ,μ) and the penalty parameter ϱ.

Let x^{feas} be a known feasible point of (2.1) (see Assumption 1). We now describe the algorithm framework of a novel augmented Lagrangian method as follows.

Algorithm framework of augmented Lagrangian method:

Let $\{\epsilon_k\}$ be a positive decreasing sequence. Let $\lambda^0\in\Re^m_+$, $\mu^0\in\Re^p$, $\varrho_0>0$, $\tau>0$, $\sigma>1$ be given. Choose an arbitrary initial point $x^0_{\text{init}}\in X$ and constant $\Upsilon\geq\max\{f(x^{\text{feas}}), L_{\varrho_0}(x^0_{\text{init}},\lambda^0,\mu^0)\}$. Set $k=0$.

1) Find an approximate solution $x^k\in X$ for the subproblem

$$\min_{x\in X} L_{\varrho_k}(x,\lambda^k,\mu^k) \tag{2.13}$$

such that

$$\text{dist}\Big(-\nabla w(x^k), \partial P(x^k)+N_X(x^k)\Big)\leq\epsilon_k,\quad L_{\varrho_k}(x^k,\lambda^k,\mu^k)\leq\Upsilon. \tag{2.14}$$

2) Update Lagrange multipliers according to

$$\lambda^{k+1} := [\lambda^k+\varrho_k g(x^k)]^+,\quad \mu^{k+1} := \mu^k+\varrho_k h(x^k). \tag{2.15}$$

3) Set $\varrho_{k+1} := \max\left\{\sigma\varrho_k, \|\lambda^{k+1}\|^{1+\tau}, \|\mu^{k+1}\|^{1+\tau}\right\}$.

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

end

The above augmented Lagrangian method differs from the classical augmented Lagrangian method in that: i) the values of the augmented Lagrangian functions at their approximate minimizers given by the method are uniformly bounded from above (see Step 1)); and ii) the magnitude of penalty parameters outgrows that of Lagrangian multipliers (see Step 3)). These two novel properties are crucial in ensuring the convergence of our augmented Lagrangian method both theoretically and practically. In fact, we observed in our experiments that when one or both of these steps are replaced by the counterparts of the classical augmented Lagrangian method, the resulting method almost always fails to converge to even a feasible point as applied to our proposed formulation of sparse PCA.

To make the above augmented Lagrangian method complete, we need to address how to find an approximate solution $x^k \in X$ for subproblem (2.13) satisfying (2.14) as required in Step 1). We will leave this discussion to the end of this subsection. For the time being, we establish the main convergence result regarding this method for solving problem (2.1).

Theorem 2.2.1 *Assume that $\epsilon_k \to 0$. Let $\{x^k\}$ be the sequence generated by the above augmented Lagrangian method satisfying (2.14). Suppose that a subsequence $\{x^k\}_{k\in K}$ converges to x^*. Then, the following statements hold:*

(a) x^* is a feasible point of problem (2.1);

(b) Further, if Robinson's condition (2.2) is satisfied at x^*, then the subsequence $\{(\lambda^{k+1}, \mu^{k+1})\}_{k\in K}$ is bounded, and each accumulation point (λ^*, μ^*) of $\{(\lambda^{k+1}, \mu^{k+1})\}_{k\in K}$ is the vector of Lagrange multipliers satisfying the first-order optimality conditions (2.4)-(2.5) at x^*.

Proof. In view of (2.11), (2.12) and the second relation in (2.14), we have

$$f(x^k) + P(x^k) + \frac{1}{2\varrho_k}(\|[\lambda^k + \varrho_k g(x^k)]^+\|^2 - \|\lambda^k\|^2) + (\mu^k)^T h(x^k) + \frac{\varrho_k}{2}\|h(x^k)\|^2 \le \Upsilon \quad \forall k.$$

It follows that

$$\|[\lambda^k/\varrho_k + g(x^k)]^+\|^2 + \|h(x^k)\|^2 \le 2[\Upsilon - f(x^k) - g(x^k) - (\mu^k)^T h(x^k)]/\varrho_k + (\|\lambda_k\|/\varrho_k)^2.$$

Noticing that $\varrho_0 > 0$ $\tau > 0$, and $\varrho_{k+1} = \max\{\sigma\varrho_k, \|\lambda^{k+1}\|^{1+\tau}, \|\mu^{k+1}\|^{1+\tau}\}$ for $k \geq 0$, we can observe that $\varrho_k \to \infty$ and $\|(\lambda^k, \mu^k)\|/\varrho_k \to 0$. We also know that $\{x^k\}_{k\in K} \to x^*$, $\{g(x^k)\}_{k\in K} \to g(x^*)$ and $\{h(x^k)\}_{k\in K} \to h(x^*)$. Using these results, and upon taking limits as $k \in K \to \infty$ on both sides of the above inequality, we obtain that

$$\|[g(x^*)]^+\|^2 + \|h(x^*)\|^2 \leq 0,$$

which implies that $g(x^*) \leq 0$ and $h(x^*) = 0$. We also know that $x^* \in X$. It thus follows that statement (a) holds.

We next show that statement (b) also holds. Using (2.13), (2.11), (2.12), (2.15), and the first relation in (2.14), we have

$$\|\nabla f(x^k) + \sum_{i=1}^{m} \lambda_i^{k+1} \nabla g_i(x^k) + \sum_{i=1}^{p} \mu_i^{k+1} \nabla h_i(x^k) + z^k + v^k\| \leq \epsilon_k \tag{2.16}$$

for some $z^k \in \partial P(x^k)$ and $v^k \in N_X(x^k)$. Suppose for contradiction that the subsequence $\{(\lambda^{k+1}, \mu^{k+1})\}_{k\in K}$ is unbounded. By passing to a subsequence if necessary, we can assume that $\{(\lambda^{k+1}, \mu^{k+1})\}_{k\in K} \to \infty$. Let $(\bar{\lambda}^{k+1}, \bar{\mu}^{k+1}) = (\lambda^{k+1}, \mu^{k+1})/\|(\lambda^{k+1}, \mu^{k+1})\|$ and $\bar{v}^k = v^k/\|(\lambda^{k+1}, \mu^{k+1})\|$. Recall that $\{x^k\}_{k\in K} \to x^*$. It together with Theorem 6.2.7 of [73] implies that $\cup_{k\in K}\partial P(x^k)$ is bounded, and so is $\{z^k\}_{k\in K}$. In addition, $\{g(x^k)\}_{k\in K} \to g(x^*)$ and $\{h(x^k)\}_{k\in K} \to h(x^*)$. Then, we can observe from (2.16) that $\{\bar{v}^k\}_{k\in K}$ is bounded. Without loss of generality, assume that $\{(\bar{\lambda}^{k+1}, \bar{\mu}^{k+1})\}_{k\in K} \to (\bar{\lambda}, \bar{\mu})$ and $\{\bar{v}^k\}_{k\in K} \to \bar{v}$ (otherwise, one can consider their convergent subsequences). Clearly, $\|(\bar{\lambda}, \bar{\mu})\| = 1$. Dividing both sides of (2.16) by $\|(\lambda^{k+1}, \mu^{k+1})\|$ and taking limits as $k \in k \to \infty$, we obtain that

$$\sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(x^*) + \sum_{i=1}^{p} \bar{\mu}_i \nabla h_i(x^*) + \bar{v} = 0. \tag{2.17}$$

Further, using the identity $\lambda^{k+1} = [\lambda^k + \varrho_k g(x^k)]^+$ and the fact that $\varrho_k \to \infty$ and $\|\lambda^k\|/\varrho_k \to 0$, we observe that $\lambda^{k+1} \in \Re^m_+$ and $\lambda_i^{k+1} = 0$ for $i \notin \mathcal{A}(x^*)$ when $k \in K$ is sufficiently large, which imply that $\bar{\lambda} \in \Re^m_+$ and $\bar{\lambda}_i = 0$ for $i \notin \mathcal{A}(x^*)$. Moreover, we have $\bar{v} \in N_X(x^*)$ since $N_X(x^*)$ is a closed cone. Using these results, (2.17), Robinson's condition (2.2) at x^*, and a similar argument as that in the proof of Theorem 2.1.2, we can obtain that $(\bar{\lambda}, \bar{\mu}) = (0, 0)$, which contradicts the identity $\|(\bar{\lambda}, \bar{\mu})\| = 1$. Therefore, the subsequence $\{(\lambda^{k+1}, \mu^{k+1})\}_{k\in K}$ is bounded. Using this result together with (2.16) and the fact $\{z^k\}_{k\in K}$ is bounded, we immediately see that $\{v^k\}_{k\in K}$ is bounded. Using semicontinuity of $\partial P(\cdot)$ and $N_X(\cdot)$ (see

Theorem 24.4 of [109] and Lemma 2.42 of [113]), and the fact $\{x^k\}_{k\in K} \to x^*$, we conclude that every accumulation point of $\{z^k\}_{k\in K}$ and $\{v^k\}_{k\in K}$ belongs to $\partial P(x^*)$ and $N_X(x^*)$, respectively. Using these results and (2.16), we further see that for every accumulation point (λ^*, μ^*) of $\{(\lambda^{k+1}, \mu^{k+1})\}_{k\in K}$, there exists some $z^* \in \partial P(x^*)$ and $v^* \in N_X(x^*)$ such that

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) + \sum_{i=1}^{p} \mu_i^* \nabla h_i(x^*) + z^* + v^* = 0.$$

Moreover, using the identity $\lambda^{k+1} = [\lambda^k + \varrho_k g(x^k)]^+$ and the fact that $\varrho_k \to \infty$ and $\|\lambda^k\|/\varrho_k \to 0$, we easily see that $\lambda^* \in \Re^m_+$ and $\lambda_i^* = 0$ for $i \notin \mathcal{A}(x^*)$. Thus, (λ^*, μ^*) satisfies the first-order optimality conditions (2.4)-(2.5) at x^*. ∎

Before ending this subsection, we now briefly discuss how to find an approximate solution $x^k \in X$ for subproblem (2.13) satisfying (2.14) as required in Step 1) of the above augmented Lagrangian method. In particular, we are interested in applying the nonmonotone gradient methods proposed in Subsection 2.3 to (2.13). As shown in Subsection 2.3 (see Theorems 2.3.6 and 2.3.10), these methods are able to find an approximate solution $x^k \in X$ satisfying the first relation of (2.14). Moreover, if an initial point for these methods is properly chosen, the obtained approximate solution x^k also satisfies the second relation of (2.14). For example, given $k \geq 0$, let $x^k_{\text{init}} \in X$ denote the initial point for solving the kth subproblem (2.13), and we define x^k_{init} for $k \geq 1$ as follows

$$x^k_{\text{init}} = \begin{cases} x^{\text{feas}}, & \text{if } L_{\varrho_k}(x^{k-1}, \lambda^k, \mu^k) > \Upsilon; \\ x^{k-1}, & \text{otherwise,} \end{cases}$$

where x^{k-1} is the approximate solution to the $(k-1)$th subproblem (2.13) satisfying (2.14) (with k replaced by $k-1$). Recall from Assumption 1 that x^{feas} is a feasible solution of (2.1). Thus, $g(x^{\text{feas}}) \leq 0$, and $h(x^{\text{feas}}) = 0$, which together with (2.11), (2.12) and the definition of Υ implies that

$$L_{\varrho_k}(x^{\text{feas}}, \lambda^k, \mu^k) \;\leq\; f(x^{\text{feas}}) \;\leq\; \Upsilon.$$

It follows from this inequality and the above choice of x^k_{init} that $L_{\varrho_k}(x^k_{\text{init}}, \lambda^k, \mu^k) \leq \Upsilon$. Additionally, the nonmonotone gradient methods proposed in Subsection 2.3 possess a natural property that the objective function values at all subsequent iterates are bounded above by the one at the initial point. Therefore, we have

$$L_{\varrho_k}(x^k, \lambda^k, \mu^k) \;\leq\; L_{\varrho_k}(x^k_{\text{init}}, \lambda^k, \mu^k) \;\leq\; \Upsilon,$$

and so the second relation of (2.14) is satisfied at x^k.

2.3 Nonmonotone gradient methods for nonsmooth minimization

In this subsection we propose two nonmonotone gradient methods for minimizing a class of nonsmooth functions over a closed convex set, which can be suitably applied to the subproblems arising in our augmented Lagrangian method detailed in Subsection 2.2. We also establish global convergence and local linear rate of convergence for these methods.

Throughout this subsection, we consider the following problem

$$\min_{x\in X}\{F(x) := f(x) + P(x)\}, \tag{2.18}$$

where $f : \Re^n \to \Re$ is continuously differentiable, $P : \Re^n \to \Re$ is convex but not necessarily smooth, and $X \subseteq \Re^n$ is closed and convex.

In the literature [128, 134, 97, 6], several gradient methods were proposed for solving problem (2.18) or its special case. In particular, Tseng and Yun [128] studied a block coordinate descent method for (2.18). Under the assumption that the gradient of f is Lipschitz continuous, Wright et al. [134] proposed a globally convergent nonmonotone gradient method for (2.18). In addition, for the case where f is convex and its gradient is Lipschitz continuous, Nesterov [97] and Beck and Teboulle [6] developed optimal gradient methods for (2.18). In this subsection, we propose two nonmonotone gradient methods for (2.18). These two methods are closely related to the ones proposed in [134] and [128], but they are not the same (see the remarks below for details). In addition, these methods can be viewed as an extension of the well-known projected gradient methods studied in [9] for smooth problems, but the methods proposed in [134] and [128] cannot. Before proceeding, we introduce some notations and establish some technical lemmas as follows that will be used subsequently.

We say that $x \in \Re^n$ is a *stationary point* of problem (2.18) if $x \in X$ and

$$0 \in \nabla f(x) + \partial P(x) + N_X(x). \tag{2.19}$$

Given a point $x \in \Re^n$ and $H \succ 0$, we denote by $d_H(x)$ the solution of the following problem:

$$d_H(x) := \arg\min_d \left\{\nabla f(x)^T d + \frac{1}{2} d^T H d + P(x+d) :\ x + d \in X\right\}. \tag{2.20}$$

The following lemma provides an alternative characterization of stationarity that will be used in our subsequent analysis.

Lemma 2.3.1 *For any $H \succ 0$, $x \in X$ is a stationary point of problem (2.18) if and only if $d_H(x) = 0$.*

Proof. We first observe that (2.20) is a convex problem, and moreover its objective function is strictly convex. The conclusion of this lemma immediately follows from this observation and the first-order optimality condition of (2.20). ∎

The next lemma shows that $\|d_H(x)\|$ changes not too fast with H. It will be used to prove Theorems 2.3.7 and 2.3.11.

Lemma 2.3.2 *For any $x \in \Re^n$, $H \succ 0$, and $\tilde{H} \succ 0$, let $d = d_H(x)$ and $\tilde{d} = d_{\tilde{H}}(x)$. Then*

$$\|\tilde{d}\| \leq \frac{1 + \lambda_{\max}(Q) + \sqrt{1 - 2\lambda_{\min}(Q) + \lambda_{\max}(Q)^2}}{2\lambda_{\min}(\tilde{H})} \lambda_{\max}(H)\|d\|, \tag{2.21}$$

where $Q = H^{-1/2}\tilde{H}H^{-1/2}$.

Proof. The conclusion immediately follows from Lemma 3.2 of [128] with $J = \{1, \ldots, n\}$, $c = 1$, and $P(x) := P(x) + I_X(x)$, where I_X is the indicator function of X. ∎

The following lemma will be used to prove Theorems 2.3.7 and 2.3.11.

Lemma 2.3.3 *Given $x \in \Re^n$ and $H \succ 0$, let $g = \nabla f(x)$ and $\Delta_d = g^T d + P(x + d) - P(x)$ for all $d \in \Re^n$. Let $\sigma \in (0, 1)$ be given. The following statements hold:*

(a) If $d = d_H(x)$, then

$$-\Delta_d \;\geq\; d^T H d \;\geq\; \lambda_{\min}(H)\|d\|^2.$$

(b) For any $\bar{x} \in \Re^n$, $\alpha \in (0, 1]$, $d = d_H(x)$, and $x' = x + \alpha d$, then

$$(g + Hd)^T(x' - \bar{x}) + P(x') - P(\bar{x}) \leq (\alpha - 1)(d^T H d + \Delta_d).$$

(c) If f satisfies

$$\|\nabla f(y) - \nabla f(z)\| \leq L\|y - z\| \quad \forall y, z \in \Re^n \tag{2.22}$$

for some $L > 0$, then the descent condition

$$F(x + \alpha d) \leq F(x) + \sigma\alpha\Delta_d$$

is satisfied for $d = d_H(x)$, provided $0 \leq \alpha \leq \min\{1, 2(1 - \sigma)\lambda_{\min}(H)/L\}$.

(d) If f satisfies (2.22), then the descent condition

$$F(x+d) \le F(x) + \sigma\Delta_d$$

is satisfied for $d = d_{H(\theta)}(x)$, where $H(\theta) = \theta H$, provided $\theta \ge L/[2(1-\sigma)\lambda_{\min}(H)]$.

Proof. The statements (a)-(c) follow from Theorem 4.1 (a) and Lemma 3.4 of [128] with $J = \{1, \ldots, n\}$, $\gamma = 0$, and $\underline{\lambda} = \lambda_{\min}(H)$. We now prove statement (d). Letting $\alpha = 1$, $d = d_{H(\theta)}(x)$ and using statement (c), we easily see that when $2(1-\sigma)\lambda_{\min}(H(\theta)) \ge 1$, $F(x+d) \le F(x) + \sigma\Delta_d$ is satisfied, which together with the definition of $H(\theta)$ implies statement (d) holds. ∎

We now present the first nonmonotone gradient method for (2.18) as follows.

Nonmonotone gradient method I:

Choose parameters $\eta > 1$, $0 < \sigma < 1$, $0 < \underline{\theta} < \bar{\theta}$, $0 < \underline{\lambda} \le \bar{\lambda}$, and integer $M \ge 0$. Set $k = 0$ and choose $x^0 \in X$.

1) Choose $\theta_k^0 \in [\underline{\theta}, \bar{\theta}]$ and $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$.

2) For $j = 0, 1, \ldots$

2a) Let $\theta_k = \theta_k^0 \eta^j$. Solve (2.20) with $x = x^k$ and $H = \theta_k H_k$ to obtain $d^k = d_H(x)$.

2b) If d^k satisfies

$$F(x^k + d^k) \ \le \ \max_{[k-M]^+ \le i \le k} F(x^i) + \sigma\Delta_k, \tag{2.23}$$

go to step 3), where

$$\Delta_k := \nabla f(x^k)^T d^k + P(x^k + d^k) - P(x^k). \tag{2.24}$$

3) Set $x^{k+1} = x^k + d^k$ and $k \leftarrow k+1$.

end

Remark. The above method is closely related to the one proposed in [134]. They differ from each other only in that the distinct Δ_k's are used inequality (2.23). In particular, the method [134] uses $\Delta_k = -\theta_k \|d^k\|^2/2$. For global convergence, the method [134], however, requires a strong assumption that the gradient of f is *Lipschitz* continuous, which is not

needed for our method (see Theorem 2.3.6). In addition, our method can be viewed as an extension of one projected gradient method (namely, SPG1) studied in [9] for smooth problems, but their method cannot. Finally, local convergence is established for our method (see Theorem 2.3.7), but not studied for the methods in [134] and [9]. ∎

We next prove global convergence of the nonmonotone gradient method I. Before proceeding, we establish two technical lemmas below. The first lemma shows that if $x^k \in X$ is a nonstationary point, there exists an $\theta_k > 0$ in step 2a) so that (2.23) is satisfied, and hence the above method is well defined.

Lemma 2.3.4 *Suppose that $H_k \succ 0$ and $x^k \in X$ is a nonstationary point of problem (2.18). Then, there exists $\tilde{\theta} > 0$ such that $d^k = d_{H_k(\theta_k)}(x^k)$, where $H_k(\theta_k) = \theta_k H_k$, satisfies (2.23) whenever $\theta_k \geq \tilde{\theta}$.*

Proof. For simplicity of notation, let $d(\theta) = d_{H_k(\theta)}(x^k)$, where $H_k(\theta) = \theta H_k$ for any $\theta > 0$. Then, it follows from (2.20) that for all $\theta > 0$,

$$\theta\|d(\theta)\| \leq -\frac{2[\nabla f(x^k)^T d(\theta) + P(x^k + d(\theta)) - P(x^k)]}{\lambda_{\min}(H_k)\|d(\theta)\|} \leq -\frac{2F'(x^k, d(\theta)/\|d(\theta)\|)}{\lambda_{\min}(H_k)}, \quad (2.25)$$

where the second inequality follows from the fact that $P(x^k + d(\theta)) - P(x^k) \geq P'(x^k, d(\theta))$ and $F'(x^k, d(\theta)) = \nabla f(x^k)^T d(\theta) + P'(x^k, d(\theta))$. Thus, we easily see that the set $\tilde{S} := \{\theta\|d(\theta)\| : \theta > 0\}$ is bounded. It implies that $\|d(\theta)\| \to 0$ as $\theta \to \infty$. We claim that

$$\liminf_{\theta\to\infty} \theta\|d(\theta)\| \ > \ 0. \quad (2.26)$$

Suppose not. Then there exists a sequence $\{\bar{\theta}_l\} \uparrow \infty$ such that $\bar{\theta}_l\|d(\bar{\theta}_l)\| \to 0$ as $l \to \infty$. Invoking that $d(\bar{\theta}_l)$ is the optimal solution of (2.20) with $x = x^k$, $H = \bar{\theta}_l H_k$ and $\theta = \bar{\theta}_l$, we have

$$0 \in \nabla f(x^k) + \bar{\theta}_l H_k d(\bar{\theta}_l) + \partial P(x^k + d(\bar{\theta}_l)) + N_X(x^k + d(\bar{\theta}_l)).$$

Upon taking limits on both sides as $l \to \infty$, and using semicontinuity of $\partial P(\cdot)$ and $N_X(\cdot)$ (see Theorem 24.4 of [109] and Lemma 2.42 of [113]), and the relations $\|d(\bar{\theta}_l)\| \to 0$ and $\bar{\theta}_l\|d(\bar{\theta}_l)\| \to 0$, we see that (2.19) holds at x^k, which contradicts the nonstationarity of x^k. Hence, (2.26) holds. We observe that

$$\theta d(\theta)^T H_k d(\theta) \geq \lambda_{\min}(H_k)\theta\|d(\theta)\|^2,$$

which together with (2.26) and $H_k \succ 0$, implies that

$$\|d(\theta)\| = O\left(\theta d(\theta)^T H_k d(\theta)\right) \text{ as } \theta \to \infty. \tag{2.27}$$

This relation together with Lemma 2.3.3(a) implies that as $\theta \to \infty$,

$$\|d(\theta)\| = O\left(\theta d(\theta)^T H_k d(\theta)\right) = O\left(P(x^k) - \nabla f(x^k)^T d(\theta) - P(x^k + d(\theta))\right). \tag{2.28}$$

Using this result and the relation $\|d(\theta)\| \to 0$ as $\theta \to \infty$, we further have

$$\begin{aligned} F(x^k + d(\theta)) - \max_{[k-M]^+ \le i \le k} F(x^i) &\le F(x^k + d(\theta)) - F(x^k) \\ &= f(x^k + d(\theta)) - f(x^k) + P(x^k + d(\theta)) - P(x^k) \\ &= \nabla f(x^k)^T d(\theta) + P(x^k + d(\theta)) - P(x^k) + o(\|d(\theta)\|) \\ &\le \sigma[\nabla f(x^k)^T d(\theta) + P(x^k + d(\theta)) - P(x^k)], \end{aligned} \tag{2.29}$$

provided θ is sufficiently large. It implies that the conclusion holds. ∎

The following lemma shows that the search directions $\{d^k\}$ approach zero, and the sequence of objective function values $\{F(x^k)\}$ also converges.

Lemma 2.3.5 *Suppose that F is bounded below in X and uniformly continuous in the the level set $\mathcal{L} = \{x \in X : F(x) \le F(x^0)\}$. Then, the sequence $\{x^k\}$ generated by the nonmonotone gradient method I satisfies $\lim_{k\to\infty} d^k = 0$. Moreover, the sequence $\{F(x^k)\}$ converges.*

Proof. We first observe that $\{x^k\} \subseteq \mathcal{L}$. Let $l(k)$ be an integer such that $[k - M]^+ \le l(k) \le k$ and

$$F(x^{l(k)}) = \max\{F(x^i) : [k - M]^+ \le i \le k\}$$

for all $k \ge 0$. We clearly observe that $F(x^{k+1}) \le F(x^{l(k)})$ for all $k \ge 0$, which together with the definition of $l(k)$ implies that the sequence $\{F(x^{l(k)})\}$ is monotonically nonincreasing. Further, since F is bounded below in X, we have

$$\lim_{k\to\infty} F(x^{l(k)}) = F^* \tag{2.30}$$

for some $F^* \in \Re$. We next prove by induction that the following limits hold for all $j \ge 1$:

$$\lim_{k\to\infty} d^{l(k)-j} = 0, \qquad \lim_{k\to\infty} F(x^{l(k)-j}) = F^*. \tag{2.31}$$

Using (2.23) and (2.24) with k replaced by $l(k)-1$, we obtain that

$$F(x^{l(k)}) \leq F(x^{l(l(k)-1)}) + \sigma\Delta_{l(k)-1}. \tag{2.32}$$

Replacing k and θ by $l(k)-1$ and $\theta_{l(k)-1}$ in (2.28), respectively, and using $H_{l(k)-1} \succeq \underline{\lambda} I$ and the definition of $\Delta_{l(k)-1}$ (see (2.24)), we have

$$\Delta_{l(k)-1} \leq -\underline{\lambda}\theta_{l(k)-1}\|d^{l(k)-1}\|^2.$$

The above two inequalities yield that

$$F(x^{l(k)}) \leq F(x^{l(l(k)-1)}) - \sigma\underline{\lambda}\theta_{l(k)-1}\|d^{l(k)-1}\|^2, \tag{2.33}$$

which together with (2.30) implies that $\lim_{k\to\infty} \theta_{l(k)-1}\|d^{l(k)-1}\|^2 = 0$. Further, noticing that $\theta_k \geq \underline{\theta}$ for all k, we obtain that $\lim_{k\to\infty} d^{l(k)-1} = 0$. Using this result and (2.30), we have

$$\lim_{k\to\infty} F(x^{l(k)-1}) = \lim_{k\to\infty} F(x^{l(k)} - d^{l(k)-1}) = \lim_{k\to\infty} F(x^{l(k)}) = F^*, \tag{2.34}$$

where the second equality follows from uniform continuity of F in $\mathcal{L}$. Therefore, (2.31) holds for $j = 1$. We now need to show that if (2.31) holds for j, then it also holds for $j+1$. Using a similar argument as that leading to (2.33), we have

$$F(x^{l(k)-j}) \leq F(x^{l(l(k)-j-1)}) - \sigma\underline{\lambda}\theta_{l(k)-j-1}\|d^{l(k)-j-1}\|^2,$$

which together with (2.30), the induction assumption $\lim_{k\to\infty} F(x^{l(k)-j}) = F^*$, and the fact that $\theta_{l(k)-j-1} \geq \underline{\theta}$ for all k, yields $\lim_{k\to\infty} d^{l(k)-j-1} = 0$. Using this result, the induction assumption $\lim_{k\to\infty} F(x^{l(k)-j}) = F^*$, and a similar argument as that leading to (2.34), we can show that $\lim_{k\to\infty} F(x^{l(k)-j-1}) = F^*$. Hence, (2.31) holds for $j+1$.

Finally, we will prove that $\lim_{k\to\infty} d^k = 0$ and $\lim_{k\to\infty} F(x^k) = F^*$. By the definition of $l(k)$, we see that for $k \geq M+1$, $k-M-1 = l(k)-j$ for some $1 \leq j \leq M+1$, which together with the first limit in (2.31), implies that $\lim_{k\to\infty} d^k = \lim_{k\to\infty} d^{k-M-1} = 0$. Additionally, we observe that

$$x^{l(k)} = x^{k-M-1} + \sum_{j=1}^{\bar{l}_k} d^{l(k)-j} \quad \forall k \geq M+1,$$

where $\bar{l}_k = l(k) - (k-M-1) \leq M+1$. Using the above identity, (2.31), and uniform continuity of F in $\mathcal{L}$, we see that $\lim_{k\to\infty} F(x^k) = \lim_{k\to\infty} F(x^{k-M-1}) = F^*$. Thus, the conclusion of this lemma holds. ∎

We are now ready to show that the nonmonotone gradient method I is globally convergent.

Theorem 2.3.6 *Suppose that F is bounded below in X and uniformly continuous in the level set $\mathcal{L} = \{x \in X : F(x) \leq F(x^0)\}$. Then, any accumulation point of the sequence $\{x^k\}$ generated by the nonmonotone gradient method I is a stationary point of (2.18).*

Proof. Suppose for contradiction that x^* is an accumulation point of $\{x^k\}$ that is a nonstationary point of (2.18). Let K be the subsequence such that $\{x^k\}_{k\in K} \to x^*$. We first claim that $\{\theta_k\}_{k\in K}$ is bounded. Suppose not. Then there exists a subsequence of $\{\theta_k\}_{k\in K}$ that goes to ∞. Without loss of generality, we assume that $\{\theta_k\}_{k\in K} \to \infty$. For simplicity of notations, let $\bar{\theta}_k = \theta_k/\eta$, $d^k(\theta) = d_{H_k(\theta)}(x^k)$ for $k \in K$ and $\theta > 0$, where $H_k(\theta) = \theta H_k$. Since $\{\theta_k\}_{k\in K} \to \infty$ and $\theta_k^0 \leq \bar{\theta}$, there exists some index $\bar{k} \geq 0$ such that $\theta_k > \theta_k^0$ for all $k \in K$ with $k \geq \bar{k}$. By the particular choice of θ_k specified in steps (2a) and (2b), we have

$$F(x^k + d^k(\bar{\theta}_k)) > \max_{[k-M]^+\leq i\leq k} F(x^i) + \sigma[\nabla f(x^k)^T d^k(\bar{\theta}_k) + P(x^k + d^k(\bar{\theta}_k)) - P(x^k)], \quad (2.35)$$

Using a similar argument as that leading to (2.25), we have

$$\bar{\theta}_k \|d^k(\bar{\theta}_k)\| \leq -\frac{2F'(x^k, d^k(\bar{\theta}_k)/\|d^k(\bar{\theta}_k)\|)}{\lambda_{\min}(H_k)} \quad \forall k \in K,$$

which along with the relations $H_k \succeq \underline{\lambda} I$ and $\{x^k\}_{k\in K} \to x^*$, implies that $\{\bar{\theta}_k \|d^k(\bar{\theta}_k)\|\}_{k\in K}$ is bounded. Since $\{\bar{\theta}_k\}_{k\in K} \to \infty$, we further have $\{\|d^k(\bar{\theta}_k)\|\}_{k\in K} \to 0$. We now claim that

$$\liminf_{k\in K, k\to\infty} \bar{\theta}_k \|d^k(\bar{\theta}_k)\| \ > \ 0. \quad (2.36)$$

Suppose not. By passing to a subsequence if necessary, we can assume that $\{\bar{\theta}_k \|d^k(\bar{\theta}_k)\|\}_{k\in K} \to 0$. Invoking that $d^k(\bar{\theta}_k)$ is the optimal solution of (2.20) with $x = x^k$ and $H = \bar{\theta}_k H_k$, we have

$$0 \in \nabla f(x^k) + \bar{\theta}_k H_k d^k(\bar{\theta}_k) + \partial P(x^k + d^k(\bar{\theta}_k)) + N_X(x^k + d^k(\bar{\theta}_k)) \quad \forall k \in K.$$

Upon taking limits on both sides as $k \in K \to \infty$, and using semicontinuity of $\partial P(\cdot)$ and $N_X(\cdot)$ (see Theorem 24.4 of [109] and Lemma 2.42 of [113]), the relations $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$, $\{\|d^k(\bar{\theta}_k)\|\}_{k\in K} \to 0$, $\{\bar{\theta}_k \|d^k(\bar{\theta}_k)\|\}_{k\in K} \to 0$ and $\{x^k\}_{k\in K} \to x^*$, we see that (2.19) holds at x^*, which contradicts nonstationarity of x^*. Thus, (2.36) holds. Now, using (2.36), the relation $H_k \succeq \underline{\lambda} I$, and a similar argument as for deriving (2.27), we obtain that $\|d^k(\bar{\theta}_k)\| = O\left(\bar{\theta}_k d^k(\bar{\theta}_k)^T H_k d^k(\bar{\theta}_k)\right)$ as $k \in K \to \infty$. Using this result and a similar argument as the one leading to (2.29), we have

$$F(x^k + d^k(\bar{\theta}_k)) \leq \max_{[k-M]^+\leq i\leq k} F(x^i) + \sigma[\nabla f(x^k)^T d^k(\bar{\theta}_k) + P(x^k + d^k(\bar{\theta}_k)) - P(x^k)],$$

provided that $k \in K$ is sufficiently large. The above inequality evidently contradicts (2.35). Thus, $\{\theta_k\}_{k\in K}$ is bounded.

Finally, invoking that $d^k = d^k(\theta_k)$ is the optimal solution of (2.20) with $x = x^k$, $H = \theta_k H_k$, we have

$$0 \in \nabla f(x^k) + \theta_k H_k d^k + \partial P(x^k + d^k) + N_X(x^k + d^k) \quad \forall k \in K. \tag{2.37}$$

By Lemma 2.3.5, we have $\{d^k\}_{k\in K} \to 0$. Upon taking limits on both sides of (2.37) as $k \in K \to \infty$, and using semicontinuity of $\partial P(\cdot)$ and $N_X(\cdot)$ (see Theorem 24.4 of [109] and Lemma 2.42 of [113]), and the relations $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$, $\{d^k\}_{k\in K} \to 0$ and $\{x^k\}_{k\in K} \to x^*$, we see that (2.19) holds at x^*, which contradicts the nonstationarity of x^* that is assumed at the beginning of this proof. Therefore, the conclusion of this theorem holds. ∎

We next analyze the asymptotic convergence rate of the nonmonotone gradient method I under the following assumption, which is the same as the one made in [128]. In what follows, we denote by $\bar{X}$ the set of stationary points of problem (2.18).

Assumption 2 *(a) $\bar{X} \neq \emptyset$ and, for any $\zeta \geq \min_{x\in X} F(x)$, there exists $\varpi > 0$ and $\epsilon > 0$ such that*

$$\operatorname{dist}(x, \bar{X}) \leq \varpi \|d_I(x)\| \quad \textit{whenever} \quad F(x) \leq \zeta,\ \|d_I(x)\| \leq \epsilon.$$

(b) There exists $\delta > 0$ such that

$$\|x - y\| \geq \delta \quad \textit{whenever} \quad x \in \bar{X}, y \in \bar{X}, F(x) \neq F(y).$$

We are ready to establish local linear rate of convergence for the nonmonotone gradient method I described above. The proof of the following theorem is inspired by the work of Tseng and Yun [128], who analyzed a similar local convergence for a coordinate gradient descent method for a class of nonsmooth minimization problems.

Theorem 2.3.7 *Suppose that Assumption 2 holds, f satisfies (2.22), and F is bounded below in X and uniformly continuous in the level set $\mathcal{L} = \{x \in X : F(x) \leq F(x^0)\}$. Then, the sequence $\{x^k\}$ generated by the nonmonotone gradient method I satisfies*

$$F(x^{l(k)}) - F^* \ \leq \ c(F(x^{l(l(k))-1}) - F^*),$$

provided k is sufficiently large, where $F^* = \lim_{k\to\infty} F(x^k)$ (see Lemma 2.3.5), and c is some constant in $(0, 1)$.

Proof. Invoking $\theta_k^0 \leq \bar{\theta}$ and the specific choice of θ_k, we see from Lemma 2.3.3(d) that $\hat{\theta} := \sup_k \theta_k < \infty$. Let $H_k(\theta) = \theta H_k$. Then, it follows from $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$ and $\theta_k \geq \underline{\theta}$ that $(\underline{\theta} \cdot \underline{\lambda}) I \preceq H_k(\theta_k) \preceq \hat{\theta}\bar{\lambda} I$. Using this relation, Lemma 2.3.2, $H_k \succeq \underline{\lambda} I$, and $d^k = d_{H_k(\theta_k)}(x^k)$, we obtain that

$$\|d_I(x^k)\| = O\left(\|d^k\|\right), \tag{2.38}$$

which together with Lemma 2.3.5 implies $\{d_I(x^k)\} \to 0$. Thus, for any $\epsilon > 0$, there exists some index $\bar{k}$ such that $d_I(x^{l(k)-1}) \leq \epsilon$ for all $k \geq \bar{k}$. In addition, we clearly observe that $F(x^{l(k)-1}) \leq F(x^0)$. Then, by Assumption 2(a) and (2.38), there exists some index k' such that

$$\|x^{l(k)-1} - \bar{x}^{l(k)-1}\| \leq c_1 \|d^{l(k)-1}\| \quad \forall k \geq k' \tag{2.39}$$

for some $c_1 > 0$ and $\bar{x}^{l(k)-1} \in \bar{X}$. Note that

$$\|x^{l(k+1)-1} - x^{l(k)-1}\| \leq \sum_{i=l(k)-1}^{l(k+1)-2} \|d^i\| \leq \sum_{i=[k-M-1]^+}^{[k-1]^+} \|d^i\|,$$

which together with $\{d^k\} \to 0$, implies that $\|x^{l(k+1)-1} - x^{l(k)-1}\| \to 0$. Using this result, (2.39), and Lemma 2.3.5, we obtain

$$\begin{aligned} \|\bar{x}^{l(k+1)-1} - \bar{x}^{l(k)-1}\| &\leq \|x^{l(k+1)-1} - \bar{x}^{l(k+1)-1}\| + \|x^{l(k)-1} - \bar{x}^{l(k)-1}\| \\ &\quad + \|x^{l(k+1)-1} - \bar{x}^{l(k)-1}\| \\ &\leq c_1 \|d^{l(k+1)-1}\| + c_1 \|d^{l(k)-1}\| + \|x^{l(k+1)-1} - \bar{x}^{l(k)-1}\| \ \to \ 0. \end{aligned}$$

It follows from this relation and Assumption 2(b) that there exists an index $\hat{k} \geq k'$ and $v \in \Re$ such that

$$F(\bar{x}^{l(k)-1}) = v \quad \forall k \geq \hat{k}. \tag{2.40}$$

Then, by Lemma 5.1 of [128], we see that

$$F^* = \lim_{k\to\infty} F(x^k) = \liminf_{k\to\infty} F(x^{l(k)-1}) \ \geq \ v. \tag{2.41}$$

Further, using the definition of F, (2.22), (2.40), Lemma 2.3.3(b), and $H_k(\theta_k) \preceq \hat{\theta}\bar{\lambda} I$, we

have for $k \geq \hat{k}$,

$$\begin{aligned} F(x^{l(k)}) - v &= f(x^{l(k)}) + P(x^{l(k)}) - f(\bar{x}^{l(k)-1}) - P(\bar{x}^{l(k)-1}) \\ &= \nabla f(\tilde{x}^k)^T(x^{l(k)} - \bar{x}^{l(k)-1}) + P(x^{l(k)}) - P(\bar{x}^{l(k)-1}) \\ &= (\nabla f(\tilde{x}^k) - \nabla f(x^{l(k)-1})^T(x^{l(k)} - \bar{x}^{l(k)-1}) \qquad (2.42) \\ &\quad -(H_{l(k)-1}(\theta_{l(k)-1})d^{l(k)-1})^T(x^{l(k)} - \bar{x}^{l(k)-1}) \\ &\quad +[(\nabla f(x^{l(k)-1}) + H_{l(k)-1}(\theta_{l(k)-1})d^{l(k)-1})^T(x^{l(k)} - \bar{x}^{l(k)-1}) \qquad (2.43) \\ &\quad +P(x^{l(k)}) - P(\bar{x}^{l(k)-1})] \\ &\leq L\|\tilde{x}^k - x^{l(k)-1}\|\|x^{l(k)} - \bar{x}^{l(k)-1}\| + \hat{\theta}\bar{\lambda}\|d^{l(k)-1}\|\|x^{l(k)} - \bar{x}^{l(k)-1}\|, \qquad (2.44) \end{aligned}$$

where $\tilde{x}^k$ is some point lying on the segment joining $x^{l(k)}$ with $\bar{x}^{l(k)-1}$. It follows from (2.39) that, for $k \geq \hat{k}$,

$$\|\tilde{x}^k - x^{l(k)-1}\| \leq \|x^{l(k)} - x^{l(k)-1}\| + \|x^{l(k)-1} - \bar{x}^{l(k)-1}\| = (1 + c_1)\|d^{l(k)-1}\|.$$

Similarly, $\|x^{l(k)} - \bar{x}^{l(k)-1}\| \leq (1 + c_1)\|d^{l(k)-1}\|$ for $k \geq \hat{k}$. Using these inequalities, Lemma 2.3.3(a), $H_k(\theta_k) \succeq (\underline{\theta} \cdot \underline{\lambda})I$, and (2.44), we see that for $k \geq \hat{k}$,

$$F(x^{l(k)}) - v \;\leq\; -c_2\Delta_{l(k)-1}$$

for some constant $c_2 > 0$. This inequality together with (2.32) gives

$$F(x^{l(k)}) - v \;\leq\; c_3\left(F(x^{l(l(k)-1)}) - F(x^{l(k)})\right) \quad \forall k \geq \hat{k}, \qquad (2.45)$$

where $c_3 = c_2/\sigma$. Using $\lim_{k\to\infty} F(x^{l(k)}) = F^*$, and upon taking limits on both sides of (2.45), we see that $F^* \leq v$, which together with (2.41) implies that $v = F^*$. Using this result and upon rearranging terms of (2.45), we have

$$F(x^{l(k)}) - F^* \;\leq\; c(F(x^{l(l(k))-1}) - F^*) \quad \forall k \geq \hat{k},$$

where $c = c_3/(1 + c_3)$. ∎

We next present the second nonmonotone gradient method for (2.18) as follows.

Nonmonotone gradient method II:

Choose parameters $0 < \eta < 1$, $0 < \sigma < 1$, $0 < \underline{\alpha} < \bar{\alpha}$, $0 < \underline{\lambda} \leq \bar{\lambda}$, and integer $M \geq 0$. Set $k = 0$ and choose $x^0 \in X$.

1) Choose $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$.

2) Solve (2.20) with $x = x^k$ and $H = H_k$ to obtain $d^k = d_H(x)$, and compute Δ_k according to (2.24).

3) Choose $\alpha_k^0 \in [\underline{\alpha}, \bar{\alpha}]$. Find the smallest integer $j \geq 0$ such that $\alpha_k = \alpha_k^0 \eta^j$ satisfies

$$x^k + \alpha_k d^k \in X, \quad F(x^k + \alpha_k d^k) \ \leq \ \max_{[k-M]^+ \leq i \leq k} F(x^i) + \sigma \alpha_k \Delta_k, \tag{2.46}$$

where Δ_k is defined in (2.24).

4) Set $x^{k+1} = x^k + \alpha_k d^k$ and $k \leftarrow k + 1$.

end

Remark. The above method is closely related to the one proposed in [128]. In particular, when the entire coordinate block, that is, $J = \{1, \ldots, n\}$, is chosen for the method [128], it becomes a special case of our method with $M = 0$, which is actually a gradient descent method. Given that our method is generally a nonmonotone method when $M \geq 1$, most proofs of global and local convergence for the method [128] do not hold for our method. In addition, our method can be viewed as an extension of one projected gradient method (namely, SPG2) studied in [9] for smooth problems, but the method [128] generally cannot. ■

We next prove global convergence of the nonmonotone gradient method II. Before proceeding, we establish two technical lemmas below. The first lemma shows that if $x^k \in X$ is a nonstationary point, there exists an $\alpha_k > 0$ in step 3) so that (2.46) is satisfied, and hence the above method is well defined.

Lemma 2.3.8 *Suppose that $H_k \succ 0$ and $x^k \in X$ is a nonstationary point of problem (2.18). Then, there exists $\tilde{\alpha} > 0$ such that $d^k = d_{H_k}(x^k)$ satisfies (2.46) whenever $0 < \alpha_k \leq \tilde{\alpha}$.*

Proof. In view of Lemma 2.1 of [128] with $J = \{1, \ldots, n\}$, $c = 1$, $x = x^k$, and $H = H_k$, we have

$$\begin{aligned} F(x^k + \alpha d^k) &\leq F(x^k) + \alpha \Delta_k + o(\alpha) \\ &\leq \max_{[k-M]^+ \leq i \leq k} F(x^i) + \alpha \Delta_k + o(\alpha) \quad \forall \alpha \in (0, 1], \end{aligned}$$

where Δ_k is defined in (2.24). Using the assumption of this lemma, we see from Lemma 2.3.1 that $d^k \neq 0$, which together with $H_k \succ 0$ and Lemma 2.3.3(a) implies $\Delta_k < 0$. The conclusion of this lemma immediately follows from this relation and the above inequality. ∎

The following lemma shows that the scaled search directions $\{\alpha_k d^k\}$ approach zero, and the sequence of objective function values $\{F(x^k)\}$ also converges.

Lemma 2.3.9 *Suppose that F is bounded below in X and uniformly continuous in the level set $\mathcal{L} = \{x \in X : F(x) \leq F(x^0)\}$. Then, the sequence $\{x^k\}$ generated by the nonmonotone gradient method II satisfies $\lim_{k\to\infty} \alpha_k d^k = 0$. Moreover, the sequence $\{F(x^k)\}$ converges.*

Proof. Let $l(k)$ be defined in the proof of Lemma 2.3.5. We first observe that $\{x^k\} \subseteq \mathcal{L}$. Using (2.24), the definition of d^k, and $H_k \succeq \underline{\lambda} I$, we have

$$\Delta_k = \nabla f(x^k)^T d^k + P(x^k + d^k) - P(x^k) \leq -\frac{1}{2}(d^k)^T H_k d^k \leq -\frac{1}{2}\underline{\lambda}\|d^k\|^2, \tag{2.47}$$

which together with the relation $\alpha_k \leq \alpha_k^0 \leq \bar{\alpha}$, implies that

$$\alpha_k^2 \|d^k\|^2 \leq -2\bar{\alpha}\alpha_k \Delta_k / \underline{\lambda}. \tag{2.48}$$

By a similar argument as that leading to (2.30), we see that $\{x^k\}$ satisfies (2.30) for some F^*. We next show by induction that the following limits hold for all $j \geq 1$:

$$\lim_{k\to\infty} \alpha_{l(k)-j} d^{l(k)-j} = 0, \qquad \lim_{k\to\infty} F(x^{l(k)-j}) = F^*. \tag{2.49}$$

Indeed, using (2.46) with k replaced by $l(k) - 1$, we obtain that

$$F(x^{l(k)}) \leq F(x^{l(l(k)-1)}) + \sigma\alpha_{l(k)-1}\Delta_{l(k)-1}.$$

It together with (2.30) immediately yields $\lim_{k\to\infty} \alpha_{l(k)-1}\Delta_{l(k)-1} = 0$. Using this result and (2.48), we see that the first identity of (2.49) holds for $j = 1$. Further, in view of this identity, (2.30), and uniform continuity of F in $\mathcal{L}$, we can easily see that the second identity of (2.49) also holds $j = 1$. We now need to show that if (2.49) holds for j, then it also holds for $j + 1$. First, it follows from (2.46) that

$$F(x^{l(k)-j}) \leq F(x^{l(l(k)-j-1)}) + \sigma\alpha_{l(k)-j-1}\Delta_{l(k)-j-1},$$

which together with (2.30) and the induction assumption that $\lim_{k\to\infty} F(x^{l(k)-j}) = F^*$, yields $\lim_{k\to\infty} \alpha_{l(k)-j-1}\Delta_{l(k)-j-1} = 0$. Using this result and (2.48), we have

$$\lim_{k\to\infty} \alpha_{l(k)-j-1} d^{l(k)-j-1} = 0.$$

In view of this identity, uniform continuity of F in $\mathcal{L}$ and the induction assumption

$$\lim_{k\to\infty} F(x^{l(k)-j}) = F^*,$$

we can easily show that $\lim_{k\to\infty} F(x^{l(k)-j-1}) = F^*$. Hence, (2.49) holds for $j+1$. The conclusion of this lemma then follows from (2.49) and a similar argument as that in the proof of Lemma 2.3.5. ∎

We are now ready to show that the nonmonotone gradient method II is globally convergent.

Theorem 2.3.10 *Suppose that F is bounded below in X and uniformly continuous in the level set $\mathcal{L} = \{x \in X : F(x) \le F(x^0)\}$. Then, any accumulation point of the sequence $\{x^k\}$ generated by the nonmonotone gradient method II is a stationary point of (2.18).*

Proof. Suppose for contradiction that x^* is an accumulation point of $\{x^k\}$ that is a nonstationary point of (2.18). Let K be the subsequence such that $\{x^k\}_{k\in K} \to x^*$. We first claim that $\liminf_{k\in K, k\to\infty} \|d^k\| > 0$. Suppose not. By passing to a subsequence if necessary, we can assume that $\{\|d^k\|\}_{k\in K} \to 0$. Invoking that d^k is the optimal solution of (2.20) with $x = x^k$ and $H = H_k$, we have

$$0 \in \nabla f(x^k) + H_k d^k + \partial P(x^k + d^k) + N_X(x^k + d^k) \quad \forall k \in K.$$

Upon taking limits on both sides as $k \in K \to \infty$, and using semicontinuity of $\partial P(\cdot)$ and $N_X(\cdot)$ (see Theorem 24.4 of [109] and Lemma 2.42 of [113]) the relations $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$, $\{\|d^k\|\}_{k\in K} \to 0$ and $\{x^k\}_{k\in K} \to x^*$, we see that (2.19) holds at x^*, which contradicts the nonstationarity of x^*. Thus, $\liminf_{k\in K, k\to\infty} \|d^k\| > 0$ holds. Further, using a similar argument as that leading to (2.25), we have

$$\|d^k\| \le -\frac{2F'(x^k, d^k/\|d^k\|)}{\lambda_{\min}(H_k)} \quad \forall k \in K,$$

which together with $\{x^k\}_{k\in K} \to x^*$, $H_k \succeq \underline{\lambda} I$ and $\liminf_{k\in K, k\to\infty} \|d^k\| > 0$, implies that $\{d^k\}_{k\in K}$ is bounded. Further, using (2.47), we see that $\limsup_{k\in K, k\to\infty} \Delta_k < 0$. Now, it follows from Lemma 2.3.9 and the relation $\liminf_{k\in K, k\to\infty} \|d^k\| > 0$ that $\{\alpha_k\}_{k\in K} \to 0$. Since $\alpha_k^0 \ge \underline{\alpha} > 0$, there exists some index $\bar{k} \ge 0$ such that $\alpha_k < \alpha_k^0$ and $\alpha_k < \eta$ for all

$k \in K$ with $k \geq \bar{k}$. Let $\bar{\alpha}_k = \alpha_k/\eta$. Then, $\{\bar{\alpha}_k\}_{k\in K} \to 0$ and $0 < \bar{\alpha}_k \leq 1$ for all $k \in K$. By the stepsize rule used in step (3), we have, for all $k \in K$ with $k \geq \bar{k}$,

$$F(x^k + \bar{\alpha}_k d^k) > \max_{[k-M]^+ \leq i \leq k} F(x^i) + \sigma \bar{\alpha}_k \Delta_k, \tag{2.50}$$

On the other hand, in view of the definition of F, (2.24), the boundedness of $\{d^k\}_{k\in K}$, the relation $\limsup_{k\in K, k\to\infty} \Delta_k < 0$, and the monotonicity of $(P(x^k + \alpha d^k) - P(x^k))/\alpha$, we obtain that, for sufficiently large $k \in K$,

$$\begin{aligned}
F(x^k + \bar{\alpha}_k d^k) &= f(x^k + \bar{\alpha}_k d^k) + P(x^k + \bar{\alpha}_k d^k) \\
&= f(x^k + \bar{\alpha}_k d^k) - f(x^k) + P(x^k + \bar{\alpha}_k d^k) - P(x^k) + F(x^k) \\
&= \bar{\alpha}_k \nabla f(x^k)^T d^k + o(\bar{\alpha}_k \|d^k\|) + P(x^k + \bar{\alpha}_k d^k) - P(x^k) + F(x^k) \\
&\leq \bar{\alpha}_k \nabla f(x^k)^T d^k + o(\bar{\alpha}_k) + \bar{\alpha}_k [P(x^k + d^k) - P(x^k)] + \max_{[k-M]^+ \leq i \leq k} F(x^i) \\
&= \max_{[k-M]^+ \leq i \leq k} F(x^i) + \bar{\alpha}_k \Delta_k + o(\bar{\alpha}_k) \\
&< \max_{[k-M]^+ \leq i \leq k} F(x^i) + \sigma \bar{\alpha}_k \Delta_k,
\end{aligned}$$

which clearly contradicts (2.50). Therefore, the conclusion of this theorem holds. ∎

We next establish local linear rate of convergence for the nonmonotone gradient method II described above. The proof of the following theorem is inspired by the work of Tseng and Yun [128].

Theorem 2.3.11 *Suppose that Assumption 2 holds, $\bar{\alpha} \leq 1$, f satisfies (2.22), and F is bounded below in X and uniformly continuous in the level set $\mathcal{L} = \{x \in X : F(x) \leq F(x^0)\}$. Then, the sequence $\{x^k\}$ generated by the nonmonotone gradient method II satisfies*

$$F(x^{l(k)}) - F^* \ \leq \ c(F(x^{l(l(k))-1}) - F^*)$$

provided k is sufficiently large, where $F^* = \lim_{k\to\infty} F(x^k)$ (see Lemma 2.3.9), and c is some constant in $(0, 1)$.

Proof. Since α_k is chosen by the stepsize rule used in step (3) with $\alpha_k^0 \geq \underline{\alpha} > 0$, we see from Lemma 2.3.3(c) that $\inf_k \alpha_k > 0$. It together with Lemma 2.3.9 implies that $\{d^k\} \to 0$. Further, using Lemma 2.3.2 and the fact that $d^k = d_{H_k}(x^k)$ and $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$, we obtain

that $\|d_I(x^k)\| = \Theta(\|d^k\|)$, and hence $\{d_I(x^k)\} \to 0$. Then, by a similar argument as that in the proof of Theorem 2.3.7, there exist $c_1 > 0$, $v \in \Re$, and $\bar{x}^{l(k)-1} \in \bar{X}$ such that

$$\|x^{l(k)-1} - \bar{x}^{l(k)-1}\| \le c_1 \|d^{l(k)-1}\|, \quad F(\bar{x}^{l(k)-1}) = v \quad \forall k \ge \hat{k},$$

where $\hat{k}$ is some index. Then, by Lemma 5.1 of [128], we see that (2.41) holds for $\{x^k\}$, and the above F^* and v. Further, using the definition of F, (2.22), Lemma 2.3.3(b), and $\underline{\lambda} I \preceq H_k \preceq \bar{\lambda} I$, we have, for $k \ge \hat{k}$,

$$\begin{aligned}
F(x^{l(k)}) - v &= f(x^{l(k)}) + P(x^{l(k)}) - f(\bar{x}^{l(k)-1}) - P(\bar{x}^{l(k)-1}) \\
&= \nabla f(\tilde{x}^k)^T (x^{l(k)} - \bar{x}^{l(k)-1}) + P(x^{l(k)}) - P(\bar{x}^{l(k)-1}) \\
&= (\nabla f(\tilde{x}^k) - \nabla f(x^{l(k)-1})^T (x^{l(k)} - \bar{x}^{l(k)-1}) - (H_{l(k)-1} d^{l(k)-1})^T (x^{l(k)} - \bar{x}^{l(k)-1}) \\
&\quad + \left[(\nabla f(x^{l(k)-1}) + H_{l(k)-1} d^{l(k)-1})^T (x^{l(k)} - \bar{x}^{l(k)-1}) + P(x^{l(k)}) - P(\bar{x}^{l(k)-1}) \right] \\
&\le L \|\tilde{x}^k - x^{l(k)-1}\| \|x^{l(k)} - \bar{x}^{l(k)-1}\| + \bar{\lambda} \|d^{l(k)-1}\| \|x^{l(k)} - \bar{x}^{l(k)-1}\| \\
&\quad + (\alpha_{l(k)-1} - 1) \left[(d^{l(k)-1})^T H_{l(k)-1} d^{l(k)-1} + \Delta_{l(k)-1} \right], \qquad (2.51)
\end{aligned}$$

where $\tilde{x}^k$ is some point lying on the segment joining $x^{l(k)}$ with $\bar{x}^{l(k)-1}$. It follows from (2.39) and $\alpha_k \le 1$ that, for $k \ge \hat{k}$,

$$\|\tilde{x}^k - x^{l(k)-1}\| \le \|x^{l(k)} - x^{l(k)-1}\| + \|x^{l(k)-1} - \bar{x}^{l(k)-1}\| \le (1 + c_1) \|d^{l(k)-1}\|.$$

Similarly, $\|x^{l(k)} - \bar{x}^{l(k)-1}\| \le (1 + c_1)\|d^{l(k)-1}\|$ for $k \ge \hat{k}$. Using these inequalities, Lemma 2.3.3(a), $H_k \succeq \underline{\lambda} I$, $\alpha_k \le 1$, and (2.51), we see that, for $k \ge \hat{k}$,

$$F(x^{l(k)}) - v \ \le \ -c_2 \Delta_{l(k)-1}$$

for some constant $c_2 > 0$. The remaining proof follows similarly as that of Theorem 2.3.7. ■

2.4 Concluding Remark

In this chapter, we developed a novel globally convergent augmented Lagrangian method for solving a class of nonsmooth constrained optimization problems. Additionally, we proposed two nonmonotone gradient methods for solving the augmented Lagrangian subproblems, and established their global and local convergence.

In addition, Burer and Monteiro [13] recently applied the classical augmented Lagrangian method to a nonconvex nonlinear program (NLP) reformulation of semidefinite programs (SDP) via low-rank factorization, and they obtained some nice computational results especially for the SDP relaxations of several hard combinatorial optimization problems. However, the classical augmented Lagrangian method generally cannot guarantee the convergence to a feasible point when applied to a nonconvex NLP. Due to this and [96], their approach [13] at least theoretically may not converge to a feasible point of the primal SDP. Given that the augmented Lagrangian method proposed in this chapter converges globally under some mild assumptions, it would be interesting to apply it to the NLP reformulation of SDP and compare the performance with the approach studied in [13].

Chapter 3

The augmented Lagrangian approach for sparse PCA

In this chapter we propose a new formulation for sparse Principal Component Analysis(PCA) by taking into account the three nice properties of the standard PCA, that is, maximal total explained variance, uncorrelation of principal components, and orthogonality of loading vectors. We also explore the connection of this formulation with the standard PCA and show that it can be viewed as a certain perturbation of the standard PCA. We apply the augmented Lagrangian method proposed in Chpater 2 to solve this new formulation on synthetic [137], Pitprops [74], and gene expression data [38] and compare the results with other exsiting methods.

This chapter is based on the paper [90] co-authored with Zhaosong Lu.

3.1 Introduction to Sparse PCA

Principal component analysis (PCA) is a popular tool for data processing and dimension reduction. It has been widely used in numerous applications in science and engineering such as biology, chemistry, image processing, machine learning and so on. For example, PCA has recently been applied to human face recognition, handwritten zip code classification and gene expression data analysis (see [69, 71, 1, 70]).

In essence, PCA aims at finding a few linear combinations of the original variables, called *principal components* (PCs), which point in orthogonal directions capturing as much

of the variance of the variables as possible. It is well known that PCs can be found via the eigenvalue decomposition of the covariance matrix Σ. However, Σ is typically unknown in practice. Instead, the PCs can be approximately computed via the singular value decomposition (SVD) of the data matrix or the eigenvalue decomposition of the sample covariance matrix. In detail, let $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})$ be a p-dimensional random vector, and X be an $n \times p$ data matrix, which records the n observations of ξ. Without loss of generality, assume X is centered, that is, the column means of X are all 0. Then the commonly used sample covariance matrix is $\hat{\Sigma} = X^T X/(n-1)$. Suppose the eigenvalue decomposition of $\hat{\Sigma}$ is

$$\hat{\Sigma} = V D V^T.$$

Then $\eta = \xi V$ gives the PCs, and the columns of V are the corresponding loading vectors. It is worth noting that V can also be obtained by performing the SVD of X (see, for example, [137]). Clearly, the columns of V are orthonormal vectors, and moreover $V^T \hat{\Sigma} V$ is diagonal. We thus immediately see that if $\hat{\Sigma} = \Sigma$, the corresponding PCs are uncorrelated; otherwise, they can be correlated with each other (see Subsection 3.2 for details). We now describe several important properties of the PCs obtained by the standard PCA when Σ is well estimated by $\hat{\Sigma}$ (see also [137]):

1. The PCs sequentially capture the maximum variance of the variables approximately, thus encouraging minimal information loss as much as possible;

2. The PCs are nearly uncorrelated, so the explained variance by different PCs has small overlap;

3. The PCs point in orthogonal directions, that is, their loading vectors are orthogonal to each other.

In practice, typically the first few PCs are enough to represent the data, thus a great dimensionality reduction is achieved. In spite of the popularity and success of PCA due to these nice features, PCA has an obvious drawback, that is, PCs are usually linear combinations of all p variables and the loadings are typically nonzero. This makes it often difficult to interpret the PCs, especially when p is large. Indeed, in many applications, the original variables have concrete physical meaning. For example in biology, each variable might represent the expression level of a gene. In these cases, the interpretation of PCs would be facilitated if they were composed only from a small number of the original variables, namely, each PC

involved a small number of nonzero loadings. It is thus imperative to develop sparse PCA techniques for finding the PCs with sparse loadings while enjoying the above three nice properties as much as possible.

Sparse PCA has been an active research topic for more than a decade. The first class of approaches are based on ad-hoc methods by post-processing the PCs obtained from the standard PCA mentioned above. For example, Jolliffe [76] applied various rotation techniques to the standard PCs for obtaining sparse loading vectors. Cadima and Jolliffe [14] proposed a simple thresholding approach by artificially setting to zero the standard PCs' loadings with absolute values smaller than a threshold. In recent years, optimization approaches have been proposed for finding sparse PCs. They usually formulate sparse PCA into an optimization problem, aiming at achieving the sparsity of loadings while maximizing the explained variance as much as possible. For instance, Jolliffe et al. [78] proposed an interesting algorithm, called SCoTLASS, for finding sparse orthogonal loading vectors by sequentially maximizing the approximate variance explained by each PC under the l_1-norm penalty on loading vectors. Zou et al. [137] formulated sparse PCA as a regression-type optimization problem and imposed a combination of l_1- and l_2-norm penalties on the regression coefficients. d'Aspremont et al. [42] proposed a method, called DSPCA, for finding sparse PCs by solving a sequence of semidefinite program relaxations of sparse PCA. Shen and Huang [117] recently developed an approach for computing sparse PCs by solving a sequence of rank-one matrix approximation problems under several sparsity-inducing penalties. Very recently, Journée et al. [77] formulated sparse PCA as nonconcave maximization problems with l_0- or l_1-norm sparsity-inducing penalties. They showed that these problems can be reduced into maximization of a convex function on a compact set, and they also proposed a simple but computationally efficient gradient method for finding a stationary point of the latter problems. Additionally, greedy methods were investigated for sparse PCA by Moghaddam et al. [95] and d'Aspremont et al. [40].

The PCs obtained by the above methods [76, 14, 78, 137, 42, 117, 95, 40, 77] are usually sparse. However, the aforementioned nice properties of the standard PCs are lost to some extent in these sparse PCs. Indeed, the likely correlation among the sparse PCs are not considered in these methods. Therefore, their sparse PCs can be quite correlated with each other. Also, the total explained variance that these methods attempt to maximize can be too optimistic as there may be some overlap among the individual variances of sparse PCs. Finally, the loading vectors of the sparse PCs given by these methods lack orthogonality

except SCoTLASS [78].

In this chapter we propose a new formulation for sparse PCA by taking into account the three nice properties of the standard PCA, that is, maximal total explained variance, uncorrelation of PCs, and orthogonality of loading vectors. We also explore the connection of this formulation with the standard PCA and show that it can be viewed as a certain perturbation of the standard PCA. Then the new formulation of sparse PCA is solved by the augmented Lagrangian method proposed in Chapter 2. We then compare the proposed sparse PCA approach with several existing methods on synthetic [137], Pitprops [74], and gene expression data [38].

3.2 Formulation for sparse PCA

In this subsection we propose a new formulation for sparse PCA by taking into account sparsity and orthogonality of loading vectors, and uncorrelation of PCs. We also address the connection of our formulation with the standard PCA.

Let $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})$ be a p-dimensional random vector with covariance matrix Σ. Suppose X is an $n \times p$ data matrix, which records the n observations of ξ. Without loss of generality, assume the column means of X are 0. Then the commonly used sample covariance matrix of ξ is $\hat{\Sigma} = X^T X/(n-1)$. For any r loading vectors represented as $V = [V_1, \ldots, V_r] \in \Re^{p \times r}$ where $1 \leq r \leq p$, the corresponding components are given by $\eta = (\eta^{(1)}, \ldots, \eta^{(r)}) = \xi V$, which are linear combinations of $\xi^{(1)}, \ldots, \xi^{(p)}$. Clearly, the covariance matrix of η is $V^T \Sigma V$, and thus the components $\eta^{(i)}$ and $\eta^{(j)}$ are uncorrelated if and only if the ijth entry of $V^T \Sigma V$ is zero. Also, the total explained variance by the components $\eta^{(i)}$'s equals, if they are uncorrelated, the sum of the individual variances of $\eta^{(i)}$'s, that is,

$$\sum_{i=1}^{r} V_i^T \Sigma V_i = \mathrm{Tr}(V^T \Sigma V).$$

Recall that our aim is to find a set of sparse and orthogonal loading vectors V so that the corresponding components $\eta^{(1)}, \ldots, \eta^{(r)}$ are uncorrelated and explain as much variance of the original variables $\xi^{(1)}, \ldots, \xi^{(p)}$ as possible. It appears that our goal can be achieved by

solving the following problem:

$$\begin{array}{ll} \max\limits_{V\in\Re^{n\times r}} & \mathrm{Tr}(V^T\Sigma V) - \rho \bullet |V| \\ \text{s.t.} & V^T\Sigma V \text{ is diagonal}, \\ & V^TV = I, \end{array} \tag{3.1}$$

where $\rho \in \Re^{p\times r}_+$ is a tunning parameter for controlling the sparsity of V. However, the covariance matrix Σ is typically unknown and can only be approximated by the sample covariance matrix $\hat{\Sigma}$. It looks plausible to modify (3.1) by simply replacing Σ with $\hat{\Sigma}$ at a glance. Nevertheless, such a modification would eliminate all optimal solutions V^* of (3.1) from consideration since $(V^*)^T\hat{\Sigma}V^*$ is generally non-diagonal. For this reason, given a sample covariance $\hat{\Sigma}$, we consider the following formulation for sparse PCA, which can be viewed as a modification of problem (3.1),

$$\begin{array}{ll} \max\limits_{V\in\Re^{n\times r}} & \mathrm{Tr}(V^T\hat{\Sigma}V) - \rho \bullet |V| \\ \text{s.t.} & |V_i^T\hat{\Sigma}V_j| \le \Delta_{ij} \quad \forall i \ne j, \\ & V^TV = I, \end{array} \tag{3.2}$$

where $\Delta_{ij} \ge 0$ $(i \ne j)$ are the parameters for controlling the correlation of the components corresponding to V. Clearly, $\Delta_{ij} = \Delta_{ji}$ for all $i \ne j$.

We next explore the connection of formulation (3.2) with the standard PCA. Before proceeding, we state a technical lemma as follows that will be used subsequently. Its proof can be found in [101].

Lemma 3.2.1 *Given any $\hat{\Sigma} \in \mathcal{S}^n$ and integer $1 \le r \le n$, define*

$$\underline{i}_r = \max\{1 \le i \le n : \lambda_i(\hat{\Sigma}) > \lambda_r(\hat{\Sigma})\}, \qquad \bar{i}_r = \max\{1 \le i \le n : \lambda_i(\hat{\Sigma}) = \lambda_r(\hat{\Sigma})\}, \tag{3.3}$$

and let f^* be the optimal value of

$$\max\{\mathrm{Tr}(\hat{\Sigma}Y) : \ 0 \preceq Y \preceq I, \ \mathrm{Tr}(Y) = r\}. \tag{3.4}$$

Then, $f^* = \sum_{i=1}^r \lambda_i(\hat{\Sigma})$, and Y^* is an optimal solution of (3.4) if and only if $Y^* = U_1^*U_1^{*T} + U_2^*P^*U_2^{*T}$, where $P^* \in \mathcal{S}^{\bar{i}_r - \underline{i}_r}$ satisfies $0 \preceq P^* \preceq I$ and $\mathrm{Tr}(P^*) = r - \underline{i}_r$, and $U_1^* \in \Re^{n\times \underline{i}_r}$ and $U_2^* \in \Re^{n\times(\bar{i}_r - \underline{i}_r)}$ are the matrices whose columns consist of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to the eigenvalues $(\lambda_1(\hat{\Sigma}), \ldots, \lambda_{\underline{i}_r}(\hat{\Sigma}))$ and $(\lambda_{\underline{i}_r+1}(\hat{\Sigma}), \ldots, \lambda_{\bar{i}_r}(\hat{\Sigma}))$, respectively.

We next address the relation between the eigenvectors of $\hat{\Sigma}$ and the solutions of problem (3.2) when $\rho = 0$ and $\Delta_{ij} = 0$ for all $i \neq j$.

Theorem 3.2.2 *Suppose for problem (3.2) that $\rho = 0$ and $\Delta_{ij} = 0$ for all $i \neq j$. Let f^* be the optimal value of (3.2). Then, $f^* = \sum_{i=1}^{r} \lambda_i(\hat{\Sigma})$, and $V^* \in \Re^{n\times r}$ is an optimal solution of (3.2) if and only if the columns of V^* consist of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to r largest eigenvalues of $\hat{\Sigma}$.*

Proof. We first show that $f^* = \sum_{i=1}^{r} \lambda_i(\hat{\Sigma})$. Indeed, let U be an $n \times r$ matrix whose columns consist of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to r largest eigenvalues of $\hat{\Sigma}$. We then see that U is a feasible solution of (3.2) and $\mathrm{Tr}(U^T\hat{\Sigma}U) = \sum_{i=1}^{r} \lambda_i(\hat{\Sigma})$. It follows that $f^* \geq \sum_{i=1}^{r} \lambda_i(\hat{\Sigma})$. On the other hand, we observe that f^* is bounded above by the optimal value of

$$\max\{\mathrm{Tr}(V^T\hat{\Sigma}V) : \ V^TV = I, \ V \in \Re^{n\times r}\}.$$

We know from [58] that its optimal value equals $\sum_{i=1}^{r} \lambda_i(\hat{\Sigma})$. Therefore, $f^* = \sum_{i=1}^{r} \lambda_i(\hat{\Sigma})$ holds and U is an optimal solution of (3.2). It also implies that the "if" part of this proposition holds. We next show that the "only if" part also holds. Let $V^* \in \Re^{n\times r}$ be an optimal solution of (3.2), and define $Y^* = V^*V^{*T}$. Then, we have $V^{*T}V^* = I$, which yields $0 \preceq Y^* \preceq I$ and $\mathrm{Tr}(Y^*) = r$. Hence, Y^* is a feasible solution of (3.4). Using the fact that $f^* = \sum_{i=1}^{r} \lambda_i(\hat{\Sigma})$, we then have

$$\mathrm{Tr}(\hat{\Sigma}Y^*) = \mathrm{Tr}(V^{*T}\hat{\Sigma}V^*) = \sum_{i=1}^{r} \lambda_i(\hat{\Sigma}),$$

which together with Lemma 3.2.1 implies that Y^* is an optimal solution of (3.4). Let $\underline{i}_r$ and $\bar{i}_r$ be defined in (3.3). Then, it follows from Lemma 3.2.1 that $Y^* = U_1^*U_1^{*T} + U_2^*P^*U_2^{*T}$, where $P^* \in \mathcal{S}^{\bar{i}_r - \underline{i}_r}$ satisfies $0 \preceq P^* \preceq I$ and $\mathrm{Tr}(P^*) = r - \underline{i}_r$, and $U_1^* \in \Re^{n\times \underline{i}_r}$ and $U_2^* \in \Re^{n\times(\bar{i}_r - \underline{i}_r)}$ are the matrices whose columns consist of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to the eigenvalues $(\lambda_1(\hat{\Sigma}), \ldots, \lambda_{\underline{i}_r}(\hat{\Sigma}))$ and $(\lambda_{\underline{i}_r+1}(\hat{\Sigma}), \ldots, \lambda_{\bar{i}_r}(\hat{\Sigma}))$, respectively. Thus, we have

$$\hat{\Sigma}U_1^* = U_1^*\Lambda, \quad \hat{\Sigma}U_2^* = \lambda_r(\hat{\Sigma})U_2^*, \tag{3.5}$$

where $\Lambda = \mathscr{D}(\lambda_1(\hat{\Sigma}), \ldots, \lambda_{\underline{i}_r}(\hat{\Sigma}))$. In addition, it is easy to show that $\mathrm{rank}(Y^*) = \underline{i}_r + \mathrm{rank}(P^*)$. Since $Y^* = V^*V^{*T}$ and $V^{*T}V^* = I$, we can observe that $\mathrm{rank}(Y^*) = r$. Hence, $\mathrm{rank}(P^*) = r - \underline{i}_r$, which implies that P^* has only $r - \underline{i}_r$ nonzero eigenvalues. Using this

fact and the relations $0 \preceq P^* \preceq I$ and $\mathrm{Tr}(P^*) = r - \underline{i}_r$, we can further conclude that $r - \underline{i}_r$ eigenvalues of P^* are 1 and the rest are 0. Therefore, there exists $W \in \Re^{(\bar{i}_r - \underline{i}_r)\times(r-\underline{i}_r)}$ such that

$$W^T W = I, \quad P^* = WW^T. \tag{3.6}$$

It together with $Y^* = U_1^* U_1^{*T} + U_2^* P^* U_2^{*T}$ implies that $Y^* = U^* U^{*T}$, where $U^* = [U_1^* \;\; U_2^* W]$. In view of (3.6) and the identities $U_1^{*T} U_1^* = I$, $U_2^{*T} U_2^* = I$ and $U_1^{*T} U_2^* = 0$, we see that $U^{*T} U^* = I$. Using this result, and the relations $V^{*T} V^* = I$ and $Y^* = U^* U^{*T} = V^* V^{*T}$, it is not hard to see that the columns of U^* and V^* form an orthonormal basis for the range space of Y^*, respectively. Thus, $V^* = U^* Q$ for some $Q \in \Re^{r\times r}$ satisfying $Q^T Q = I$. Now, let $D = V^{*T} \hat{\Sigma} V^*$. By the definition of V^*, we know that D is an $r \times r$ diagonal matrix. Moreover, in view of (3.5), (3.6), the definition of U^*, and the relations $V^* = U^* Q$, $U_1^{*T} U_1^* = I$, $U_2^{*T} U_2^* = I$ and $U_1^{*T} U_2^* = 0$, we have

$$\begin{aligned} D &= V^{*T} \hat{\Sigma} V^* = Q^T U^{*T} \hat{\Sigma} U^* Q = Q^T \begin{bmatrix} U_1^{*T} \\ W^T U_2^{*T} \end{bmatrix} \hat{\Sigma} \, [U_1^* \;\; U_2^* W] \, Q \\ &= Q^T \begin{bmatrix} \Lambda & 0 \\ 0 & \lambda_r(\hat{\Sigma}) I \end{bmatrix} Q, \end{aligned} \tag{3.7}$$

which together with $Q^T Q = I$ implies that D is similar to the diagonal matrix appearing on the right-hand side of (3.7). Hence, the diagonal elements of D consist of r largest eigenvalues of $\hat{\Sigma}$. In addition, let $Q_1 \in \Re^{\underline{i}_r \times r}$ and $Q_2 \in \Re^{(r-\underline{i}_r)\times r}$ be the submatrices corresponding to the first $\underline{i}_r$ and the last $r - \underline{i}_r$ rows of Q, respectively. Then, in view of the definition of U^* and $V^* = U^* Q$, we have

$$[U_1^* \;\; U_2^* W] = U^* = V^* Q^T = [V^* Q_1^T \;\; V^* Q_2^T].$$

Thus, we obtain that $U_1^* = V^* Q_1^T$ and $U_2^* W = V^* Q_2^T$. Using these identities, (3.5), (3.7), and the relation $V^* = U^* Q$, we have

$$\begin{aligned} \hat{\Sigma} V^* &= \hat{\Sigma} U^* Q = \hat{\Sigma} \, [U_1^* \;\; U_2^* W] \, Q = [U_1^* \Lambda \;\; \lambda_r(\hat{\Sigma}) U_2^* W] Q \\ &= [V^* Q_1^T \Lambda \;\; \lambda_r(\hat{\Sigma}) V^* Q_2^T] Q = V^* Q^T \begin{bmatrix} \Lambda & 0 \\ 0 & \lambda_r(\hat{\Sigma}) I \end{bmatrix} Q = V^* D. \end{aligned}$$

It follows that the columns of V^* consist of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to r largest eigenvalues of $\hat{\Sigma}$, and thus the "only if" part of this proposition holds. ∎

From the above theorem, we see that when $\rho = 0$ and $\Delta_{ij} = 0$ for all $i \neq j$, each solution of (3.2) consists of the orthonormal eigenvectors of $\hat{\Sigma}$ corresponding to r largest eigenvalues of $\hat{\Sigma}$, which can be computed from the eigenvalue decomposition of $\hat{\Sigma}$. Therefore, the loading vectors obtained from (3.2) are the same as those given by the standard PCA when applied to $\hat{\Sigma}$. On the other hand, when ρ and Δ_{ij} for all $i \neq j$ are small, the loading vectors found by (3.2) can be viewed as an approximation to the ones provided by the standard PCA.

3.3 Augmented Lagrangian method for sparse PCA

In this subsection we discuss the applicability and implementation details of the augmented Lagrangian method proposed in Chapter 2 for solving sparse PCA (3.2).

3.3.1 Applicability of augmented Lagrangian method for (3.2)

We first observe that problem (3.2) can be reformulated as

$$\begin{array}{ll} \min\limits_{V \in \Re^{n\times r}} & -\mathrm{Tr}(V^T\hat{\Sigma}V) + \rho \bullet |V| \\ \text{s.t.} & V_i^T\hat{\Sigma}V_j \leq \Delta_{ij} \quad \forall i \neq j, \\ & -V_i^T\hat{\Sigma}V_j \leq \Delta_{ij} \quad \forall i \neq j, \\ & V^TV = I. \end{array} \tag{3.8}$$

Clearly, problem (3.8) has the same form as (2.1). From Subsection 2.2, we know that the sufficient conditions for convergence of our augmented Lagrangian method include: i) a feasible point is explicitly given; and ii) Robinson's condition (2.2) holds at an accumulation point. It is easy to observe that any $V \in \Re^{n\times r}$ consisting of r orthonormal eigenvectors of $\hat{\Sigma}$ is a feasible point of (3.8), and thus the first condition is trivially satisfied. Given that the accumulation points are not known beforehand, it is hard to check the second condition directly. Instead, we may check Robinson's condition at all feasible points of (3.8). However, due to complication of the constraints, we are only able to verify Robinson's condition at a set of feasible points below. Before proceeding, we establish a technical lemma as follows that will be used subsequently.

Lemma 3.3.1 *Let $V \in \Re^{n\times r}$ be a feasible solution of (3.8). Given any W_1, $W_2 \in \mathcal{S}^r$, the system of*

$$\begin{aligned} \delta V^T\hat{\Sigma}V + V^T\hat{\Sigma}\,\delta V + \delta D &= W_1, &(3.9)\\ \delta V^T V + V^T\delta V &= W_2 &(3.10)\end{aligned}$$

has at least one solution $(\delta V, \delta D) \in \Re^{n\times r} \times \mathcal{D}^r$ if one of the following conditions holds:

a) $V^T\hat{\Sigma}V$ is diagonal and $V_i^T\hat{\Sigma}V_i \neq V_j^T\hat{\Sigma}V_j$ for all $i \neq j$;

b) $V^T\hat{\Sigma}(I - VV^T)\hat{\Sigma}V$ is nonsingular.

Proof. Note that the columns of V consist of r orthonormal eigenvectors. Therefore, there exist $\bar{V} \in \Re^{n\times(n-r)}$ such that $[V\ \bar{V}] \in \Re^{n\times n}$ is an orthogonal matrix. It follows that for any $\delta V \in \Re^{n\times r}$, there exists $\delta P \in \Re^{r\times r}$ and $\delta\bar{P} \in \Re^{(n-r)\times r}$ such that $\delta V = V\delta P + \bar{V}\delta\bar{P}$. Performing such a change of variable for δV, and using the fact that the matrix $[V\ \bar{V}]$ is orthogonal, we can show that the system of (3.9) and (3.10) is equivalent to

$$\begin{aligned} \delta P^T G + G\delta P + \delta\bar{P}^T\bar{G} + \bar{G}^T\delta\bar{P} + \delta D &= W_1, &(3.11)\\ \delta P^T + \delta P &= W_2, &(3.12)\end{aligned}$$

where $G = V^T\hat{\Sigma}V$ and $\bar{G} = \bar{V}^T\hat{\Sigma}V$. The remaining proof of this lemma reduces to show that the system of (3.11) and (3.12) has at least a solution $(\delta P, \delta\bar{P}, \delta D) \in \Re^{r\times r} \times \Re^{(n-r)\times r} \times \mathcal{D}^r$ if one of conditions (a) or (b) holds.

First, we assume that condition (a) holds. Then, G is a diagonal matrix and $G_{ii} \neq G_{jj}$ for all $i \neq j$. It follows that there exists a unique $\delta P^* \in \Re^{n\times r}$ satisfying $\delta P_{ii} = (W_2)_{ii}/2$ for all i and

$$\begin{aligned} \delta P_{ij}G_{jj} + G_{ii}\delta P_{ij} &= (W_1)_{ij} \quad \forall i \neq j,\\ \delta P_{ij} + \delta P_{ji} &= (W_2)_{ij} \quad \forall i \neq j.\end{aligned}$$

Now, let $\delta\bar{P}^* = 0$ and $\delta D^* = \widetilde{\mathscr{D}}(W_1 - GW_2)$. It is easy to verify that $(\delta P^*, \delta\bar{P}^*, \delta D^*)$ is a solution of the system of (3.11) and (3.12).

We next assume that condition (b) holds. Given any $\delta\bar{P} \in \Re^{(n-r)\times r}$, there exist $\delta Y \in \Re^{(n-r)\times r}$ and $\delta Z \in \Re^{r\times r}$ such that $\bar{G}^T\delta Y = 0$ and $\delta\bar{P} = \delta Y + \bar{G}\delta Z$. Peforming such a change of variable for $\delta\bar{P}$, we see that (3.11) can be rewritten as

$$\delta P^T G + G\delta P + \delta Z^T\bar{G}^T\bar{G} + \bar{G}^T\bar{G}\delta Z + \delta D = W_1. \tag{3.13}$$

Thus, it suffices to show that the system of (3.12) and (3.13) has at least a solution $(\delta P, \delta Z, \delta D) \in \Re^{r\times r} \times \Re^{r\times r} \times \mathcal{D}^r$. Using the definition of $\bar{G}$ and the fact that the matrix $[V\ \bar{V}]$ is orthogonal, we see that

$$\bar{G}^T\bar{G} = V^T\hat{\Sigma}\bar{V}\bar{V}^T\hat{\Sigma}V = V^T\hat{\Sigma}(I - VV^T)\hat{\Sigma}V,$$

which together with condition (b) implies that $\bar{G}^T\bar{G}$ is nonsingular. Now, let

$$\delta P^* = W_2/2, \quad \delta Z^* = (\bar{G}^T\bar{G})^{-1}(2W_1 - W_2G - GW_2)/4, \quad \delta D^* = 0.$$

It is easy to verify that $(\delta P^*, \delta Z^*, \delta D^*)$ is a solution of the system of (3.13) and (3.12). Therefore, the conclusion holds. ∎

We are now ready to show that Robinson's condition (2.2) holds at a set of feasible points of (3.8).

Proposition 3.3.2 *Let $V \in \Re^{n\times r}$ be a feasible solution of (3.8). The Robinson's condition (2.2) holds at V if one of the following conditions hold:*

a) $\Delta_{ij} = 0$ and $V_i^T\hat{\Sigma}V_i \neq V_j^T\hat{\Sigma}V_j$ for all $i \neq j$;

b) There is at least one active and one inactive inequality constraint of (3.8) at V, and $V^T\hat{\Sigma}(I - VV^T)\hat{\Sigma}V$ is nonsingular;

c) All inequality constraints of (3.8) are inactive at V.

Proof. We first suppose that condition (a) holds. Then, it immediately implies that $V^T\hat{\Sigma}V$ is diagonal, and hence the condition (a) of Lemma 3.3.1 holds. In addition, we observe that all constraints of (3.8) become equality ones. Using these facts and Lemma 3.3.1, we see that Robinson's condition (2.2) holds at V. Next, we assume that condition (b) holds. It implies that condition (b) of Lemma 3.3.1 holds. The conclusion then follows directly from Lemma 3.3.1. Finally, suppose condition (c) holds. Then, Robinson's condition (2.2) holds at V if and only if (3.10) has at least a solution $\delta V \in \Re^{n\times r}$ for any $W_2 \in \mathcal{S}^r$. Noting that $V^TV = I$, we easily see that $\delta V = VW_2/2$ is a solution of (3.10), and thus Robinson's condition (2.2) holds at V. ∎

From Proposition 3.3.2, we see that Robinson's condition (2.2) indeed holds at a set of feasible points of (3.8). Though we are not able to show that it holds at all feasible points of (3.8), we observe in our implementation that the accumulation points of our augmented Lagrangian method generally satisfy one of the conditions described in Proposition 3.3.2, and so Robinson's condition usually holds at the accumulation points. Moreover, we have never seen that our augmented Lagrangian method failed to converge for an instance in our implementation so far.

3.3.2 Implementation details of augmented Lagrangian method for (3.8)

In this subsection, we show how our augmented Lagrangian method proposed in Subsection 2.2 can be applied to solve problem (3.8) (or, equivalently, (3.2)). In particular, we will discuss the implementation details of outer and inner iterations of this method.

We first discuss how to efficiently evaluate the function and gradient involved in our augmented Lagrangian method for problem (3.8). Suppose that $\varrho > 0$ is a penalty parameter, and $\{\lambda_{ij}^+\}_{i\neq j}$ and $\{\lambda_{ij}^-\}_{i\neq j}$ are the Lagrangian multipliers for the inequality constraints of (3.8), respectively, and $\mu \in \mathcal{S}^r$ is the Lagrangian multipliers for the equality constraints of (3.8). For convenience of presentation, let $\Delta \in \mathcal{S}^r$ be the matrix whose ijth entry equals the parameter Δ_{ij} of (3.8) for all $i \neq j$ and diagonal entries are 0. Similarly, let λ^+ (resp., λ^-) be an $r \times r$ symmetric matrix whose ijth entry is λ_{ij}^+ (resp., λ_{ij}^-) for all $i \neq j$ and diagonal entries are 0. We now define $\lambda \in \Re^{2r\times r}$ by stacking λ^+ over λ^-. Using these notations, we observe that the associated Lagrangian function for problem (3.8) can be rewritten as

$$L_\varrho(V, \lambda, \mu) = w(V) + \rho \bullet |V|, \tag{3.14}$$

where

$$\begin{aligned} w(V) &= -\mathrm{Tr}(V^T\hat{\Sigma}V) + \tfrac{1}{2\varrho}\left(\left\|\left[\begin{pmatrix}\lambda^+\\ \lambda^-\end{pmatrix} + \varrho\begin{pmatrix} S-\Delta\\ -S-\Delta\end{pmatrix}\right]^+\right\|_F^2 - \left\|\begin{pmatrix}\lambda^+\\ \lambda^-\end{pmatrix}\right\|_F^2\right) \\ &\quad +\mu \bullet R + \tfrac{\varrho}{2}\|R\|_F^2, \end{aligned}$$

and

$$S = V^T\hat{\Sigma}V - \widetilde{\mathscr{D}}(V^T\hat{\Sigma}V), \qquad R = V^TV - I. \tag{3.15}$$

It is not hard to verify that the gradient of $w(V)$ can be computed according to

$$\nabla w(V) = 2\left(-\hat{\Sigma}V\left(I - [\lambda^+ + \varrho S - \varrho\Delta]^+ + [\lambda^- - \varrho S - \varrho\Delta]^+\right) + V(\mu + \varrho R)\right).$$

Clearly, the main effort for the above function and gradient evaluations lies in computing $V^T\hat{\Sigma}V$ and $\hat{\Sigma}V$. When $\hat{\Sigma} \in \mathcal{S}^p$ is explicitly given, the computational complexity for evaluating these two quantities is $O(p^2r)$. In practice, we are, however, typically given the data matrix $X \in \Re^{n\times p}$. Assuming the column means of X are 0, the sample covariance matrix $\hat{\Sigma}$ can be obtained from $\hat{\Sigma} = X^TX/(n-1)$. Nevertheless, when $p \gg n$, we observe that it is not efficient to compute and store $\hat{\Sigma}$. Also, it is much cheaper to compute $V^T\hat{\Sigma}V$ and $\hat{\Sigma}V$ by using $\hat{\Sigma}$ implicitly rather than explicitly. Indeed, we can first evaluate XV, and then compute $V^T\hat{\Sigma}V$ and $\hat{\Sigma}V$ according to

$$V^T\hat{\Sigma}V = (XV)^T(XV)/(n-1), \qquad \hat{\Sigma}V = X^T(XV)/(n-1).$$

Then, the resulting overall computational complexity is $O(npr)$, which is clearly much superior to the one by using $\hat{\Sigma}$ explicitly, that is, $O(p^2r)$.

We now address initialization and termination criterion for our augmented Lagrangian method. In particular, we choose initial point $V^0_{\rm init}$ and feasible point $V^{\rm feas}$ to be the loading vectors of the r standard PCs, that is, the orthonormal eigenvectors corresponding to r largest eigenvalues of $\hat{\Sigma}$. In addition, we set initial penalty parameter and Lagrangian multipliers to be 1, and set the parameters $\tau = 0.2$ and $\sigma = 10$. We terminate our method once the constraint violation and the relative difference between the augmented Lagrangian function and the regular objective function are sufficiently small, that is,

$$\max_{i\neq j}[|V_i^T\hat{\Sigma}V_j| - \Delta_{ij}]^+ \leq \epsilon_I, \quad \max_{i,j}|R_{ij}| \leq \epsilon_E, \quad \frac{|L_\varrho(V,\lambda,\mu) - f(V)|}{\max(|f(V)|,1)} \leq \epsilon_O, \tag{3.16}$$

where $f(V) = -\mathrm{Tr}(V^T\hat{\Sigma}V) + \rho \bullet |V|$, R is defined in (3.15), and ϵ_I, ϵ_E, ϵ_O are some prescribed accuracy parameters corresponding to inequality constraints, equality constraints and objective function, respectively.

We next discuss how to apply the nonmonotone gradient methods proposed in Subsection 2.3 for the augmented Lagrangian subproblems, which are in the form of

$$\min_V L_\varrho(V,\lambda,\mu), \tag{3.17}$$

where the function $L_\varrho(\cdot,\lambda,\mu)$ is defined in (3.14). Given that the implementation details of those nonmonotone gradient methods are similar, we only focus on the first one, that is, the nonmonotone gradient method I. First, the initial point for this method can be chosen according to the scheme described at the end of Subsection 2.2. In addition, given

the kth iterate V^k, we choose $H_k = \beta_k^{-1} I$ according to the scheme proposed by Barzilai and Borwein [5], which was also used by Birgin et al. [9] for studying a class of projected gradient methods. Indeed, let $0 < \beta_{\min} < \beta_{\max}$ be given. Initially, choose an arbitrary $\beta_0 \in [\beta_{\min}, \beta_{\max}]$. Then, β_k is updated as follows:

$$\beta_{k+1} = \begin{cases} \beta_{\max}, & \text{if } b_k \leq 0; \\ \max\{\beta_{\min}, \min\{\beta_{\max}, a_k/b_k\}\}, & \text{otherwise,} \end{cases}$$

where $a_k = \|V^k - V^{k-1}\|_F^2$ and $b_k = (V^k - V^{k-1}) \bullet (\nabla w(V^k) - \nabla w(V^{k-1}))$. The search direction d^k is then computed by solving subproblem (2.20) with $H = \theta_k H_k$ for some $\theta_k > 0$, which in the context of (2.13) and (3.14) becomes

$$d^k := \arg\min_d \left\{ \nabla w(V^k) \bullet d + \frac{1}{2\theta_k \beta_k} \|d\|_F^2 + \rho \bullet |V^k + d| \right\}. \tag{3.18}$$

It is not hard to verify that the optimal solution of problem (3.18) has a closed-form expression, which is given by

$$d^k = \operatorname{sign}(C) \odot [|C| - \theta_k \beta_k \rho]^+ - V^k,$$

where $C = V^k - \theta_k \beta_k \nabla w(V^k)$. In addition, we see from Lemma 2.3.1 that the following termination criterion is suitable for this method when applied to (3.17):

$$\frac{\max_{ij} |d_I(V^k)|_{ij}}{\max(|L_\varrho(V^k, \lambda, \mu)|, 1)} \leq \epsilon,$$

where $d_I(V^k)$ is the solution of (3.18) with $\theta_k \beta_k = 1$, and ϵ is a prescribed accuracy parameter. In our numerical implementation, we set $\beta_0 = 1/\max_{ij} |d_I(V^0)|_{ij}$, $\beta_{\max} = 10^{15}$, $\beta_{\min} = 10^{-15}$ and $\epsilon = 10^{-4}$.

Finally, it shall be mentioned that for the sake of practical performance, the numerical implementation of our augmented Lagrangian method is slightly different from the one described in Subsection 2.2. In particular, we follow a similar scheme as discussed on pp. 405 of [7] to adjust penalty parameter and Lagrangian multipliers. Indeed, they are updated separately rather than simultaneously. Roughly speaking, given $\gamma \in (0, 1)$, we adjust penalty parameter only when the constraint violation is not decreased by a factor γ over the previous minimization. Similarly, we update Lagrangian multipliers only when the constraint violation is decreased by a factor γ over the previous minimization. We choose $\gamma = 0.25$ in our implementation as recommended in [7].

3.4 Numerical results

In this subsection, we conduct numerical experiments for the augmented Lagrangian method detailed in Subsections 2.2 and 3.3.2 for formulation (3.8) (or, equivalently, (3.2)) of sparse PCA on synthetic, random, and real data. In particular, we compare the results of our approach with several existing sparse PCA methods in terms of total explained variance, correlation of PCs, and orthogonality of loading vectors, which include the generalized power methods (Journée et al. [77]), the DSPCA algorithm (d'Aspremont et al. [42]), the SPCA algorithm (Zou et al. [137]), and the sPCA-rSVD algorithm (Shen and Huang [117]). We now list all the methods used in this subsection in Table 3.1. Specifically, the methods with the prefix 'GPower' are the generalized power methods studied in [77], and the method ALSPCA is the augmented Lagrangian method proposed in this Chapter.

Table 3.1: Sparse PCA methods used for our comparison

GPower$_{l_1}$	Single-unit sparse PCA via l_1-penalty
GPower$_{l_0}$	Single-unit sparse PCA via l_0-penalty
GPower$_{l_{1,m}}$	Block sparse PCA via l_1-penalty
GPower$_{l_{0,m}}$	Block sparse PCA via l_0-penalty
DSPCA	DSPCA algorithm
SPCA	SPCA algorithm
rSVD	sPCA-rSVD algorithm with soft thresholding
ALSPCA	Augmented Lagrangian algorithm

As discussed in Subsection 3.2, the PCs obtained from the standard PCA based on sample covariance matrix $\hat{\Sigma} \in \Re^{n\times p}$ are nearly uncorrelated when the sample size is sufficiently large, and the total explained variance by the first r PCs approximately equals the sum of the individual variances of PCs, that is, $\mathrm{Tr}(V^T\hat{\Sigma}V)$, where $V \in \Re^{p\times r}$ consists of the loading vectors of these PCs. However, the PCs found by sparse PCA methods may be correlated with each other, and thus the quantity $\mathrm{Tr}(V^T\hat{\Sigma}V)$ can overestimate much the total explained variance by these PCs due to the overlap among their individual variances. In response to such an overlap, two adjusted total explained variances were proposed in [137, 117]. It is not hard to observe that they can be viewed as the total explained variance of a set of transformed variables from the estimated sparse PCs. Given that these transformed variables can distinct dramatically from those sparse PCs, their total explained variances may also differ much from each other. To alleviate this drawback while taking into account the possible correlations among PCs, we naturally introduce the following *adjusted*

total explained variance for sparse PCs:

$$\text{AdjVar}V = \text{Tr}(V^T\hat{\Sigma}V) - \sqrt{\sum_{i\neq j}(V_i^T\hat{\Sigma}V_j)^2}.$$

It is not hard to show that AdjVar ≥ 0 for any $V \in \Re^{p\times r}$ provided $\hat{\Sigma} \succeq 0$. Clearly, when the PCs are uncorrelated, it becomes the usual total explained variance, that is, $\text{Tr}(V^T\hat{\Sigma}V)$. We can also define the *cumulative percentage of adjusted variance* (CPAV) for the first r sparse PCs as the quotient of the adjusted total explained variance of these PCs and the total explained variance by all standard PCs, that is, $\text{AdjVar}V/\text{Tr}(\hat{\Sigma})$.

Finally, we shall stress that the main purpose of this subsection is to compare the performance of those methods listed in Table 3.1 for finding the sparse PCs that nearly enjoy the three important properties possessed by the standard PCA (see Subsection 3.1). Therefore, we will not compare the speed of these methods. Nevertheless, it shall be mentioned that our method, that is, ALSPCA, is a first-order method and capable of solving large-scale problems within a reasonable amount of time as demonstrated in our experiments presented in Subsection 3.4.4.

3.4.1 Synthetic data

In this subsection we use the synthetic data introduced by Zou et al. [137] to test the effectiveness of our approach ALSPCA for finding sparse PCs.

The synthetic example [137] considers three hidden factors:

$$V_1 \sim N(0, 290), \qquad V_2 \sim N(0, 300), \qquad V_3 = -0.3V_1 + 0.925V_2 + \epsilon, \qquad \epsilon \sim N(0, 1),$$

where V_1, V_2 and ϵ are independent. Then the 10 observable variables are generated as follows:

$$\begin{aligned} X_i &= V_1 + \epsilon_i^1, \qquad \epsilon_i^1 \sim N(0,1), \qquad i = 1,2,3,4, \\ X_i &= V_2 + \epsilon_i^2, \qquad \epsilon_i^2 \sim N(0,1), \qquad i = 5,6,7,8, \\ X_i &= V_3 + \epsilon_i^3, \qquad \epsilon_i^3 \sim N(0,1), \qquad i = 9,10, \end{aligned}$$

where ϵ_i^j are independent for $j = 1, 2, 3$ and $i = 1, \ldots, 10$. We will use the actual covariance matrix of $(X_1, \ldots, X_{10})$ to find the standard and sparse PCs, respectively.

We first observe that V_1 and V_2 are independent, but V_3 is a linear combination of V_1 and V_2. Moreover, the variances of these three underlying factors V_1, V_2 and V_3 are 290,

Table 3.2: Loadings of the first two PCs by standard PCA and ALSPCA

Variable	PCA		ALSPCA	
	PC1	PC2	PC1	PC2
X_1	0.1158	0.4785	0	0.5000
X_2	0.1158	0.4785	0	0.5000
X_3	0.1158	0.4785	0	0.5000
X_4	0.1158	0.4785	0	0.5000
X_5	-0.3955	0.1449	-0.5000	0
X_6	-0.3955	0.1449	-0.5000	0
X_7	-0.3955	0.1449	-0.5000	0
X_8	-0.3955	0.1449	-0.5000	0
X_9	-0.4005	-0.0095	0	0
X_{10}	-0.4005	-0.0095	0	0
CPAV (%)	99.72		80.46	

Synthetic data

300, and 283.8, respectively. Thus V_2 is slightly more important than V_1, and they both are more important than V_3. In addition, the first two standard PCs together explain 99.72% of the total variance (see Table 3.2). These observations suggest that: i) the first two sparse PCs may be sufficient to explain most of the variance; and ii) the first sparse PC recovers the most important factor V_2 using (X_5, X_6, X_7, X_8), and the second sparse PC recovers the second important factor V_1 using (X_1, X_2, X_3, X_4). Given that (X_5, X_6, X_7, X_8) and (X_1, X_2, X_3, X_4) are independent, these sparse PCs would be uncorrelated and orthogonal each other.

In our test, we set $r = 2$, $\Delta_{ij} = 0$ for all $i \neq j$, and $\rho = 4$ for formulation (3.8) of sparse PCA. In addition, we choose (3.16) as the termination criterion for ALSPCA with $\epsilon_I = \epsilon_O = 0.1$ and $\epsilon_E = 10^{-3}$. The results of standard PCA and ALSPCA for this example are presented in Table 3.2. The loadings of standard and sparse PCs are given in columns two and three, respectively, and their CPAVs are given in the last row. We clearly see that our sparse PCs are consistent with the ones predicted above. Interestingly, they are identical with the ones obtained by SPCA and DSPCA reported in [137, 42]. For general data, however, these methods may perform quite differently (see Subsection 3.4.2).

3.4.2 Pitprops data

In this subsection we test the performance of our approach ALSPCA for finding sparse PCs on the Pitprops data introduced by Jeffers [74]. We also compare the results with several existing methods [137, 42, 117, 77].

The Pitprops data [74] has 180 observations and 13 measured variables. It is a classic example that illustrates the difficulty of interpreting PCs. Recently, several sparse PCA

Table 3.3: Loadings of the first six PCs by standard PCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	0.4038	0.2178	0.2073	0.0912	0.0826	0.1198
length	0.4055	0.1861	0.2350	0.1027	0.1128	0.1629
moist	0.1244	0.5406	-0.1415	-0.0784	-0.3498	-0.2759
testsg	0.1732	0.4556	-0.3524	-0.0548	-0.3558	-0.0540
ovensg	0.0572	-0.1701	-0.4812	-0.0491	-0.1761	0.6256
ringtop	0.2844	-0.0142	-0.4753	0.0635	0.3158	0.0523
ringbut	0.3998	-0.1897	-0.2531	0.0650	0.2151	0.0026
bowmax	0.2936	-0.1892	0.2431	-0.2856	-0.1853	-0.0551
bowdist	0.3566	0.0171	0.2076	-0.0967	0.1061	0.0342
whorls	0.3789	-0.2485	0.1188	0.2050	-0.1564	-0.1731
clear	-0.0111	0.2053	0.0704	-0.8036	0.3430	0.1753
knots	-0.1151	0.3432	-0.0920	0.3008	0.6003	-0.1698
diaknot	-0.1125	0.3085	0.3261	0.3034	-0.0799	0.6263

Pitprops data

methods [78, 137, 117, 42] have been applied to this data set for finding *six* sparse PCs by using the actual covariance matrix. For ease of comparison, we present the standard PCs, and the sparse PCs by some of those methods in Tables 3.3-3.6, respectively. We shall mention that two groups of sparse PCs were found in [42] by DSPCA with the parameter $k_1 = 5$ or 6, and they have similar sparsity and total explained variance (see [42] for details). Thus we only present the latter one (i.e., the one with $k_1 = 6$) in Table 3.6. Also, we applied the GPower methods [77] to this data set for finding the PCs with the sparsity given by the largest one of those found in [137, 117, 42], and observed that the best result was given by GPower$_{l_0}$. Thus we only report the sparse PCs obtained by GPower$_{l_0}$ in Table 3.7. In addition, we present sparsity, CPAV, non-orthogonality and correlation of the PCs obtained by the standard PCA and sparse PCA methods [137, 117, 42, 77] in columns two to five of Table 3.11, respectively. In particular, the second and fifth columns of this table respectively give sparsity (measured by the number of zero loadings) and CPAV. The third column reports non-orthogonality, which is measured by the maximum absolute difference between 90° and the angles formed by all pairs of loading vectors. Clearly, the smaller value in this column implies the better orthogonality. The fourth column presents the maximum correlation of PCs. Though the PCs given by these sparse PCA methods all have nice sparsity, we observe from Tables 3.11 that they are highly correlated and moreover, almost all of them are far from orthogonal except the ones given by SPCA [137]. To improve the quality of sparse PCs, we next apply our approach ALSPCA, and compare the results with these methods. For all tests below, we choose (3.16) as the termination criterion for ALSPCA with $\epsilon_O = 0.1$ and $\epsilon_I = \epsilon_E = 10^{-3}$.

Table 3.4: Loadings of the first six PCs by SPCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.477	0	0	0	0	0
length	-0.476	0	0	0	0	0
moist	0	0.785	0	0	0	0
testsg	0	0.620	0	0	0	0
ovensg	0.177	0	0.640	0	0	0
ringtop	0	0	0.589	0	0	0
ringbut	-0.250	0	0.492	0	0	0
bowmax	-0.344	-0.021	0	0	0	0
bowdist	-0.416	0	0	0	0	0
whorls	-0.400	0	0	0	0	0
clear	0	0	0	-1	0	0
knots	0	0.013	0	0	-1	0
diaknot	0	0	-0.015	0	0	1

Pitprops data

Table 3.5: Loadings of the first six PCs by rSVD

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.449	0	0	-0.114	0	0
length	-0.460	0	0	-0.102	0	0
moist	0	-0.707	0	0	0	0
testsg	0	-0.707	0	0	0	0
ovensg	0	0	0.550	0	0	-0.744
ringtop	-0.199	0	0.546	-0.176	0	0
ringbut	-0.399	0	0.366	0	0	0
bowmax	-0.279	0	0	0.422	0	0
bowdist	-0.380	0	0	0.283	0	0
whorls	-0.407	0	0	0	0.231	0
clear	0	0	0	-0.785	-0.973	0
knots	0	0	0	-0.265	0	0.161
diaknot	0	0	-0.515	0	0	-0.648

Pitprops data

In the first experiment, we aim to find six nearly uncorrelated and orthogonal sparse PCs by ALSPCA while explaining most of variance. In particular, we set $r = 6$, $\Delta_{ij} = 0.07$ for all $i \neq j$ and $\rho = 0.8$ for formulation (3.8) of sparse PCA. The resulting sparse PCs are presented in Table 3.8, and their sparsity, CPAV, non-orthogonality and correlation are reported in row seven of Table 3.11. We easily observe that our method ALSPCA overall outperforms the other sparse PCA methods substantially in all aspects except sparsity. Naturally, we can improve the sparsity by increasing the values of ρ, yet the total explained variance may be sacrificed as demonstrated in our next experiment.

We now attempt to find six PCs with similar correlation and orthogonality but higher sparsity than those given in the above experiment. For this purpose, we set $\Delta_{ij} = 0.07$ for all $i \neq j$ and choose $\rho = 2.1$ for problem (3.8) in this experiment. The resulting sparse PCs are presented in Table 3.9, and their CPAV, non-orthogonality and correlation of these PCs

Table 3.6: Loadings of the first six PCs by DSPCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.4907	0	0	0	0	0
length	-0.5067	0	0	0	0	0
moist	0	0.7071	0	0	0	0
testsg	0	0.7071	0	0	0	0
ovensg	0	0	0	0	-1.0000	0
ringtop	-0.0670	0	-0.8731	0	0	0
ringbut	-0.3566	0	-0.4841	0	0	0
bowmax	-0.2335	0	0	0	0	0
bowdist	-0.3861	0	0	0	0	0
whorls	-0.4089	0	0	0	0	0
clear	0	0	0	0	0	1.0000
knots	0	0	0	1.0000	0	0
diaknot	0	0	0.0569	0	0	0

Pitprops data

Table 3.7: Loadings of the first six PCs by GPower$_{l_0}$

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.4182	0	0	0	0	0
length	-0.4205	0	0	0	0	0
moist	0	-0.7472	0	0	0	0
testsg	-0.1713	-0.6646	0	0	0	0
ovensg	0	0	0	0	-0.7877	0
ringtop	-0.2843	0	0	0	-0.6160	0
ringbut	-0.4039	0	0	0	0	0
bowmax	-0.3002	0	0	0	0	0
bowdist	-0.3677	0	0	0	0	0
whorls	-0.3868	0	0	0	0	0
clear	0	0	0	0	0	1.0000
knots	0	0	0	1.0000	0	0
diaknot	0	0	1.0000	0	0	0

Pitprops data

are given in row eight of Table 3.11. Compared to the PCs found in the above experiment, the ones obtained in this experiment are much more sparse while retaining almost same correlation and orthogonality. However, their CPAV goes down dramatically. Combining the results of these two experiments, we deduce that for the Pitprops data, it seems not possible to extract six highly sparse (e.g., around 60 zero loadings), nearly orthogonal and uncorrelated PCs while explaining most of variance as they may not exist. The following experiment further sustains such a deduction.

Finally we are interested in exploring how the correlation controlling parameters $\Delta_{ij}(i \neq j)$ affect the performance of the sparse PCs. In particular, we set $\Delta_{ij} = 0.5$ for all $i \neq j$ and choose $\rho = 0.7$ for problem (3.8). The resulting sparse PCs are presented in Table 3.10, and their CPAV, non-orthogonality and correlation of these PCs are given in the last row of Table 3.11. We see that these PCs are highly sparse, orthogonal, and explain good

Table 3.8: Loadings of the first six PCs by ALSPCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	0.4394	0	0	0	0	0
length	0.4617	0	0	0	0	0
moist	0.0419	0.4611	-0.1644	0.0688	-0.3127	0
testsg	0.1058	0.7902	0	0	0	0
ovensg	0.0058	0	0	0	0	0
ringtop	0.1302	0	0.2094	0	0	0.9999
ringbut	0.3477	0	0.0515	0	0.3240	0
bowmax	0.2256	-0.3566	0	0	0	0
bowdist	0.4063	0	0	0	0	0
whorls	0.4606	0	0	0	0	-0.0125
clear	0	0.0369	0	-0.9973	0	0
knots	-0.1115	0.1614	-0.0762	0.0239	0.8929	0
diaknot	-0.0487	0.0918	0.9595	0.0137	0	0

Pitprops data: Test I

Table 3.9: Loadings of the first six PCs by ALSPCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	1.0000	0	0	0	0	0
length	0	-0.2916	-0.1421	0	0	-0.0599
moist	0	0.9565	-0.0433	0	0	-0.0183
testsg	0	0	0	0.0786	-0.1330	0
ovensg	0	0	-0.9683	0	0	0
ringtop	0	0	0	0	0	0
ringbut	0	0	0.1949	0	0.2369	0
bowmax	0	0	0	0	0	0
bowdist	0	0	0	0	0	0
whorls	0	0	0	0	0	0
clear	0	0	0	-0.9969	0	0
knots	0	0	-0.0480	0.0109	0.9624	0
diaknot	0	0	-0.0093	0	0	0.9980

Pitprops data: Test II

amount of variance. However, they are quite correlated each other, which is actually not surprising since $\Delta_{ij}(i \neq j)$ are not small. Despite such a drawback, these sparse PCs still overall outperform those obtained by SPCA, rSVD, DSPCA and GPower$_{l_1}$.

From the above experiments, we may conclude that for the Pitprops data, there do not exist six highly sparse, nearly orthogonal and uncorrelated PCs while explaining most of variance. Therefore, the most acceptable sparse PCs seem to be the ones given in Table 3.8.

3.4.3 Gene expression data

In this subsection we test the performance of our approach ALSPCA for finding sparse PCs on the gene expression data. We also compare the results with the GPower methods [77], which are superior to the other existing methods [137, 42, 117] as demonstrated in [77].

The data set used in this subsection is the publicly available gene expression data from

Table 3.10: Loadings of the first six PCs by ALSPCA

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	0.4051	0	0	0	0	0
length	0.4248	0	0	0	0	0
moist	0	0.7262	0	0	0	0
testsg	0.0018	0.6875	0	0	0	0
ovensg	0	0	-1.0000	0	0	0
ringtop	0.1856	0	0	0	0	0
ringbut	0.4123	0	0	0	0	0
bowmax	0.3278	0	0	0	0	0
bowdist	0.3830	0	0	0	0	0
whorls	0.4437	-0.0028	0	0	0	0
clear	0	0	0	-1.0000	0	0
knots	0	0	0	0	1.0000	0
diaknot	0	0	0	0	0	1.0000

Pitprops data: Test III

Table 3.11: Comparison of SPCA, rSVD, DSPCA, GPower$_{l_0}$ and ALSPCA

Method	Sparsity	Non-orthogonality	Correlation	CPAV (%)
PCA	0	0	0	87.00
SPCA	60	0.86	0.395	66.21
rSVD	53	14.76	0.459	67.04
DSPCA	63	13.63	0.573	60.97
GPower$_{l_0}$	63	10.09	0.353	64.15
ALSPCA-1	46	0.03	0.082	69.55
ALSPCA-2	60	0.03	0.084	39.42
ALSPCA-3	63	0.00	0.222	65.97

Pitprops data

http://icbp.lbl.gov/breastcancer/, and described in Chin et al. [38], consisting of 19672 gene expression measurements on 89 samples (that is, $p = 19672$, $n = 89$). We aim to extract r number of PCs with around 80% zeros by ALSPCA and GPower methods [77] for $r = 5, 10, 15, 20, 25$, respectively. For all tests below, we set $\Delta_{ij} = 3$ for all $i \neq j$ for problem (3.8) and choose (3.16) as the termination criterion for ALSPCA with $\epsilon_E = 0.5$ and $\epsilon_O = 0.1$.

The sparsity, CPAV, non-orthogonality and correlation of the PCs obtained by the standard PCA, ALSPCA and GPower methods are presented in columns two to five of Tables 3.12-3.16 for $r = 5, 10, 15, 20, 25$, respectively. In particular, the second and fifth columns of these tables respectively give sparsity (that is, the percentage of zeros in loadings) and CPAV. The third column reports non-orthogonality, which is measured by the maximum absolute difference between 90° and the angles formed by all pairs of loading vectors. Evidently, the smaller value in this column implies the better orthogonality. The fourth column presents the maximum correlation of PCs. It is clear that the standard PCs are completely dense. We also observe that the sparse PCs given by our method are almost uncorrelated

and their loading vectors are nearly orthogonal, which are consistently much superior to the GPower methods. Though the CPAV for GPower methods is better than our method, the CPAV for GPower methods may not be a close measurement of the actual total explained variance as their sparse PCs are highly correlated. But for our method, the sparse PCs are almost uncorrelated and thus the CPAV can measure well their actual total explained variance.

Table 3.12: Performance on the gene expression data for $r = 5$

Method	Sparsity (%)	Non-orthogonality	Correlation	CPAV (%)
PCA	0	0	0	34.77
GPower$_{l_1}$	80.14	7.56	0.348	22.17
GPower$_{l_0}$	79.70	5.47	0.223	22.79
GPower$_{l_1,m}$	79.64	7.39	0.274	22.68
GPower$_{l_0,m}$	80.36	12.47	0.452	22.23
ALSPCA	80.43	0.07	0.010	20.56

Table 3.13: Performance on the gene expression data for $r = 10$

Method	Sparsity (%)	Non-orthogonality	Correlation	CPAV (%)
PCA	0	0	0	46.16
GPower$_{l_1}$	80.11	4.93	0.387	31.16
GPower$_{l_0}$	79.84	4.62	0.375	31.45
GPower$_{l_1,m}$	79.95	6.31	0.332	31.80
GPower$_{l_0,m}$	80.36	6.45	0.326	31.59
ALSPCA	80.51	0.01	0.017	29.85

Table 3.14: Performance on the gene expression data for $r = 15$

Method	Sparsity (%)	Non-orthogonality	Correlation	CPAV (%)
PCA	0	0	0	53.27
GPower$_{l_1}$	79.56	4.73	0.253	38.29
GPower$_{l_0}$	79.84	4.02	0.284	38.32
GPower$_{l_1,m}$	79.39	5.94	0.347	38.31
GPower$_{l_0,m}$	79.99	5.18	0.307	38.19
ALSPCA	80.16	0.01	0.014	33.92

3.4.4 Random data

In this subsection we conduct experiments on a set of randomly generated data to test how the size of data matrix X, the sparsity controlling parameter ρ, and the number of components r affect the computational speed of our ALSPCA method.

First, we randomly generate 100 centered data matrices X with size $n \times p$ that is specified in the tables below. For all tests, we set $\Delta_{ij} = 0.1$ for all $i \neq j$ for problem (3.8) and choose (3.16) as the termination criterion for ALSPCA with $\epsilon_E = 0.1$ and $\epsilon_O = 0.1$. In the first

Table 3.15: Performance on the gene expression data for $r = 20$

Method	Sparsity (%)	Non-orthogonality	Correlation	CPAV (%)
PCA	0	0	0	59.60
GPower$_{l_1}$	79.51	4.37	0.280	43.30
GPower$_{l_0}$	80.16	4.52	0.245	43.12
GPower$_{l_1,m}$	79.61	4.48	0.317	42.98
GPower$_{l_0,m}$	80.40	4.18	0.255	43.25
ALSPCA	80.66	0.11	0.037	39.59

Table 3.16: Performance on the gene expression data for $r = 25$

Method	Sparsity (%)	Non-orthogonality	Correlation	CPAV (%)
PCA	0	0	0	64.67
GPower$_{l_1}$	79.48	3.60	0.237	47.74
GPower$_{l_0}$	79.94	3.05	0.296	47.76
GPower$_{l_1,m}$	79.49	5.05	0.275	47.85
GPower$_{l_0,m}$	80.39	5.00	0.237	47.45
ALSPCA	80.68	0.02	0.021	43.66

test, we aim to extract five sparse PCs by ALSPCA with $\rho = 0.001, 0.01, 0.1, 1$, respectively. In the second test, we aim to extract 5 to 25 PCs with a fixed $\rho = 0.1$ by ALSPCA. In the third test, we fix the sparsity (that is, percentage of zeros) of the PC loadings to 80% and find r number of sparse PCs by ALSPCA with $r = 5, 10, 15, 20, 25$, respectively. The average CPU times (in seconds) of ALSPCA over the above 100 instances are reported in Tables 3.17-3.19. We observe that ALSPCA is capable of solving all problems within reasonable amount of time. It seems that the CPU time grows linearly as the problem size, sparsity controlling parameter ρ, and number of components r increase.

Table 3.17: Average CPU time of ALSPCA on random data for $r = 5$

$n \times p$	$\rho = 0.001$	$\rho = 0.01$	$\rho = 0.1$	$\rho = 1$
50×500	0.4	0.8	1.2	4.9
100×1000	1.2	1.5	2.4	9.5
250×2500	3.7	4.4	13.3	38.8
500×5000	8.8	13.4	15.6	65.6
750×7500	13.6	24.0	33.2	96.3

3.5 Concluding remarks

In this chapter we proposed a new formulation of sparse PCA for finding sparse and nearly uncorrelated principal components (PCs) with orthogonal loading vectors while explaining as much of the total variance as possible. We also applied the augmented Lagrangian method proposed in Chapter 2 for solving a class of nonsmooth constrained optimization problems, which is well suited for our formulation of sparse PCA. Finally, we

Table 3.18: Average CPU time of ALSPCA on random data for $\rho = 0.1$

$n \times p$	$r = 5$	$r = 10$	$r = 15$	$r = 20$	$r = 25$
50×500	1.2	12.8	24.0	37.6	48.8
100×1000	2.4	16.9	28.7	40.8	144.0
250×2500	13.4	64.2	94.8	125.1	373.6
500×5000	16.5	85.5	141.9	186.6	553.1
750×7500	38.1	96.6	217.6	328.6	798.2

Table 3.19: Average CPU time of ALSPCA on random data for 80% sparsity

$n \times p$	$r = 5$	$r = 10$	$r = 15$	$r = 20$	$r = 25$
50×500	11.5	26.5	33.6	43.0	49.7
100×1000	15.2	29.3	57.8	83.7	102.7
250×2500	20.7	39.5	79.7	98.0	120.0
500×5000	41.5	60.3	91.4	143.1	197.0
750×7500	55.3	90.4	141.7	208.3	255.1

compared our sparse PCA approach with several existing methods on synthetic and real data, respectively. The computational results demonstrate that the sparse PCs produced by our approach substantially outperform those by other methods in terms of total explained variance, correlation of PCs, and orthogonality of loading vectors.

As observed in our experiments, formulation (3.2) is very effective in finding the desired sparse PCs. However, there remains a natural theoretical question for it. Given a set of random variables, suppose there exist sparse and uncorrelated PCs with orthogonal loading vectors while explaining most of variance of the variables. In other words, their actual covariance matrix Σ has few dominant eigenvalues and the associated orthonormal eigenvectors are sparse. Since Σ is typically unknown and only approximated by a sample covariance matrix $\hat{\Sigma}$, one natural question is whether or not there exist some suitable parameters ρ and Δ_{ij} $(i \neq j)$ so that (3.2) is able to recover those sparse PCs almost surely as the sample size becomes sufficiently large.

Chapter 4

Penalty Decomposition Methods

In this Chapter, we first establish the first-order optimality conditions for general l_0 minimization problems and study a class of special l_0 minimization problems. Then we develop the PD methods for general l_0 minimization problems and establish some convergence results for them. Finally, we conduct numerical experiments to test the performance of our PD methods for solving sparse logistic regression, sparse inverse covariance selection, and compressed sensing and present some concluding remarks.

This chapter is based on the paper [91] co-authored with Zhaosong Lu.

4.1 First-order optimality conditions

In this subsection we study the first-order optimality conditions for problems (1.6) and (1.7). 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 (1.6) and (1.7).

Theorem 4.1.1 *Assume that x^* is a local minimizer of problem (1.6). 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

$$\left\{ \begin{bmatrix} g'(x^*)d - v \\ h'(x^*)d \\ (I_{\bar{J}^*})^T d \end{bmatrix} : d \in \mathcal{T}_{\mathcal{X}}(x^*), v \in \Re^m, v_i \leq 0, i \in \mathcal{A}(x^*) \right\} = \Re^m \times \Re^p \times \Re^{|J|-r} \tag{4.1}$$

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

$$\mathcal{A}(x^*) = \{1 \leq i \leq m : g_i(x^*) = 0\}. \tag{4.2}$$

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

$$\begin{aligned} -\nabla f(x^*) - \nabla g(x^*)\lambda^* - \nabla h(x^*)\mu^* - z^* \in \mathcal{N}_{\mathcal{X}}(x^*), \\ \lambda_i^* \geq 0,\ \lambda_i^* g_i(x^*) = 0,\ i = 1, \ldots, m; \qquad z_j^* = 0,\ j \in \bar{J} \cup J^*. \end{aligned} \tag{4.3}$$

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

Proof. By the assumption that x^* is a local minimizer of problem (1.6), one can observe that x^* is also a local minimizer of the following problem:

$$\min_{x \in \mathcal{X}} \{f(x) : g(x) \leq 0,\ h(x) = 0,\ x_{\bar{J}^*} = 0\}. \tag{4.4}$$

Using this observation, (4.1) and Theorem 3.25 of [113], we see that the conclusion holds. ■

Theorem 4.1.2 *Assume that x^* is a local minimizer of problem (1.7). Let $J^* = \{j \in J : x_j^* \neq 0\}$ and $\bar{J}^* = J \setminus J^*$. suppose that the following Robinson condition*

$$\left\{ \begin{bmatrix} g'(x^*)d - v \\ h'(x^*)d \\ (I_{\bar{J}^*})^T d \end{bmatrix} : d \in \mathcal{T}_{\mathcal{X}}(x^*), v \in \Re^m, v_i \leq 0, i \in \mathcal{A}(x^*) \right\} = \Re^m \times \Re^p \times \Re^{|\bar{J}^*|} \tag{4.5}$$

holds, where $\mathcal{A}(x^*)$ is defined in (4.2). Then, there exists $(\lambda^*, \mu^*, z^*) \in \Re^m \times \Re^p \times \Re^n$ together with x^* satisfying (4.3).

Proof. It is not hard to observe that x^* is a local minimizer of problem (1.7) if and only if x^* is a local minimizer of problem (4.4). Using this observation, (4.5) and Theorem 3.25 of [113], we see that the conclusion holds. ■

We next establish the first-order sufficient optimality conditions for problems (1.6) and (1.7) when the l_0 part is the only nonconvex part.

Theorem 4.1.3 *Assume that h's are affine functions, and f and g's are convex functions. Let x^* be a feasible point of problem (1.6), 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* (4.3) *holds. Then, x^* is a local minimizer of problem (1.6).*

Proof. It follows from the above assumptions and Theorem 3.34 of [113] that x^* is a minimizer of problem (4.4) 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

$$\mathcal{O}_{J^*}(x^*; \epsilon) = \{x \in \mathcal{X} : g(x) \leq 0,\ h(x) = 0,\ x_{\bar{J}^*} = 0,\ \|x - x^*\|_2 < \epsilon\}$$

with $\bar{J}^* = J \setminus J^*$. One can observe from (1.6) that for any $x \in \mathcal{O}(x^*; \epsilon)$, where

$$\mathcal{O}(x^*; \epsilon) = \{x \in \mathcal{X} : g(x) \leq 0,\ h(x) = 0,\ \|x_J\|_0 \leq r,\ \|x - x^*\|_2 < \epsilon\},$$

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 4.1.4 *Assume that h's are affine functions, and f and g's are convex functions. Let x^* be a feasible point of problem (1.7), 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* (4.3) *holds. Then, x^* is a local minimizer of problem (1.7).*

Proof. By virtue of the above assumptions and Theorem 3.34 of [113], we know that x^* is a minimizer of problem (4.4) with $\bar{J}^* = J \setminus J^*$. Also, we observe that any point is a local minimizer of problem (1.7) if and only if it is a local minimizer of problem (4.4). It then implies that x^* is a local minimizer of (1.7). ∎

Remark. The second-order necessary or sufficient optimality conditions for problems (1.6) and (1.7) can be similarly established as above. ∎

4.2 A class of special l_0 minimization

In this subsection 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 4.2.1 *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:*

$$\min \left\{ \phi(x) = \sum_{i=1}^{n} \phi_i(x_i) : \|x\|_0 \leq r,\ x \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \right\}. \tag{4.6}$$

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 (4.6), 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 (4.6). It remains to show that $\phi(x) \geq \phi(x^*)$ for any feasible point x of (4.6). 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 (4.6), which implies that the conclusion holds. ∎

It is straightforward to establish the following result.

Proposition 4.2.2 *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:*

$$\min \left\{ \nu \|x\|_0 + \sum_{i=1}^{n} \phi_i(x_i) : \quad x \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \right\}. \tag{4.7}$$

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 (4.7), where x^* is defined as follows:

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

4.3 Penalty decomposition methods for general l_0 minimization

In this subsection we propose penalty decomposition (PD) methods for solving general l_0 minimization problems (1.6) and (1.7) and establish their convergence. Throughout this subsection, we make the following assumption for problems (1.6) and (1.7).

Assumption 3 *Problems (1.6) and (1.7) are feasible, and moreover, at least a feasible solution, denoted by x^{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^{feas} is readily available or can be observed from the physical background of problems. For example, all application problems discussed in Subsection 4.4 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 (1.6) can be approximately solved as the following problem:

$$\min_{x \in \mathcal{X}} \{ f(x) + \rho(\|u^+\|_2^2 + \|v\|_2^2) : g(x) - u \leq 0,\ h(x) - v = 0,\ \|x_J\|_0 \leq r \}$$

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

4.3.1 Penalty decomposition method for problem (1.6)

In this subsection we propose a PD method for solving problem (1.6) and establish its convergence.

We observe that (1.6) can be equivalently reformulated as

$$\min_{x\in\mathcal{X},y\in\mathcal{Y}}\{f(x):\ g(x)\le 0,\ h(x)=0,\ x_J-y=0\}, \tag{4.8}$$

where

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

The associated quadratic penalty function is defined as follows:

$$q_\varrho(x,y)=f(x)+\frac{\varrho}{2}(\|[g(x)]^+\|_2^2+\|h(x)\|_2^2+\|x_J-y\|_2^2)\quad\forall x\in\mathcal{X},y\in\mathcal{Y} \tag{4.9}$$

for some penalty parameter $\varrho>0$.

We are now ready to propose a PD method for solving problem (4.8) (or equivalently, (1.6)) in which each penalty subproblem is approximately solved by a block coordinate descent (BCD) method.

Penalty decomposition method for (1.6):

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

$$\min\{q_{\varrho_k}(x,y):\ x\in\mathcal{X},\ y\in\mathcal{Y}\} \tag{4.10}$$

by performing steps 1a)-1d):

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

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

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

$$\|\mathcal{P}_\mathcal{X}(x^k-\nabla_x q_{\varrho_k}(x^k,y^k))-x^k\|_2\le\epsilon_k, \tag{4.11}$$

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 (4.11) 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,

$$\max\left\{\frac{\|x_l^k - x_{l-1}^k\|_\infty}{\max(\|x_l^k\|_\infty, 1)}, \frac{\|y_l^k - y_{l-1}^k\|_\infty}{\max(\|y_l^k\|_\infty, 1)}\right\} \;\le\; \epsilon_I \tag{4.12}$$

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

$$\|x^k - y^k\|_\infty \;\le\; \epsilon_O \tag{4.13}$$

for some $\epsilon_O > 0$. Given that problem (4.10) 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 (4.10) 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 (4.10). 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 4.2.1 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 (4.10)

satisfying (4.11) can be found by the BCD method described in steps 1a)-1d). For notational convenience, we omit the index k from (4.10) and consider the BCD method for solving the problem

$$\min\{q_\varrho(x,y) : x \in \mathcal{X},\ y \in \mathcal{Y}\} \tag{4.14}$$

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

Block coordinate descent method for (4.14):

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 4.3.1 *Suppose that $(x^*, y^*) \in \Re^n \times \Re^{|J|}$ is a block coordinate minimizer of problem (4.14), that is,*

$$x^* \in \operatorname{Arg}\min_{x\in\mathcal{X}} q_\varrho(x, y^*), \quad y^* \in \operatorname{Arg}\min_{y\in\mathcal{Y}} q_\varrho(x^*, y). \tag{4.15}$$

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 (4.14).

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 $i \in K$. We claim that

$$(y^* - x_J^*)^T h_y = 0. \tag{4.16}$$

If $\|x_J^*\|_0 > r$, we observe from the second relation of (4.15) and Proposition 4.2.1 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 (4.16) holds. On the other hand, if $\|x_J^*\|_0 \leq r$, one can observe that $y^* = x_J^*$ and thus (4.16) 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 (4.15) and the first-order

optimality condition that $[\nabla_x q_\varrho(x^*, y^*)]^T h_x \geq 0$. Using this inequality along with (4.16) 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 (4.14). ∎

Theorem 4.3.2 *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 block coordinate minimizer of problem (4.14).

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

$$\|\mathcal{P}_{\mathcal{X}}(x^l - \nabla_x q_\varrho(x^l, y^l)) - x^l\|_2 < \epsilon.$$

(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 (4.14).

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

$$q_\varrho(x^{l+1}, y^l) \leq q_\varrho(x, y^l) \quad \forall x \in \mathcal{X}, \tag{4.17}$$

$$q_\varrho(x^l, y^l) \leq q_\varrho(x^l, y) \quad \forall y \in \mathcal{Y}. \tag{4.18}$$

It follows that

$$q_\varrho(x^{l+1}, y^{l+1}) \;\leq\; q_\varrho(x^{l+1}, y^l) \;\leq\; q_\varrho(x^l, y^l) \quad \forall l \geq 1. \tag{4.19}$$

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, (4.19) 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 (4.17) and (4.18) as $l \in L \to \infty$, we have

$$q_\varrho(x^*, y^*) \leq q_\varrho(x, y^*) \quad \forall x \in \mathcal{X}, \tag{4.20}$$

$$q_\varrho(x^*, y^*) \leq q_\varrho(x^*, y) \quad \forall y \in \mathcal{Y}. \tag{4.21}$$

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 (4.20) and (4.21) implies that (x^*, y^*) is a block coordinate minimizer of (4.14) and hence statement (a) holds. Using (4.20) and the first-order optimality condition, we have

$$\|\mathcal{P}_\mathcal{X}(x^* - \nabla_x q_\varrho(x^*, y^*)) - x^*\|_2 = 0.$$

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

$$\lim_{l\in L\to\infty} \|\mathcal{P}_\mathcal{X}(x^l - \nabla_x q_\varrho(x^l, y^l)) - x^l\|_2 = 0,$$

and hence, statement (b) immediately follows. In addition, statement (c) holds due to statement (a) and Lemma 4.3.1. ∎

The following theorem establishes the convergence of the outer iterations of the PD method for solving problem (1.6). 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 (1.6). 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 (1.6).

Theorem 4.3.3 *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 (1.6). 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* (4.1) *holds at x^* for such $\bar{J}^*$. Then, $\{(\lambda^k, \mu^k, \varpi^k)\}_{k\in K}$ is bounded, where*

$$\lambda^k = \varrho_k[g(x^k)]^+, \quad \mu^k = \varrho_k h(x^k), \quad \varpi^k = \varrho_k(x^k_J - y^k). \tag{4.22}$$

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 (4.3) *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 (1.6).*

Proof. In view of (4.9) and our choice of y^k_0 that is specified in step 3), one can observe that

$$f(x^k) + \frac{\varrho_k}{2}(\|[g(x^k)]^+\|_2^2 + \|h(x^k)\|_2^2 + \|x^k_J - y^k\|_2^2) = q_{\varrho_k}(x^k, y^k) \le \min_{x\in\mathcal{X}} q_{\varrho_k}(x, y^k_0) \le \Upsilon \quad \forall k. \tag{4.23}$$

It immediately implies that $\{x^k\} \subseteq \mathcal{X}_\Upsilon$, and hence, $\{x^k\}$ is bounded. Moreover, we can obtain from (4.23) that

$$\|x^k_J - y^k\|_2^2 \le 2[\Upsilon - f(x^k)]/\varrho_k \le 2[\Upsilon - \min_{x\in\mathcal{X}_\Upsilon} f(x)]/\varrho_0,$$

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 (4.11) that $\|s^k\|_2 \le \epsilon_k$ 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

$$-\nabla_x q_{\varrho_k}(x^k, y^k) - s^k \in \mathcal{N}_{\mathcal{X}}(x^k + s^k). \tag{4.24}$$

Using this relation, (4.24), (4.22) and the definition of q_ϱ, we have

$$-\nabla f(x^k) - \nabla g(x^k)\lambda^k - \nabla h(x^k)\mu^k - I_J\varpi^k - s^k \in \mathcal{N}_{\mathcal{X}}(x^k + s^k). \tag{4.25}$$

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 $\{\|(\lambda^k, \mu^k, \varpi^k)\|_2\}_{k\in K} \to \infty$. Let $(\bar\lambda^k, \bar\mu^k, \bar\varpi^k) = (\lambda^k, \mu^k, \varpi^k)/\|(\lambda^k, \mu^k, \varpi^k)\|_2$. 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, $\|(\bar\lambda, \bar\mu, \bar\varpi)\|_2 = 1$. Dividing both sides of (4.25) by $\|(\lambda^k, \mu^k, \varpi^k)\|_2$, 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

$$-\nabla g(x^*)\bar\lambda - \nabla h(x^*)\bar\mu - I_J\bar\varpi \in \mathcal{N}_{\mathcal{X}}(x^*). \tag{4.26}$$

We can see from (4.2) and (4.22) that $\bar\lambda \in \Re^m_+$, and $\bar\lambda_i = 0$ for $i \notin \mathcal{A}(x^*)$. Also, from Proposition 4.2.1 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 (4.1) 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, (4.26), and the facts that $d \in \mathcal{T}_{\mathcal{X}}(x^*)$ and $\bar\varpi_{I^*} = 0$, we have

$$\begin{aligned}
\|\bar\lambda\|_2^2 + \|\bar\mu\|_2^2 + \|\bar\varpi\|_2^2 &= -[(-\bar\lambda)^T\bar\lambda + (-\bar\mu)^T\bar\mu + (-\bar\varpi_{\bar I^*})^T\bar\varpi_{\bar I^*}] \\
&= -[(g'(x^*)d - v)^T\bar\lambda + (h'(x^*)d)^T\bar\mu + ((I_{\bar J^*})^T d)^T\bar\varpi_{\bar I^*}] \\
&= d^T(-\nabla g(x^*)\bar\lambda - \nabla h(x^*)\bar\mu - I_J\bar\varpi) + v^T\bar\lambda \;\le\; 0.
\end{aligned}$$

It yields $(\bar\lambda, \bar\mu, \bar\varpi) = (0, 0, 0)$, which contradicts the identity $\|(\bar\lambda, \bar\mu, \bar\varpi)\|_2 = 1$. 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 (4.25) 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 (4.3) 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 (4.22) that $\lambda^k_i \geq 0$ and $\lambda^k_i g_i(x^k) = 0$ for all i, which immediately lead to the second relation of (4.3). Hence, $(\lambda^*, \mu^*, \varpi^*)$ together with x^* satisfies (4.3). 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 4.1.3 hold. It then follows from Theorem 4.1.3 that x^* is a local minimizer of (1.6). ∎

4.3.2 Penalty decomposition method for problem (1.7)

In this subsection we propose a PD method for solving problem (1.7) and establish some convergence results for it.

We observe that problem (1.7) can be equivalently reformulated as

$$\min_{x\in\mathcal{X}, y\in\Re^{|J|}} \{f(x) + \nu\|y\|_0 : \ g(x) \leq 0, \ h(x) = 0, \ \ x_J - y = 0\}. \tag{4.27}$$

The associated quadratic penalty function for (4.27) is defined as

$$p_\varrho(x, y) := f(x) + \nu\|y\|_0 + \frac{\varrho}{2}(\|[g(x)]^+\|_2^2 + \|h(x)\|_2^2 + \|x_J - y\|_2^2) \quad \forall x \in \mathcal{X}, y \in \Re^{|J|} \tag{4.28}$$

for some penalty parameter $\varrho > 0$.

We are now ready to present the PD method for solving (4.27) (or, equivalently, (1.7)) in which each penalty subproblem is approximately solved by a BCD method.

Penalty decomposition method for (1.7):

Let $\{\epsilon_k\}$ be a positive decreasing sequence. Let $\varrho_0 > 0$, $\sigma > 1$ be given, and let q_ϱ be defined in (4.9). 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

$$\min\{p_{\varrho_k}(x, y) : \ x \in \mathcal{X}, \ y \in \Re^{|J|}\} \tag{4.29}$$

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

$$\|\mathcal{P}_{\mathcal{X}}(x^k - \nabla_x q_{\varrho_k}(x^k, y^k)) - x^k\|_2 \le \epsilon_k, \tag{4.30}$$

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 $y_0^{k+1} := x^{\mathrm{feas}}$. 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 4.3.1 can also be applied to this PD method. In addition, one can apply a similar strategy as mentioned in Subsection 4.3.1 to enhance the performance of the BCD method for solving (4.29). Finally, in view of Proposition 4.2.2, 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 (4.29) satisfying (4.30) can be found by the BCD method described in steps 1a)-1d). For convenience of presentation, we omit the index k from (4.29) and consider the BCD method for solving the following problem:

$$\min\{p_\varrho(x, y) : \ x \in \mathcal{X}, \ y \in \Re^{|J|}\} \tag{4.31}$$

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 4.3.1 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 4.3.4 *Suppose that $(x^*, y^*) \in \Re^n \times \Re^{|J|}$ is a block coordinate minimizer of problem* (4.31)*, that is,*

$$x^* \in \mathrm{Arg}\min_{x\in\mathcal{X}} p_\varrho(x, y^*), \quad y^* \in \mathrm{Arg}\min_{y\in\Re^{|J|}} p_\varrho(x^*, y). \tag{4.32}$$

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 (4.31)*.*

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 (4.32) and Proposition 4.2.2 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

$$\rho(y^* - x_J^*)^T h_y + \nu\|y^* + h_y\|_0 - \nu\|y^*\|_0 \;=\; -\rho\sum_{i\notin K} x^*_{J(i)}(h_y)_i + \nu\sum_{i\notin K}\|(h_y)_i\|_0 \;\geq\; 0. \tag{4.33}$$

Let q_ϱ be defined in (4.9). By the first relation of (4.32) and a similar argument as in Lemma 4.3.1, we have $[\nabla_x q_\varrho(x^*, y^*)]^T h_x \geq 0$. Using this inequality along with (4.33) 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 our choice of h_x and h_y implies that (x^*, y^*) is a local minimizer of (4.31). ∎

Theorem 4.3.5 *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 block coordinate minimizer of problem (4.31)*.*

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

$$\|\mathcal{P}_\mathcal{X}(x^l - \nabla_x q_\varrho(x^l, y^l)) - x^l\|_2 < \epsilon$$

where the function q_ϱ is defined in (4.9)*.*

(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 (4.31).

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

$$p_\varrho(x^{l+1}, y^l) \leq p_\varrho(x, y^l) \quad \forall x \in \mathcal{X}, \tag{4.34}$$

$$p_\varrho(x^l, y^l) \leq p_\varrho(x^l, y) \quad \forall y \in \Re^{|J|}. \tag{4.35}$$

Using these relations and a similar argument as in the proof of Theorem 4.3.2, one can show that $\lim_{l\to\infty} p_\varrho(x^l, y^l)$ exists, and moreover,

$$\lim_{l\to\infty} p_\varrho(x^l, y^l) = \lim_{l\to\infty} p_\varrho(x^{l+1}, y^l). \tag{4.36}$$

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 $\mathcal{X}$. For notational convenience, let

$$F(x) := f(x) + \frac{\varrho}{2}(\|[g(x)]^+\|_2^2 + \|h(x)\|_2^2).$$

It then follows from (4.28) that

$$p_\varrho(x, y) = F(x) + \nu\|y\|_0 + \frac{\varrho}{2}\|x_J - y\|_2^2, \quad \forall x \in \mathcal{X}, y \in \Re^{|J|}. \tag{4.37}$$

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, (4.35) and (4.37), we obtain that, when $l \in L$ is sufficiently large,

$$p_\varrho(x^l, y) \;\geq\; p_\varrho(x^l, y^l) \;=\; F(x^l) + \nu\|y^l\|_0 + \frac{\varrho}{2}\|x_J^l - y^l\|_2^2 \;\geq\; F(x^l) + \nu\|y^*\|_0 + \frac{\varrho}{2}\|x_J^l - y^l\|_2^2.$$

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

$$p_\varrho(x^*, y) \;\geq\; F(x^*) + \nu\|y^*\|_0 + \frac{\varrho}{2}\|x_J^* - y^*\|_2^2 \;=\; p_\varrho(x^*, y^*), \quad \forall y \in \Re^{|J|}. \tag{4.38}$$

In addition, it follows from (4.34) and (4.37) that

$$\begin{array}{rcl} F(x) + \frac{1}{2}\|x_J - y^l\|_2^2 & = & p_\varrho(x, y^l) - \nu\|y^l\|_0 \;\geq\; p_\varrho(x^{l+1}, y^l) - \nu\|y^l\|_0 \\ & = & F(x^{l+1}) + \frac{1}{2}\|x_J^{l+1} - y^l\|_2^2, \quad \forall x \in \mathcal{X}. \end{array} \tag{4.39}$$

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

$$\begin{aligned}
&\lim_{l\in\bar{L}\to\infty} F(x^{l+1}) + \tfrac{1}{2}\|x_J^{l+1} - y^l\|_2^2 \;=\; \lim_{l\in\bar{L}\to\infty} p_\varrho(x^{l+1}, y^l) - \nu\|y^l\|_0 \\
&= \lim_{l\in\bar{L}\to\infty} p_\varrho(x^{l+1}, y^l) - \nu \lim_{l\in\bar{L}\to\infty} \|y^l\|_0 \;=\; \lim_{l\in\bar{L}\to\infty} p_\varrho(x^l, y^l) - \nu \lim_{l\in\bar{L}\to\infty} \|y^l\|_0 \\
&= \lim_{l\in\bar{L}\to\infty} p_\varrho(x^l, y^l) - \nu\|y^l\|_0 = \lim_{l\in\bar{L}\to\infty} F(x^l) + \tfrac{1}{2}\|x_J^l - y^l\|_2^2 \;=\; F(x^*) + \tfrac{1}{2}\|x_J^* - y^*\|_2^2,
\end{aligned}$$

where the third equality is due to (4.36). Using this relation and taking limits on both sides of (4.39) as $l \in \bar{L} \to \infty$, we further have

$$F(x) + \frac{1}{2}\|x_J - y^*\|_2^2 \;\geq\; F(x^*) + \frac{1}{2}\|x_J^* - y^*\|_2^2, \quad \forall x \in \mathcal{X},$$

which together with (4.28) yields

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

This relation along with (4.38) implies that (x^*, y^*) is a block coordinate minimizer of (4.31) and hence statement (a) holds. Statement (b) can be similarly proved as that of Theorem 4.3.2. In addition, statement (c) holds due to statement (a) and Lemma 4.3.4. ∎

We next establish the convergence of the outer iterations of the PD method for solving problem (1.7). 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 (1.7). Moreover, when the l_0 part is the only nonconvex part, we show that the accumulation point is a local minimizer of (1.7).

Theorem 4.3.6 *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 (1.7).

(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 (4.5) 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* (4.3) *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 (1.7).*

Proof. Statement (a) and (b) can be similarly proved as those of Theorem 4.3.3. We now show that statement (c) holds. Let $I^* = \{i : J(i) \in J^*\}$. From Proposition 4.2.2 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 4.3.3. The last conclusion of this statement holds due to its second conclusion and Theorem 4.1.4. ∎

4.4 Numerical results

In this subsection, we conduct numerical experiments to test the performance of our PD methods proposed in Subsection 4.3 by applying them to sparse logistic regression, sparse inverse covariance selection, and compressed sensing problems. The codes of all the methods implemented in this subsection 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).

4.4.1 Sparse logistic regression problem

In this subsection, we apply the PD method studied in Subsection 4.3.1. In the literature, one common approach for finding an approximate solution to (1.5) is by solving the following l_1 regularization problem:

$$\min_{v,w} l_{\rm avg}(v, w) + \lambda\|w\|_1 \tag{4.40}$$

for some regularization parameter $\lambda \geq 0$ (see, for example, [80, 54, 102, 82, 85, 119]). Our aim below is to apply the PD method studied in Subsection 4.3.1 to solve (1.5) directly and compare the results with one of the l_1 relaxation solvers, that is, SLEP [85].

Letting $x = (v, w)$, $J = \{2, \ldots, p + 1\}$ and $f(x) = l_{\rm avg}(x_1, x_J)$, we can see that problem (1.5) is in the form of (1.6). Therefore, the PD method proposed in Subsection 4.3.1 can

be suitably applied to solve (1.5). Also, we observe that the main computation effort of the PD method when applied to (1.5) lies in solving the subproblem arising in step 1a), which is in the form of

$$\min_x \left\{ l_{\text{avg}}(x_1, x_J) + \frac{\varrho}{2}\|x - c\|_2^2 : \ x \in \Re^{p+1} \right\} \tag{4.41}$$

for some $\varrho > 0$ and $c \in \Re^{p+1}$. To efficiently solve (4.41), we apply the nonmonotone projected gradient method proposed in [9, Algorithm 2.2]; in particular, we set its parameter $M = 2$ and terminate the method when $\|\nabla F(x)\|_2/\max\{|F(x)|, 1\} \leq 10^{-4}$, where $F(x)$ denotes the objective function of (4.41).

We now address the initialization and the termination criteria for our PD method when applied to (1.5). In particular, we randomly generate $z \in \Re^{p+1}$ such that $\|z_J\|_0 \leq r$ 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 (4.12) and (4.13) 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 (1.5) on some real and random data. We also compare the quality of the approximate solutions of (1.5) obtained by our method with that of (4.40) found by a first-order solver SLEP [85]. 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 [98] and other sources [66]. The first data set is the colon tumor gene expression data [66] with more features than samples; the second one is the ionosphere data [98] with less features than samples; and the third one is the Internet advertisements data [98] 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 (4.40) with four different values of λ, which are the same ones as used in [80], 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 [80]. For each such λ, let w_λ^* be the approximate optimal w obtained by SLEP. We then apply our PD method to solve problem (1.5) 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

Table 4.1: 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

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

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

where

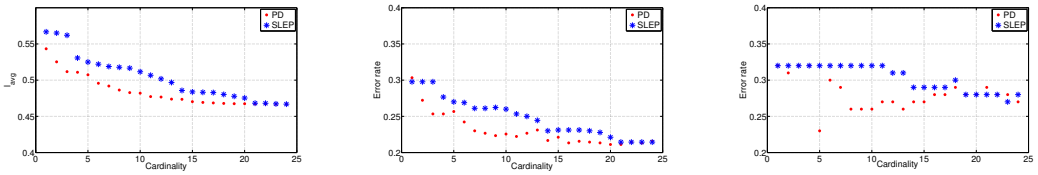
$$\text{sgn}(t) = \begin{cases} +1 & \text{if } t > 0, \\ -1 & \text{otherwise.} \end{cases}$$

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

$$\text{Error} := \left\{ \sum_{i=1}^{n} \|\phi(z^i) - b_i\|_0 / n \right\} \times 100\%. \tag{4.42}$$

The computational results are presented in Table 4.1. 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 cardinality of both solutions is the same.

The out-of-sample error rate is often used to evaluate the quality of a model vector, which is a slight modification of (4.42) by taking sum over the testing samples rather than the training samples. It usually depends on the quality and amount of training samples. For example, when the ratio between number of training samples and features is small, the out-of-sample error rate is usually high for most of models. Due to this reason, the above data



(a) Average logistic loss (b) Error rate for training data (c) Error rate for testing data

Figure 4.1: Sparse recovery.

sets may not be appropriate for evaluating out-of-sample error rate. Instead, we download a real data called "German" from the UCI machine learning bench market repository [98]. This data set contains $1,000$ samples and 24 features, which has a reasonably high sample-to-feature ratio. It is thus a suitable data set to evaluate out-of-sample error rate. We randomly partition those samples into two parts: one consisting of 900 samples used as training data and another consisting of 100 samples used as testing data. Similarly as above, we first apply SLEP to (4.40) with a sequence of suitably chosen λ to obtain solutions with cardinalities from 1 to 24. For PD, we simply set r to be 1 to 24. In this way, the solutions of PD and SLEP are of the same cardinality. The results of this experiment are presented in Figure 4.1. We can see that PD generally outperforms SLEP in terms of solution quality since it achieves smaller average logistic loss and lower error rate for both training and testing data.

In next 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 [80]. 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) 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 (4.40) 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 (1.5) with $r = \|w^*_\lambda\|_0$ so that the resulting approximate optimal

Table 4.2: Computational results on random data sets

Size			SLEP			PD		
$n \times p$	$\lambda/\lambda_{\max}$	r	l_{avg}	Error (%)	Time	l_{avg}	Error (%)	Time
2000×5000	0.9	27.7	0.6359	8.01	4.3	0.1802	6.98	89.6
	0.7	91.4	0.5046	3.43	11.2	0.0550	2.37	159.6
	0.5	161.1	0.3827	1.98	17.7	0.0075	0.07	295.3
	0.3	238.9	0.2639	1.23	21.8	0.0022	0	216.1
	0.1	330.1	0.1289	0.46	19.6	0.0015	0	130.5
5000×2000	0.9	17.7	0.6380	8.49	3.5	0.2254	8.09	154.3
	0.7	65.7	0.5036	3.16	10.1	0.0372	1.55	296.6
	0.5	121.0	0.3764	1.48	16.4	0.0042	0	299.8
	0.3	180.8	0.2517	0.57	22.3	0.0018	0	190.0
	0.1	255.2	0.1114	0.04	23.7	0.0013	0	124.2
5000×5000	0.9	30.7	0.6341	7.02	4.8	0.1761	6.55	125.8
	0.7	105.7	0.5022	2.95	12.9	0.0355	1.47	255.3
	0.5	192.0	0.3793	1.63	20.3	0.0042	0	325.4
	0.3	278.3	0.2592	0.88	25.1	0.0020	0	187.4
	0.1	397.0	0.1231	0.19	24.4	0.0015	0	113.4

w is at least as sparse as w^*_λ. The average results of each data set over 100 instances are reported in Table 4.2. 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 (1.5) obtained by our PD method is generally better than that of (4.40) found by SLEP when the same cardinality is considered. This observation is actually not surprising as (4.40) is a relaxation of (1.5).

4.4.2 Sparse inverse covariance selection problem

In this subsection, we apply the PD method proposed in Subsection 4.3.1. In the literature, one common approach for finding an approximate solution to (1.4) is by solving the following l_1 regularization problem:

$$\begin{array}{ll} \max\limits_{X \succeq 0} & \log\det X - \mathbf{\Sigma^t} \bullet X - \sum\limits_{(i,j)\in\bar{\Omega}} \rho_{ij}|X_{ij}| \\ \text{s.t.} & X_{ij} = 0 \quad \forall (i,j) \in \Omega, \end{array} \tag{4.43}$$

where $\{\rho_{ij}\}_{(i,j)\in\bar{\Omega}}$ is a set of regularization parameters (see, for example, [41, 43, 4, 87, 88, 62, 132, 86]). Our goal below is to apply the PD method studied in Subsection 4.3.1 to solve (1.4) directly and compare the results with one of the l_1 relaxation methods, that is, the proximal point algorithm (PPA) [132].

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 (1.4) is in the form of (1.6) and thus it can be suitably solved by the PD method proposed in Subsection 4.3.1 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 (1.4) lies in solving the subproblem arising in step 1a), which is in the form of

$$\min_{X\succeq 0} \left\{ -\log\det X + \frac{\varrho}{2}\|X - C\|_F^2 : X_{ij} = 0\ \forall (i,j) \in \Omega \right\} \tag{4.44}$$

for some $\varrho > 0$ and $C \in \mathcal{S}^p$. Given that problem (4.44) 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 4.3.1 to solve (1.4). For this PD method, the subproblem arising in step 1a) is now in the form of

$$\min_{X} \left\{ -\log\det X + \frac{\varrho}{2}\|X - C\|_F^2 : X \succeq 0 \right\} \tag{4.45}$$

for some $\varrho > 0$ and $C \in \mathcal{S}^p$. It can be shown that problem (4.45) 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, [132]). Also, it follows from Proposition 4.2.1 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 $Y_0^0 = (\widetilde{\mathscr{D}}(\mathbf{\Sigma^t}))^{-1}$, the initial penalty parameter $\varrho_0 = 1$, and the parameter $\sigma = \sqrt{10}$. In addition, we use (4.13) and

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

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 (1.4) on some random and real data. We also compare the quality of the approximate solutions of (1.4) obtained by our method with that of (4.43) found by the proximal point algorithm (PPA) [132]. Both methods call the LAPACK routine **dsyevd.f** [81] 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 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 [41, 87, 88, 132, 86]. 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

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

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:

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

where ϑ is a small positive number. Specifically, we choose $\tau = 0.15$, $\vartheta = 1.0e-4$, $\delta = 10\%$, 50% and 100%, respectively. Clearly, $\delta = 100\%$ means that $\Omega = \emptyset$, that is, none of zero entries of the actual sparse inverse covariance matrix is known beforehand. 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 (4.43) 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 (1.4) 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 [86], 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:

$$\text{Loss} := \frac{1}{p}(\mathbf{\Sigma^t} \bullet X - \log\det(\mathbf{\Sigma^t}X) - p).$$

The results of PPA and the PD method on these instances are presented in Tables 4.3-4.5, respectively. In each table, the order p of $\mathbf{\Sigma^t}$ is given in column one. The size of Ω is

Table 4.3: Computational results for $\delta = 10\%$

Problem				PPA			PD		
p	$\|\Omega\|$	$\rho_{\bar{\Omega}}$	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

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 (1.4)), 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 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 [41, 88]. 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 $\mathbf{\Sigma^t}$ 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 (1.4), 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 (4.43) and (1.4) with the aforementioned ρ_{ij} and r, respectively. In Figure 4.2, we plot the sparsity patterns of the original inverse covariance matrix $(\mathbf{\Sigma^t})^{-1}$, the noisy inverse sample covariance matrix $\mathbf{\Sigma^t}^{-1}$, and the approximate solutions to (4.43) and (1.4) obtained by PPA and our PD method, respectively. We first observe that the sparsity of both solutions is the same as

Table 4.4: Computational results for $\delta = 50\%$

Problem				PPA			PD		
p	$\|\Omega\|$	$\rho_{\bar{\Omega}}$	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

$(\mathbf{\Sigma^t})^{-1}$. 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 4.6. One can see that the solution of our PD method achieves much larger log-likelihood and smaller normalized entropy loss.

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, [67, 105, 135, 48, 86]). We first pre-process the data by the same procedure as described in [86] to obtain a sample covariance matrix $\mathbf{\Sigma^t}$, and set $\Omega = \emptyset$ and $\rho_{ij} = \rho_{\bar{\Omega}}$ for some $\rho_{\bar{\Omega}} > 0$. We apply PPA to solve problem (4.43) 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 (1.4) 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 $\mathbf{\Sigma^t}$. The results of PPA and our PD method on these two data sets are presented in Table 4.7. 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

Table 4.5: Computational results for $\delta = 100\%$

Problem				PPA			PD		
p	$\|\Omega\|$	$\rho_{\bar{\Omega}}$	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

Table 4.6: Numerical results for sparse recovery

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

method is generally faster than PPA. Moreover, our PD method outperforms PPA in terms of log-likelihood and normalized entropy loss.

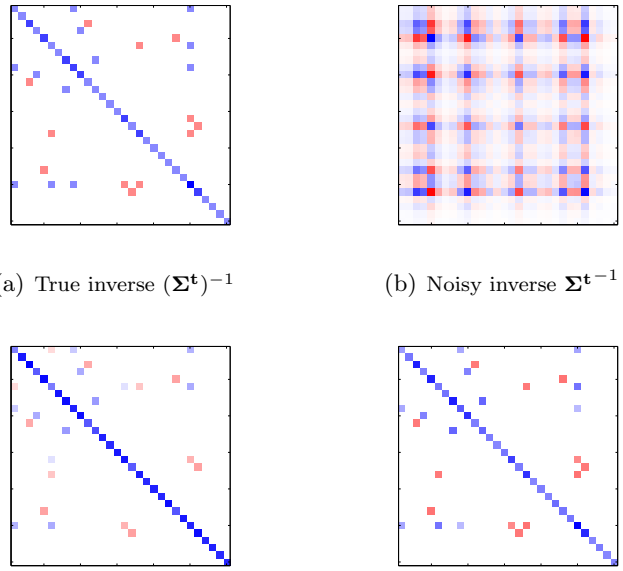
As a summary, the above experiments show that the quality of the approximate solution of (1.4) obtained by our PD method is generally better than that of (4.43) found by PPA when the same cardinality is considered.

4.4.3 Compressed sensing

In this subsection, we apply the PD methods proposed in Subsection 4.3 to solve the compressed sensing (CS) problem. One popular approach for finding an approximate solution to (1.1) is to solve the following l_1 regularization problem:

$$\min_{x \in \Re^p} \{\|x\|_1 : \ Ax = b\}, \tag{4.46}$$

where A is a full row rank matrix (see, for example, [130, 35]). Our aim below is to apply the PD method studied in Subsection 4.3.2 to solve problem (1.1) directly and compare with one of the l_1 relaxation solvers, that is, SPGL1 [130].



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

(c) Approximate solution of (4.43) (d) Approximate solution of (1.4)

Figure 4.2: Sparse recovery.

Clearly, problem (1.1) is in the form of (1.7) and thus the PD method proposed in Subsection 4.3.2 can be suitably applied to solve (1.1). Also, one can observe that the main computation effort of the PD method when applied to (1.1) lies in solving the subproblem arising in step 1a), which is in the form of

$$\min_x \{\|x - c\|_2^2 : Ax = b\} \tag{4.47}$$

for some $c \in \Re^p$. It is well known that problem (4.47) 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 (1.1) 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 (4.12) and

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

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

Table 4.7: 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

We next conduct experiment to test the performance of our PD method for finding a sparse approximate solution to problem (1.1) on the data sets from Sparco [131]. [1] We also compare the quality of such sparse approximate solution with that of the one found by a first-order solver SPGL1 [130] applied to (4.46). For the latter method, we use the default value for all parameters. To evaluate the quality of these sparse approximate solutions, we adopt a similar criterion as described in [106, 22]. Indeed, suppose that B be a basis matrix for a given signal f for which we wish to find a sparse recovery. Given a sparse vector x, the corresponding sparse approximate signal is $f_x = Bx$. The associated mean squared error is defined as

$$\text{MSE} := \|f_x - f\|_2/p,$$

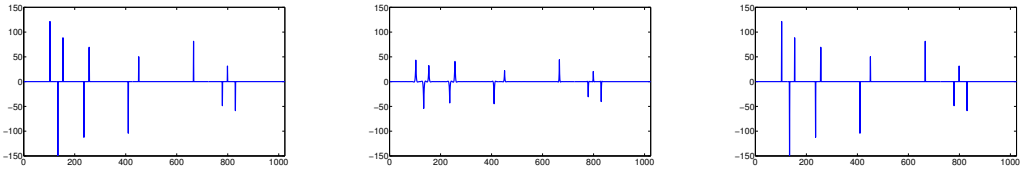
We only report the computational results for 10 data sets in Table 4.8 since the performance difference between PD and SPGL1 on the other data sets is similar to that on these data sets. In detail, the data name and the size of data are given in the first three columns. The MSE, the solution cardinality and the CPU time for both methods are reported in columns four to nine, respectively. It can be observed that SPGL1 is generally faster than PD, but PD generally provides much more sparse solution while with lower MSE. Thus, the resulting signal by PD is less noisy. For example, we plot in Figure 4.3 the results for data **blknheavi** whose actual sparse representation is pre-known. It can be seen that the signal recovered by SPGL1 has more noise than the one by PD.

In the remainder of this subsection we consider the CS problem with noisy observation. One popular approach for finding an approximate solution to (1.2) is to solve the following

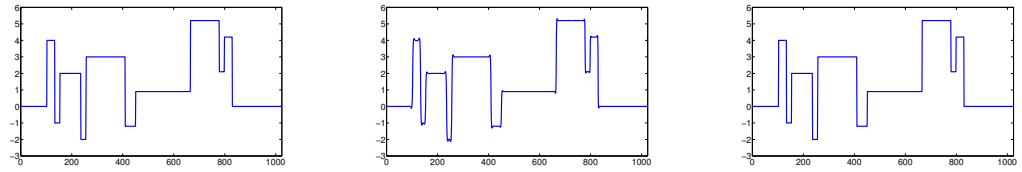
[1] Roughly speaking, a sparse approximate solution x to (1.1) means that x is sparse and $Ax \approx b$.

Table 4.8: Computational results on data from Sparco

Data	Size		PD			SPGL1		
	p	n	MSE	nnz	Time	MSE	nnz	Time
blkheavi	128	128	$1.28e-07$	12	0.1	$4.53e-03$	128	1.5
jitter	1000	200	$4.50e-08$	3	0.1	$1.38e-07$	28	0.1
gausspike	1024	256	$2.63e-07$	32	1.5	$1.09e-08$	143	0.2
sgnspike	2560	600	$3.84e-08$	20	0.2	$8.08e-08$	101	0.1
blknheavi	1024	1024	$3.19e-09$	12	0.4	$4.22e-03$	1024	14.3
cosspike	2048	1024	$8.97e-07$	121	0.3	$8.40e-08$	413	0.2
angiogram	10000	10000	$2.74e-06$	575	2.4	$5.77e-07$	1094	0.5
blurspike	16384	16384	$2.97e-03$	7906	10.1	$3.15e-03$	16384	20.6
srcsep1	57344	29166	$3.41e-05$	9736	743.6	$1.39e-08$	33887	102.5
srcsep2	86016	29166	$6.93e-04$	12485	1005.6	$2.80e-04$	52539	136.2



(a) Actual sparse solution (b) Sparse solution obtained by SPGL1 (c) Sparse Solution obtained by PD



(d) Original signal (e) Signal recovered by SPGL1 (f) Signal recovered by PD

Figure 4.3: Sparse recovery.

l_1 regularization problem:

$$\min_{x\in\Re^p} \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1, \tag{4.48}$$

where $\lambda \geq 0$ is a regularization parameter (see, for example, [61, 68, 79]). Our goal below is to apply the PD method studied in Subsection 4.3.1 to solve (1.2) directly and compare the results with one of the l_1 relaxation solvers GPSR [61] and the iterative hard-thresholding algorithm (IHT) [11, 12] which also solves (1.2) directly.

Clearly, problem (1.2) is in the form of (1.6) and thus the PD method proposed in Subsection 4.3 can be suitably applied to solve (1.2). The main computation effort of the PD method when applied to (1.2) 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 $\|y_0^0\|_0 \leq r$. Also, we set the initial penalty parameter $\varrho_0 = 1$ and the parameter $\sigma = \sqrt{10}$. In addition, we use

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

and

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

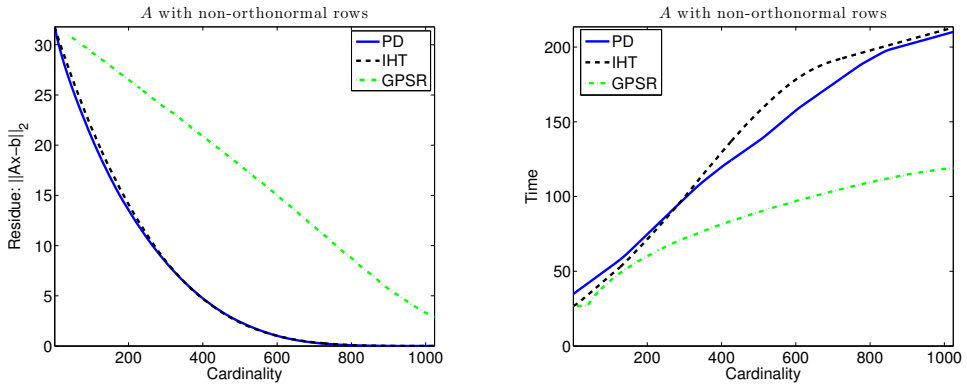
as the inner and outer termination criteria for the PD method, and set their associated accuracy parameters $\epsilon_I = 10^{-2}$ and $\epsilon_O = 10^{-3}$.

We next conduct numerical experiments to test the performance of our PD method for solving problem (1.2) on random data. We also compare the quality of the approximate solutions of (1.2) obtained by our PD method and the iterative hard-thresholding algorithm (IHT) [11, 12] with that of (4.48) found by a first-order solver GPSR [61]. 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 (4.48) 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 (1.2) 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 (1.2) (resp., (4.46)) 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 4.4. In detail, we plot the average residual $\|Ax - b\|_2$ against the cardinality in the left graph and the average accumulated CPU time [2] (in seconds) against the cardinality in the right graph. We observe that the residuals of the approximate solutions of (4.48) obtained by our PD method and IHT are almost equal and substantially smaller than that of (1.2) found by GPSR when the same cardinality is considered. In addition, we can see that GPSR is faster than the other two methods.

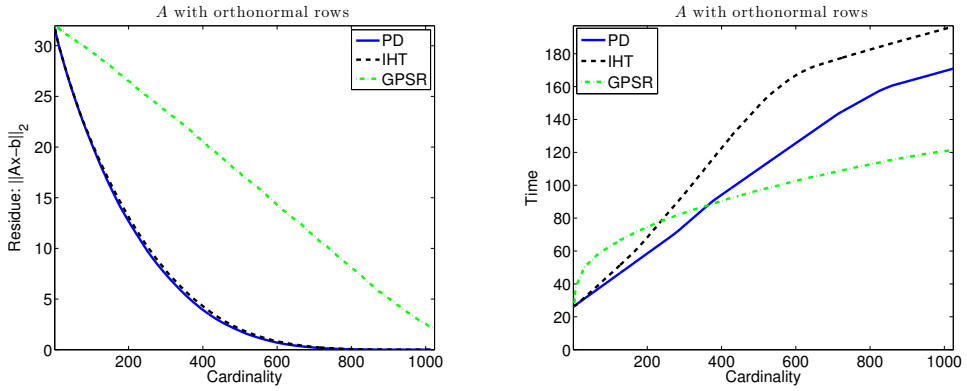
[2]For a cardinality r, the corresponding accumulated CPU time is the total CPU time used to compute approximate solutions of problem (1.2) or (4.46) with cardinality from 1 to r.

We also conduct a similar experiment as above except that A is randomly generated with orthonormal rows. The results are plotted in Figure 4.5. 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.



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

Figure 4.4: Trade-off curves.



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

Figure 4.5: Trade-off curves.

4.5 Concluding remarks

In this chapter 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 when solutions of same cardinality are sought, our approach applied to the l_0-based models generally has better solution quality and/or speed than the existing approaches that are applied to the corresponding l_1-based models.

We shall remark that the augmented Lagrangian decomposition methods can be developed for solving l_0 minimization problems (1.6) and (1.7) simply by replacing the quadratic penalty functions in the PD methods by augmented Lagrangian functions. Nevertheless, as observed in our experiments, their practical performance is generally worse than the PD methods.

Chapter 5

Wavelet Frame Based Image Restoration

The theory of (tight) wavelet frames has been extensively studied in the past twenty years and they are currently widely used for image restoration and other image processing and analysis problems. The success of wavelet frame based models, including balanced approach [28, 15] and analysis based approach [19, 55, 121], is due to their capability of sparsely approximating piecewise smooth functions like images. Motivated by the balanced approach and analysis based approach, we shall propose a wavelet frame based l_0 minimization model, where the l_0-"norm" of the frame coefficients is penalized. We adapt the penalty decomposition (PD) method of [89] to solve the proposed optimization problem. Some convergence analysis of the adapted PD method will also be provided. Numerical results showed that the proposed model solved by the PD method can generate images with better quality than those obtained by either analysis based approach or balanced approach in terms of restoring sharp features as well as maintaining smoothness of the recovered images.

This chapter is based on the paper [136] co-authored with Bin Dong and Zhaosong Lu.

5.1 Introduction to wavelet frame based image restoration

Mathematics has been playing an important role in the modern developments of image processing and analysis. Image restoration, including image denoising, deblurring, inpainting, tomography, etc., is one of the most important areas in image processing and analysis.

Its major purpose is to enhance the quality of a given image that is corrupted in various ways during the process of imaging, acquisition and communication; and enable us to see crucial but subtle objects residing in the image. Therefore, image restoration is an important step to take towards accurate interpretations of the physical world and making optimal decisions.

5.1.1 Image restoration

Image restoration is often formulated as a linear inverse problem. For the simplicity of the notations, we denote the images as vectors in $\mathbb{R}^n$ with n equals to the total number of pixels. A typical image restoration problem is formulated as

$$f = Au + \eta, \tag{5.1}$$

where $f \in \mathbb{R}^d$ is the observed image (or measurements), η denotes white Gaussian noise with variance σ^2, and $A \in \mathbb{R}^{d\times n}$ is some linear operator. The objective is to find the unknown true image $u \in \mathbb{R}^n$ from the observed image f. Typically, the linear operator in (5.1) is a convolution operator for image deconvolution problems, a projection operator for image inpainting and partial Radon transform for computed tomography.

To solve u from (5.1), one of the most natural choices is the following least square problem

$$\min_{u\in\mathbb{R}^n} \|Au - f\|_2^2,$$

where $\|\cdot\|_2$ denotes the l_2-norm. This is, however, not a good idea in general. Taking image deconvolution problem as an example, since the matrix A is ill-conditioned, the noise η possessed by f will be amplified after solving the above least squares problem. Therefore, in order to suppress the effect of noise and also preserve key features of the image, e.g., edges, various regularization based optimization models were proposed in the literature. Among all regularization based models for image restoration, variational methods and wavelet frames based approaches are widely adopted and have been proven successful.

The trend of variational methods and partial differential equation (PDE) based image processing started with the refined Rudin-Osher-Fatemi (ROF) model [112] which penalizes the total variation (TV) of u. Many of the current PDE based methods for image denoising and decomposition utilize TV regularization for its beneficial edge preserving property (see e.g., [93, 115, 100]). The ROF model is especially effective on restoring images that are

piecewise constant, e.g., binary images. Other types of variational models were also proposed after the ROF model. We refer the interested readers to [63, 27, 93, 100, 32, 3, 33, 133] and the references therein for more details.

Wavelet frame based approaches are relatively new and came from a different path. The basic idea for wavelet frame based approaches is that images can be sparsely approximated by properly designed wavelet frames, and hence, the regularization used for wavelet frame based models is the l_1-norm of frame coefficients. Although wavelet frame based approaches take similar forms as variational methods, they were generally considered as different approaches than variational methods because, among many other reasons, wavelet frame based approaches is defined for discrete data, while variational methods assume all variables are functions. This impression was changed by the recent paper [122, 17], where the authors established a rigorous connection between one of the wavelet frame based approaches, namely the analysis based approach, and variational models. It was shown in [17] that the analysis based approach can be regarded as a finite difference approximation of a certain type of general variational model, and such approximation will be exact when image resolution goes to infinity. Furthermore, the solutions of the analysis based approach also approximate, in some proper sense, the solutions of corresponding variational model. Such connections not only grant geometric interpretation to wavelet frame based approaches, but also lead to even wider applications of them, e.g., image segmentation [50] and 3D surface reconstruction from unorganized point sets [52]. On the other hand, the discretization provided by wavelet frames was shown, in e.g., [28, 30, 18, 19, 17, 51], to be superior than the standard discretizations for some of the variational models, due to the multiresolution structure and redundancy of wavelet frames which enable wavelet frame based models to adaptively choose a proper differential operators in different regions of a given image according to the order of the singularity of the underlying solutions. For these reasons, as well as the fact that digital images are always discrete, we use wavelet frames as the tool for image restoration in this chapter.

5.1.2 Wavelet frame based approaches

We now briefly introduce the concept of tight frames and tight wavelet frame, and then recall some of the frame based image restoration models. Interesting readers should consult [111, 44, 45] for theories of frames and wavelet frames, [116] for a short survey on theory and applications of frames, and [51] for a more detailed survey.

A countable set $X \subset L_2(\mathbb{R})$ is called a tight frame of $L_2(\mathbb{R})$ if

$$f = \sum_{h \in X} \langle f, h \rangle h \quad \forall f \in L_2(\mathbb{R}),$$

where $\langle \cdot, \cdot \rangle$ is the inner product of $L_2(\mathbb{R})$. The tight frame X is called a tight wavelet frame if the elements of X are generated by dilations and translations of finitely many functions called framelets. The construction of framelets can be obtained by the unitary extension principle (UEP) of [111]. In our implementations, we will mainly use the piecewise linear B-spline framelets constructed by [111]. Given a 1-dimensional framelet system for $L_2(\mathbb{R})$, the s-dimensional tight wavelet frame system for $L_2(\mathbb{R}^s)$ can be easily constructed by using tensor products of 1-dimensional framelets (see e.g., [44, 51]).

In the discrete setting, we will use $W \in \mathbb{R}^{m \times n}$ with $m \geq n$ to denote fast tensor product framelet decomposition and use W^T to denote the fast reconstruction. Then by the unitary extension principle [111], we have $W^T W = I$, i.e., $u = W^T W u$ for any image u. We will further denote an L-level framelet decomposition of u as

$$Wu = (\ldots, W_{l,j}u, \ldots)^T \quad \text{for } 0 \leq l \leq L-1, j \in \mathcal{I},$$

where $\mathcal{I}$ denotes the index set of all framelet bands and $W_{l,j}u \in \mathbb{R}^n$. Under such notation, we have $m = L \times |\mathcal{I}| \times n$. We will also use $\alpha \in \mathbb{R}^m$ to denote the frame coefficients, i.e., $\alpha = Wu$, where

$$\alpha = (\ldots, \alpha_{l,j}, \ldots)^T, \quad \text{with } \alpha_{l,j} = W_{l,j}u.$$

More details on discrete algorithms of framelet transforms can be found in [51].

Since tight wavelet frame systems are redundant systems (i.e., $m > n$), the representation of u in the frame domain is not unique. Therefore, there are mainly three formulations utilizing the sparseness of the frame coefficients, namely, analysis based approach, synthesis based approach, and balanced approach. Detailed and integrated descriptions of these three methods can be found in [51].

The wavelet frame based image processing started from [28, 29] for high-resolution image reconstructions, where the proposed algorithm was later analyzed in [15]. These work lead to the following balanced approach [16]

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{2}\|AW^T\alpha - f\|_D^2 + \frac{\kappa}{2}\|(I - WW^T)\alpha\|_2^2 + \left\| \sum_{l=0}^{L-1} \left(\sum_{j \in \mathcal{I}} \lambda_{l,j} |\alpha_{l,j}|^p \right)^{1/p} \right\|_1, \tag{5.2}$$

where $p = 1$ or 2, $0 \leq \kappa \leq \infty$, $\lambda_{l,j} \geq 0$ is a scalar parameter, and $\|\cdot\|_D$ denotes the weighted l_2-norm with D positive definite. This formulation is referred to as the balanced approach because it balances the sparsity of the frame coefficient and the smoothness of the image. The balanced approach (5.2) was applied to various applications in [26, 31, 118, 75].

When $\kappa = 0$, only the sparsity of the frame coefficient is penalized. This is called the synthesis based approach, as the image is synthesized by the sparsest coefficient vector (see e.g., [46, 56, 57, 59, 60]). When $\kappa = +\infty$, only the sparsity of canonical wavelet frame coefficients, which corresponds to the smoothness of the underlying image, is penalized. For this case, problem (5.2) can be rewritten as

$$\min_{u\in\mathbb{R}^n} \frac{1}{2}\|Au - f\|_D^2 + \left\| \sum_{l=0}^{L-1} \left(\sum_{j\in\mathcal{I}} \lambda_{l,j} |W_{l,j} u|^p \right)^{1/p} \right\|_1 . \tag{5.3}$$

This is called the analysis based approach, as the coefficient is in range of the analysis operator (see, for example, [19, 55, 121]).

Note that if we take $p = 1$ for the last term of (5.2) and (5.3), it is known as the anisotropic l_1-norm of the frame coefficients, which is the case used for earlier frame based image restoration models. The case $p = 2$, called isotropic l_1-norm of the frame coefficients, was proposed in [17] and was shown to be superior than anisotropic l_1-norm. Therefore, we will choose $p = 2$ for our simulations.

5.1.3 Motivations

For most of the variational models and wavelet frame based approaches, the choice of norm for the regularization term is the l_1-norm. Taking wavelet frame based approaches for example, the attempt of minimizing the l_1-norm of the frame coefficients is to increase their sparsity, which is the right thing to do since piecewise smooth functions like images can be sparsely approximated by tight wavelet frames. Although the l_1-norm of a vector does not directly correspond to its cardinality in contrast to l_0-"norm", it can be regarded as a convex approximation to l_0-"norm". Such approximation is also an excellent approximation for many cases. It was shown by [21], which generalizes the exciting results of compressed sensing [23, 25, 24, 53], that for a given wavelet frame, if the operator A satisfies certain conditions, and if the unknown true image can be sparsely approximated by the given wavelet frame, one can robustly recover the unknown image by penalizing the l_1-norm of the frame coefficients.

For image restoration, however, the conditions on A as required by [21] are not generally satisfied, which means penalizing l_0-"norm" and l_1-norm may produce different solutions. Although both the balanced approach (5.2) and analysis based approach (5.3) can generate restored images with very high quality, one natural question is whether using l_0-"norm" instead of l_1-norm can further improve the results.

On the other hand, it was observed, in e.g., [51] (also see Figure 5.3 and Figure 5.4), that balanced approach (5.2) generally generates images with sharper features like edges than the analysis based approach (5.3), because balanced approach emphasizes more on the sparsity of the frame coefficients. However, the recovered images from balanced approach usually contains more artifact (e.g., oscillations) than analysis based approach, because the regularization term of the analysis based approach has a direct link to the regularity of u (as proven by [17]) comparing to balanced approach. Although such trade-off can be controlled by the parameter κ in the balanced approach (5.2), it is not very easy to do in practice. Furthermore, when a large κ is chosen, some of the numerical algorithms solving (5.2) will converge slower than choosing a smaller κ (see e.g., [118, 51]).

Since penalizing l_1-norm of Wu ensures smoothness while not as much sparsity as balanced approach, we propose to penalize l_0-"norm" of Wu instead. Intuitively, this should provide us a balance between sharpness of the features and smoothness for the recovered images. The difficulty here is that l_0 minimization problems are generally hard to solve. Recently, penalty decomposition (PD) methods were proposed by [89] for a general l_0 minimization problem that can be used to solve our proposed model due to its generality. Computational results of [89] demonstrated that their methods generally outperform the existing methods for compressed sensing problems, sparse logistic regression and sparse inverse covariance selection problems in terms of quality of solutions and/or computational efficiency. This motivates us to adapt one of their PD methods to solve our proposed l_0 minimization problem. Same as proposed in [89], the block coordinate descent (BCD) method is used to solve each penalty subproblem of the PD method. However, the convergence analysis of the BCD method was missing from [89] when l_0-"norm" appears in the objective function. Indeed, the convergence of the BCD method generally requires the continuity of the objective function as discussed in [127]. In addition, the BCD method for the optimization problem with the nonconvex objective function has only been proved to converge to a stationary point which is not a local minimizer in general (see [127] for details).

We now leave the details of the model and algorithm to Subsection 5.2 and details of

simulations to Subsection 5.3.

5.2 Model and algorithm

We start by introducing some simple notations. The space of symmetric $n \times n$ matrices will be denoted by $\mathcal{S}^n$. If $X \in \mathcal{S}^n$ is positive definite, we write $X \succ 0$. We denote by I the identity matrix, whose dimension should be clear from the context. Given an index set $J \subseteq \{1, \ldots, n\}$, x_J denotes the sub-vector formed by the entries of x indexed by J. For any real vector, $\|\cdot\|_0$ and $\|\cdot\|_2$ denote the cardinality (i.e., the number of nonzero entries) and the Euclidean norm of the vector, respectively. In addition, $\|x\|_D$ denotes the weighted l_2-norm defined by $\|x\|_D = \sqrt{x^T D x}$ with $D \succ 0$.

5.2.1 Model

We now propose the following optimization model for image restoration problems,

$$\min_{u \in \mathcal{Y}} \frac{1}{2}\|Au - f\|_D^2 + \sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}} \|(Wu)_{\boldsymbol{i}}\|_0, \tag{5.4}$$

where $\mathcal{Y}$ is some convex subset of $\mathbb{R}^n$. Here we are using the multi-index $\boldsymbol{i}$ and denote $(Wu)_{\boldsymbol{i}}$ (similarly for $\lambda_{\boldsymbol{i}}$) the value of Wu at a given pixel location within a certain level and band of wavelet frame transform. Comparing to the analysis based model, we are now penalizing the number of nonzero elements of Wu. As mentioned earlier that if we emphasize too much on the sparsity of the frame coefficients as in the balanced approach or synthesis based approach, the recovered image will contain artifacts, although features like edges will be sharp; if we emphasize too much on the regularity of u like in analysis based approach, features in the recovered images will be slightly blurred, although artifacts and noise will be nicely suppressed. Therefore, by penalizing the l_0-"norm" of Wu as in (5.4), we can indeed achieve a better balance between sharpness of features and smoothness of the recovered images.

Given that the l_0-"norm" is an integer-valued, discontinuous and nonconvex function, problem (5.6) is generally hard to solve. Some algorithms proposed in the literature, e.g., iterative hard thresholding algorithms [11, 12, 72], cannot be directly applied to the proposed model (5.4) unless $W = I$. Recently, Lu and Zhang [89] proposed a penalty decomposition (PD) method which has also been introduced in Chapter 4 to solve the following general l_0

minimization problem:

$$\min_{x\in\mathcal{X}} f(x) + \nu\|x_J\|_0 \tag{5.5}$$

for some $\nu > 0$ controlling the sparsity of the solution, where $\mathcal{X}$ is a closed convex set in $\mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function, and $\|x_J\|_0$ denotes the cardinality of the subvector formed by the entries of x indexed by J. In view of [89], we reformulate (5.4) as

$$\min_{u\in\mathcal{Y},\alpha=Wu} \frac{1}{2}\|Au - f\|_D^2 + \sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}}\|\alpha_{\boldsymbol{i}}\|_0 \tag{5.6}$$

and then we can adapt the PD method of [89] to tackle problem (5.4) directly. Same as proposed in [89], the BCD method is used to solve each penalty subproblem of the PD method. In addition, we apply the non-monotone gradient projection method proposed in [9] to solve one of the subproblem in the BCD method.

5.2.2 Algorithm for Problem (5.6)

In this subsection, we discuss how the PD method proposed in [89] solving (5.5) can be adapted to solve problem (5.6). Letting $x = (u_1, \ldots, u_n, \alpha_1, \ldots, \alpha_m)$, $J = \{n+1, \ldots, n+m\}$, $\bar{J} = \{1, \ldots, n\}$, $f(x) = \frac{1}{2}\|Ax_{\bar{J}} - f\|_D^2$ and $\mathcal{X} = \{x \in \mathbb{R}^{n+m} : x_J = Wx_{\bar{J}} \text{ and } x_{\bar{J}} \in \mathcal{Y}\}$, we can clearly see that the problem (5.6) takes the same form as (5.5). In addition, there obviously exists a feasible point $(u^{\text{feas}}, \alpha^{\text{feas}})$ for problem (5.6) when $\mathcal{Y} \neq \emptyset$, i.e. there exist $(u^{\text{feas}}, \alpha^{\text{feas}})$ such that $Wu^{\text{feas}} = \alpha^{\text{feas}}$ and $u^{\text{feas}} \in \mathcal{Y}$. In particular, one can choose u^{feas} and α^{feas} both to be zero vectors when applying the PD method studied in [89] to solve the problem (5.6). We now discuss the implementation details of the PD method when solving the proposed wavelet frame based model (5.6).

Given a penalty parameter $\varrho > 0$, the associated quadratic penalty function for (5.6) is defined as

$$p_\varrho(u, \alpha) := \frac{1}{2}\|Au - f\|_D^2 + \sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}}\|\alpha_{\boldsymbol{i}}\|_0 + \frac{\varrho}{2}\|Wu - \alpha\|_2^2. \tag{5.7}$$

Then we have the following PD method for problem (5.6) where each penalty subproblem is approximately solved by a BCD method (see [89] for details).

Penalty Decomposition (PD) Method for (5.6):

Let $\varrho_0 > 0$, $\delta > 1$ be given. Choose an arbitrary $\alpha^{0,0} \in \mathbb{R}^m$ and a constant Υ such that $\Upsilon \geq \max\{\frac{1}{2}\|Au^{\text{feas}} - f\|_D^2 + \sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}}\|\alpha_{\boldsymbol{i}}^{\text{feas}}\|_0, \min_{u\in\mathcal{Y}} p_{\varrho_0}(u, \alpha^{0,0})\}$. Set $k = 0$.

1) Set $q = 0$ and apply the BCD method to find an approximate solution $(u^k, \alpha^k) \in \mathcal{Y} \times \mathbb{R}^m$ for the penalty subproblem

$$\min\{p_{\varrho_k}(u, \alpha) : \ u \in \mathcal{Y}, \ \alpha \in \mathbb{R}^m\} \tag{5.8}$$

by performing steps 1a)-1d):

1a) Solve $u^{k,q+1} \in \operatorname{Arg}\min_{u \in \mathcal{Y}} p_{\varrho_k}(u, \alpha^{k,q})$.

1b) Solve $\alpha^{k,q+1} \in \operatorname{Arg}\min_{\alpha \in \mathbb{R}^n} p_{\varrho_k}(u^{k,q+1}, \alpha)$.

1c) If $(u^{k,q+1}, \alpha^{k,q+1})$ satisfies the stopping criteria of the BCD method, set

$$(u^k, \alpha^k) := (u^{k,q+1}, \alpha^{k,q+1})$$

and go to step 2).

1d) Otherwise, set $q \leftarrow q + 1$ and go to step 1a).

2) If (u^k, α^k) satisfies the stopping criteria of the PD method, stop and output u^k. Otherwise, set$\varrho_{k+1} := \delta\varrho_k$.

3) If $\min_{u \in \mathcal{Y}} p_{\varrho_{k+1}}(u, \alpha^k) > \Upsilon$, set $\alpha^{k+1,0} := \alpha^{\text{feas}}$. Otherwise, set $\alpha^{k+1,0} := \alpha^k$.

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

end

Remark. In the practical implementation, we terminate the inner iterations of the BCD method based on the relative progress of $p_{\varrho_k}(u^{k,q}, \alpha^{k,q})$ which can be described as follows:

$$\frac{|p_{\varrho_k}(u^{k,q}, \alpha^{k,q}) - p_{\varrho_k}(u^{k,q+1}, \alpha^{k,q+1})|}{\max(|p_{\varrho_k}(u^{k,q+1}, \alpha^{k,q+1})|, 1)} \ \leq \ \epsilon_I.$$

Moreover, we terminate the outer iterations of the PD method once

$$\frac{\|Wu^k - \alpha^k\|_2}{\max(|p_{\varrho_k}(u^k, \alpha^k)|, 1)} \ \leq \ \epsilon_O.$$

■

Next we discuss how to solve two subproblems arising in step 1a) and 1b) of the BCD method.

The BCD subproblem in step 1a)

The BCD subproblem in step 1a) is in the form of

$$\min_{u\in\mathcal{Y}} \frac{1}{2}\langle u, Qu\rangle - \langle c, u\rangle \tag{5.9}$$

for some $Q \succ 0$ and $c \in \mathbb{R}^n$. Obviously, when $\mathcal{Y} = \mathbb{R}^n$, problem (5.9) is an unconstrained quadratic programming problem that can be solved by the conjugate gradient method. Nevertheless, the pixel values of an image are usually bounded. For example, the pixel values of a CT image should be always greater than or equal to zero and the pixel values of a grayscale image is between $[0, 255]$. Then the corresponding $\mathcal{Y}$ of these two examples are $\mathcal{Y} = \{x \in \mathbb{R}^n : x_i \geq lb \ \forall i = 1, \ldots, n\}$ with $lb = 0$ and $\mathcal{Y} = \{x \in \mathbb{R}^n : lb \leq x_i \leq ub \ \forall i = 1, \ldots, n\}$ with $lb = 0$ and $ub = 255$. To solve these types of the constrained quadratic programming problems, we apply the nonmonotone projected gradient method proposed in [9] and terminate it using the duality gap and dual feasibility conditions (if necessary).

For $\mathcal{Y} = \{x \in \mathbb{R}^n : x_i \geq lb \ \forall i = 1, \ldots, n\}$, given a Lagrangian multiplier $\beta \in \mathbb{R}^n$, the associated Lagrangian dual function of (5.9) can be written as:

$$L(u, \beta) = w(u) + \beta^T(lb - u),$$

where $w(u) = \frac{1}{2}\langle u, Qu\rangle - \langle c, u\rangle$. Based on the Karush-Kuhn-Tucker (KKT) conditions, for an optimal solution u^* of (5.9), there exists a Lagrangian multiplier β^* such that

$$\begin{aligned} &Qu^* - c - \beta^* = 0, \\ &\beta_i^* \geq 0 \ \forall i = 1, \ldots, n, \\ &(lb - u_i^*)\beta_i^* = 0 \ \forall i = 1, \ldots, n. \end{aligned}$$

Then at the sth iteration of the projected gradient method, we let $\beta^s = Qu^s - c$. As $\{u^s\}$ approaches the solution u^* of (5.9), $\{\beta^s\}$ approaches the Lagrangian multiplier β^* and the corresponding duality gap at each iteration is given by $\sum_{i=1}^n \beta_i^s(lb - u_i^s)$. Therefore, we terminate the projected gradient method when

$$\frac{|\sum_{i=1}^n \beta_i^s(lb - u_i^s)|}{\max(|w(u^s)|, 1)} \leq \epsilon_D \text{ and } \frac{-\min(\beta^s, 0)}{\max(\|\beta^s\|_2, 1)} \leq \epsilon_F$$

for some tolerances $\epsilon_D, \epsilon_F > 0$.

For $\mathcal{Y} = \{x \in \mathbb{R}^n : lb \leq x_i \leq ub \ \forall i = 1, \ldots, n\}$, given Lagrangian multipliers $\beta, \gamma \in \mathbb{R}^n$, the associated Lagrangian function of (5.9) can be written as:

$$L(u, \beta, \gamma) = w(u) + \beta^T(lb - u) + \gamma^T(u - ub),$$

where $w(u)$ is defined as above. Based on the KKT conditions, for an optimal solution u^* of (5.9), there exist Lagrangian multipliers β^* and γ^* such that

$$\begin{aligned}
&Qu^* - c - \beta^* + \gamma^* = 0,\\
&\beta_i^* \geq 0 \ \forall i = 1, \ldots, n,\\
&\gamma_i^* \geq 0 \ \forall i = 1, \ldots, n,\\
&(lb - u_i^*)\beta_i^* = 0 \ \forall i = 1, \ldots, n,\\
&(u_i^* - ub)\gamma_i^* = 0 \ \forall i = 1, \ldots, n.
\end{aligned}$$

Then at the sth iteration of the projected gradient method, we let $\beta^s = \max(Qu^s - c, 0)$ and $\gamma^s = -\min(Qu^s - c, 0)$. As $\{u^s\}$ approaches the solution u^* of (5.9), $\{\beta^s\}$ and $\{\gamma^s\}$ approach Lagrangian multipliers β^* and γ^*. In addition, the corresponding duality gap at each iteration is given by $\sum_{i=1}^n (\beta_i^s(lb - u_i^s) + \gamma_i^s(u_i^s - ub))$ and the duality feasibility is automatically satisfied. Therefore, we terminate the projected gradient method when

$$\frac{|\sum_{i=1}^n (\beta_i^s(lb - u_i^s) + \gamma_i^s(u_i^s - ub))|}{\max(|w(u^s)|, 1)} \leq \epsilon_D$$

for some tolerance $\epsilon_D > 0$.

The BCD subproblem in step 1b)

For $\lambda_{\boldsymbol{i}} \geq 0$, $\varrho > 0$ and $c \in \mathbb{R}^m$, the BCD subproblem in step 1b) is in the form of

$$\min_{\alpha \in \mathbb{R}^m} \sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}} \|\alpha_{\boldsymbol{i}}\|_0 + \frac{\varrho}{2} \sum_{\boldsymbol{i}} (\alpha_{\boldsymbol{i}} - c_{\boldsymbol{i}})^2.$$

By [89, Proposition 2.2] (see also [2, 11] for example), the solution of the above subproblem is in the following set:

$$\alpha^* \in H_{\tilde{\lambda}}(c) \quad \text{with } \tilde{\lambda}_{\boldsymbol{i}} := \sqrt{\frac{2\lambda_{\boldsymbol{i}}}{\varrho}} \text{ for all } \boldsymbol{i}, \tag{5.10}$$

where $H_\gamma(\cdot)$ denotes a component-wise hard thresholding operator with threshold γ:

$$[H_\gamma(x)]_{\boldsymbol{i}} = \begin{cases} 0 & \text{if } |x_{\boldsymbol{i}}| < \gamma_{\boldsymbol{i}}, \\ \{0, x_{\boldsymbol{i}}\} & \text{if } |x_{\boldsymbol{i}}| = \gamma_{\boldsymbol{i}}, \\ x_{\boldsymbol{i}} & \text{if } |x_{\boldsymbol{i}}| > \gamma_{\boldsymbol{i}}. \end{cases} \tag{5.11}$$

Note that H_γ is defined as a set-valued mapping [110, Chapter 5] which is different (only when $|x_{\boldsymbol{i}}| = \gamma_{\boldsymbol{i}}$) from the conventional definition of hard thresholding operator.

5.2.3 Convergence of the BCD method

In this subsection, we establish some convergence results regarding the inner iterations, i.e., Step 1), of the PD method. In particular, we will show that the fixed point of the BCD method is a local minimizer of (5.8). Moreover, under certain conditions, we prove that the sequence $\{(u^{k,q}, \alpha^{k,q})\}$ generated by the BCD method converges and the limit is a local minimizer of (5.8).

For convenience of presentation, we omit the index k from (5.8) and consider the BCD method for solving the following problem:

$$\min\{p_\varrho(u,\alpha):\ u\in\mathcal{Y},\ \alpha\in\mathbb{R}^m\}. \tag{5.12}$$

Without loss of generality, we assume that $D = I$. We now relabel and simplify the BCD method described in step 1a)-1c) in the PD method as follows.

$$\begin{cases} u^{q+1} = \arg\min_{u\in\mathcal{Y}}\ \frac{1}{2}\|Au-f\|_2^2 + \frac{\varrho}{2}\|Wu-\alpha^q\|_2^2, \\ \alpha^{q+1} \in \operatorname{Arg\,min}_{\alpha}\ \sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}}\|\alpha_{\boldsymbol{i}}\|_0 + \frac{\varrho}{2}\|\alpha - Wu^{q+1}\|_2^2. \end{cases} \tag{5.13}$$

We first show that the fixed point of the above BCD method is a local minimizer of (5.8).

Theorem 5.2.1 *Given a fixed point of the BCD method* (5.13)*, denoted as* (u^*, α^*)*, then* (u^*, α^*) *is a local minimizer of* $p_\varrho(u, \alpha)$*.*

Proof. We first note that the first subproblem of (5.13) gives us

$$\delta\langle A^T(Au^*-f) + \varrho W^T(Wu^*-\alpha^*), v-u^*\rangle \geq 0 \quad \text{for all } v\in\mathcal{Y}. \tag{5.14}$$

By applying (5.10), the second subproblem of (5.13) leads to:

$$\alpha^* \in H_{\tilde{\lambda}}(Wu^*). \tag{5.15}$$

Define index sets

$$\Gamma_0 := \{\boldsymbol{i} : \alpha^*_{\boldsymbol{i}} = 0\} \quad \text{and} \quad \Gamma_1 := \{\boldsymbol{i} : \alpha^*_{\boldsymbol{i}} \neq 0\}.$$

It then follows from (5.15) and (5.11) that

$$\begin{cases} |(Wu^*)_{\boldsymbol{i}}| \leq \tilde{\lambda}_{\boldsymbol{i}} & \text{for } \boldsymbol{i}\in\Gamma_0 \\ (Wu^*)_{\boldsymbol{i}} = \alpha^*_{\boldsymbol{i}} & \text{for } \boldsymbol{i}\in\Gamma_1, \end{cases} \tag{5.16}$$

where $(Wu^*)_{\boldsymbol{i}}$ denotes $\boldsymbol{i}$th entry of Wu^*.

Consider a small deformation vector $(\delta h, \delta g)$ such that $u^* + \delta h \in \mathcal{Y}$. Using (5.14), we have

$$\begin{aligned}
p_\varrho(u^*+\delta h, \alpha^*+\delta g) &= \frac{1}{2}\|Au^* + A\delta h - f\|_2^2 + \sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}} \|(\alpha^* + \delta g)_{\boldsymbol{i}}\|_0 \\
&\quad + \frac{\varrho}{2}\|\alpha^* + \delta g - W(u^* + \delta h)\|_2^2 \\
&= \frac{1}{2}\|Au^* - f\|_2^2 + \langle A\delta h, Au^* - f\rangle + \frac{1}{2}\|A\delta h\|_2^2 + \sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}} \|(\alpha^* + \delta g)_{\boldsymbol{i}}\|_0 \\
&\quad + \frac{\varrho}{2}\|\alpha^* - Wu^*\|_2^2 + \varrho\langle \alpha^* - Wu^*, \delta g - W\delta h\rangle + \frac{\varrho}{2}\|\delta g - W\delta h\|_2^2 \\
&= \frac{1}{2}\|Au^* - f\|_2^2 + \sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}} \|(\alpha^* + \delta g)_{\boldsymbol{i}}\|_0 + \frac{\varrho}{2}\|\alpha^* - Wu^*\|_2^2 + \frac{1}{2}\|A\delta h\|_2^2 \\
&\quad + \langle \delta h, A^T(Au^* - f) + \varrho W^T(Wu^* - \alpha^*)\rangle + \varrho\langle \delta g, \alpha^* - Wu^*\rangle \\
&\quad + \frac{\varrho}{2}\|\delta g - W\delta h\|_2^2 \\
&\geq \frac{1}{2}\|Au^* - f\|_2^2 + \sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}} \|(\alpha^* + \delta g)_{\boldsymbol{i}}\|_0 + \frac{\varrho}{2}\|\alpha^* - Wu^*\|_2^2 \\
&\quad + \langle \delta h, A^T(Au^* - f) + \varrho W^T(Wu^* - \alpha^*)\rangle + \varrho\langle \delta g, \alpha^* - Wu^*\rangle \\
\text{(By (5.14))} \quad &\geq \frac{1}{2}\|Au^* - f\|_2^2 + \sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}} \|(\alpha^* + \delta g)_{\boldsymbol{i}}\|_0 + \frac{\varrho}{2}\|\alpha^* - Wu^*\|_2^2 \\
&\quad + \varrho\langle \delta g, \alpha^* - Wu^*\rangle \\
&= \frac{1}{2}\|Au^* - f\|_2^2 + \frac{\varrho}{2}\|\alpha^* - Wu^*\|_2^2 \\
&\quad + \sum_{\boldsymbol{i}} \Big(\lambda_{\boldsymbol{i}} \|\alpha_{\boldsymbol{i}}^* + \delta g_{\boldsymbol{i}}\|_0 + \varrho \delta g_{\boldsymbol{i}}(\alpha_{\boldsymbol{i}}^* - (Wu^*)_{\boldsymbol{i}})\Big).
\end{aligned}$$

Splitting the summation in the last equation with respect to index sets Γ_0 and Γ_1 and using (5.16), we have

$$\begin{aligned}
p_\varrho(u^*+\delta h, \alpha^*+\delta g) &\geq \frac{1}{2}\|Au^* - f\|_2^2 + \frac{\varrho}{2}\|\alpha^* - Wu^*\|_2^2 + \sum_{\boldsymbol{i}\in\Gamma_0} \Big(\lambda_{\boldsymbol{i}} \|\delta g_{\boldsymbol{i}}\|_0 - \varrho \delta g_{\boldsymbol{i}}(Wu^*)_{\boldsymbol{i}}\Big) \\
&\quad + \sum_{\boldsymbol{i}\in\Gamma_1} \lambda_{\boldsymbol{i}} \|\alpha_{\boldsymbol{i}}^* + \delta g_{\boldsymbol{i}}\|_0.
\end{aligned}$$

Notice that when $|\delta g_{\boldsymbol{i}}|$ is small enough, we then have

$$\|\alpha_{\boldsymbol{i}}^* + \delta g_{\boldsymbol{i}}\|_0 = \|\alpha_{\boldsymbol{i}}^*\|_0 \quad \text{for } \boldsymbol{i} \in \Gamma_1.$$

Therefore, we have

$$\begin{aligned} p_\varrho(u^*+\delta h,\alpha^*+\delta g) &\geq \frac{1}{2}\|Au^*-f\|_2^2+\frac{\varrho}{2}\|\alpha^*-Wu^*\|_2^2 \\ &\quad +\sum_{\boldsymbol{i}\in\Gamma_0}\Big(\lambda_{\boldsymbol{i}}\|\delta g_{\boldsymbol{i}}\|_0-\varrho\delta g_{\boldsymbol{i}}(Wu^*)_{\boldsymbol{i}}\Big)+\sum_{\boldsymbol{i}\in\Gamma_1}\lambda_{\boldsymbol{i}}\|\alpha^*_{\boldsymbol{i}}\|_0 \\ &= p_\varrho(u^*,\alpha^*)+\sum_{\boldsymbol{i}\in\Gamma_0}\Big(\lambda_{\boldsymbol{i}}\|\delta g_{\boldsymbol{i}}\|_0-\varrho\delta g_{\boldsymbol{i}}(Wu^*)_{\boldsymbol{i}}\Big). \end{aligned}$$

We now show that, for $\boldsymbol{i}\in\Gamma_0$ and $\|\delta g\|$ small enough,

$$\lambda_{\boldsymbol{i}}\|\delta g_{\boldsymbol{i}}\|_0-\varrho\delta g_{\boldsymbol{i}}(Wu^*)_{\boldsymbol{i}}\geq 0. \tag{5.17}$$

For the indices $\boldsymbol{i}$ such that $\lambda_{\boldsymbol{i}}=0$, first inequality of (5.16) implies that $(Wu^*)_{\boldsymbol{i}}=0$ and hence (5.17) holds. Therefore, we only need to consider indices $\boldsymbol{i}\in\Gamma_0$ such that $\lambda_{\boldsymbol{i}}\neq 0$. Then obviously as long as $|\delta g_{\boldsymbol{i}}|\leq\frac{\lambda_{\boldsymbol{i}}}{\varrho|(Wu^*)_{\boldsymbol{i}}|}$, we will have (5.17) hold. We now conclude that there exists $\varepsilon>0$ such that for all $(\delta h,\delta g)$ satisfying $\max(\|\delta h\|_\infty,\|\delta g\|_\infty)<\varepsilon$, we have $p_\varrho(u^*+\delta h,\alpha^*+\delta g)\geq p_\varrho(u^*,\alpha^*)$. ∎

We next show that under some suitable assumptions, the sequence $\{(u^q,\alpha^q)\}$ generated by (5.13) converges to a fixed point of the BCD method.

Theorem 5.2.2 *Assume that $\mathcal{Y}=\mathbb{R}^n$ and $A^TA\succ 0$. Let $\{(u^q,\alpha^q)\}$ be the sequence generated by the BCD method described in* (5.13)*. Then, the sequence $\{(u^q,\alpha^q)\}$ is bounded. Furthermore, any limit point of the sequence $\{(u^q,\alpha^q)\}$ is a fixed point of* (5.13)*.*

Proof. In view of $\mathcal{Y}=\mathbb{R}^n$ and the optimality condition of the first subproblem of (5.13), one can see that

$$u^{q+1}=(A^TA+\varrho I)^{-1}A^Tf+\varrho(A^TA+\varrho I)^{-1}W^T\alpha^q. \tag{5.18}$$

Let $x:=(A^TA+\varrho I)^{-1}A^Tf$, $P:=\varrho(A^TA+\varrho I)^{-1}$, equation (5.18) can be rewritten as

$$u^{q+1}=x+PW^T\alpha^q. \tag{5.19}$$

Moreover, by the assumption $A^TA\succ 0$, we have $0\prec P\prec I$.

Using (5.19) and (5.11), we observe from the second subproblem of (5.13) that

$$\alpha^{q+1}\in H_{\tilde{\lambda}}(Wu^{q+1})=H_{\tilde{\lambda}}\left(Wx+WPW^T\alpha^q\right). \tag{5.20}$$

Let $Q := I - WPW^T$, then (5.20) can be rewritten as

$$\alpha^{q+1} \in H_{\tilde{\lambda}}\left(\alpha^q + Wx - Q\alpha^q\right). \tag{5.21}$$

In addition, from $W^TW = I$ we can easily show that $0 \prec Q \preceq I$.

Let $F(\alpha,\beta) := \frac{1}{2}\langle\alpha, Q\alpha\rangle - \langle Wx,\alpha\rangle + \sum_{\boldsymbol{i}} \bar{\lambda}_{\boldsymbol{i}}\|\alpha_{\boldsymbol{i}}\|_0 - \frac{1}{2}\langle\alpha-\beta, Q(\alpha-\beta)\rangle + \frac{1}{2}\|\alpha-\beta\|_2^2$ where $\bar{\lambda} = \frac{\lambda}{\rho}$. Then we have

$$\text{Argmin}_\alpha F(\alpha,\alpha^q) = \text{Argmin}_\alpha \frac{1}{2}\|\alpha - (\alpha^q + Wx - Q\alpha^q)\|_2^2 + \sum_{\boldsymbol{i}} \bar{\lambda}_{\boldsymbol{i}}\|\alpha_{\boldsymbol{i}}\|_0. \tag{5.22}$$

In view of equation (5.21) and (5.22) and the definition of the hard thresholding operator, we can easily observe that $\alpha^{q+1} \in \text{Argmin}_\alpha F(\alpha,\alpha^q)$. By following similar arguments as in [11, Lemma 1, Lemma D.1], we have

$$\begin{aligned} F(\alpha^{q+1},\alpha^{q+1}) &\leq F(\alpha^{q+1},\alpha^{q+1}) + \frac{1}{2}\|\alpha^{q+1}-\alpha^q\|_2^2 - \frac{1}{2}\langle\alpha^{q+1}-\alpha^q, Q(\alpha^{q+1}-\alpha^q)\rangle \\ &= F(\alpha^{q+1},\alpha^q) \\ &\leq F(\alpha^q,\alpha^q), \end{aligned}$$

which leads to

$$\|\alpha^{q+1}-\alpha^q\|_2^2 - \langle\alpha^{q+1}-\alpha^q, Q(\alpha^{q+1}-\alpha^q)\rangle \leq 2F(\alpha^q,\alpha^q) - 2F(\alpha^{q+1},\alpha^{q+1}).$$

Since $P \succ 0$, we have

$$\begin{aligned} \|W^T(\alpha^{q+1}-\alpha^q)\|_2^2 &\leq \frac{1}{C_1}\langle W^T(\alpha^{q+1}-\alpha^q), PW^T(\alpha^{q+1}-\alpha^q)\rangle \\ &= \frac{1}{C_1}\langle\alpha^{q+1}-\alpha^q, (I-Q)(\alpha^{q+1}-\alpha^q)\rangle \\ &= \frac{1}{C_1}\left(\|\alpha^{q+1}-\alpha^q\|_2^2 - \langle\alpha^{q+1}-\alpha^q, Q(\alpha^{q+1}-\alpha^q)\rangle\right) \\ &\leq \frac{2}{C_1}F(\alpha^q,\alpha^q) - \frac{2}{C_1}F(\alpha^{q+1},\alpha^{q+1}) \end{aligned}$$

for some $C_1 > 0$. Telescoping on the above inequality and using the fact that $\sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}}\|\alpha_{\boldsymbol{i}}\|_0 \geq 0$, we have

$$\begin{aligned} \sum_{q=0}^{N} \|W^T(\alpha^{q+1}-\alpha^q)\|_2^2 &\leq \frac{2}{C_1}F(\alpha^0,\alpha^0) - \frac{2}{C_1}F(\alpha^{N+1},\alpha^{N+1}) \\ &\leq \frac{2}{C_1}\left(F(\alpha^0,\alpha^0) - (\frac{1}{2}\langle\alpha^{N+1}, Q\alpha^{N+1}\rangle - \langle Wx,\alpha^{N+1}\rangle)\right) \\ &\leq \frac{2}{C_1}\left(F(\alpha^0,\alpha^0) - K\right), \end{aligned}$$

where K is the optimal value of $\min_y\{\frac{1}{2}\langle y, Qy\rangle - \langle Wx, y\rangle\}$. Since $Q \succ 0$, we have $K > -\infty$. Then the last inequality implies that $\lim_{q\to\infty} \|W^T(\alpha^{q+1} - \alpha^q)\|_2 \to 0$.

By using (5.19) and $P \prec I$, we see that

$$\begin{aligned}
\|u^{q+1} - W^T\alpha^{q+1}\|_2 &= \|x + PW^T\alpha^q - W^T\alpha^{q+1} + W^T\alpha^q - W^T\alpha^q\|_2 \\
&= \|x + (P-I)W^T\alpha^q - W^T(\alpha^{q+1} - \alpha^q)\|_2 \\
&\geq \|x + (P-I)W^T\alpha^q\|_2 - \|W^T(\alpha^{q+1} - \alpha^q)\|_2 \\
&= \|(I-P)W^T\alpha^q - x\|_2 - \|W^T(\alpha^{q+1} - \alpha^q)\|_2 \\
&\geq \|(I-P)W^T\alpha^q\|_2 - \|x\|_2 - \|W^T(\alpha^{q+1} - \alpha^q)\|_2 \\
&\geq C_2\|W^T\alpha^q\|_2 - \|x\|_2 - \|W^T(\alpha^{q+1} - \alpha^q)\|_2
\end{aligned}$$

for some $C_2 > 0$. Then by rearranging the above inequality and using the fact $W^TW = I$, we have

$$\begin{aligned}
\|W^T\alpha^q\|_2 &\leq \frac{1}{C_2}(\|u^{q+1} - W^T\alpha^{q+1}\|_2 + \|x\|_2 + \|W^T(\alpha^{q+1} - \alpha^q)\|_2) \\
&= \frac{1}{C_2}(\|W^T(Wu^{q+1} - \alpha^{q+1})\|_2 + \|x\|_2 + \|W^T(\alpha^{q+1} - \alpha^q)\|_2) \\
&\leq \frac{1}{C_2}(\|Wu^{q+1} - \alpha^{q+1}\|_2 + \|x\|_2 + \|W^T(\alpha^{q+1} - \alpha^q)\|_2).
\end{aligned}$$

By the definition of the hard thresholding operator and (5.20), we can easily see that $\|Wu^{q+1} - \alpha^{q+1}\|_2$ is bounded. In addition, notice that $\|x\|_2$ is a constant and $\lim_{q\to\infty} \|W^T(\alpha^{q+1} - \alpha^q)\|_2 \to 0$. Thus $\|W^T\alpha^q\|_2$ is also bounded. By using (5.19) and the definition of the hard thresholding operator again, we can immediately see that both $\{u^{q+1}\}$ and $\{\alpha^{q+1}\}$ are bounded as well.

Suppose that (u^*, α^*) is a limit point of the sequence $\{(u^q, \alpha^q)\}$. Therefore, there exists a subsequence $\{(u^{q'}, \alpha^{q'})\}_{q'\in K}$ converging to (u^*, α^*) where the index set K contains all the indices of the sequence $\{(u^q, \alpha^q)\}$. Using (5.20) and the definition of the hard thresholding operator, we can observe that

$$\alpha^* = \lim_{q'\in K\to\infty} \alpha^{q'+1} \in H_{\tilde{\lambda}}(\lim_{q'\in K\to\infty} Wu^{q'+1}) = H_{\tilde{\lambda}}(Wu^*).$$

In addition, it follows from (5.18) that

$$u^* = (A^TA + \varrho I)^{-1}A^Tf + \varrho(A^TA + \varrho I)^{-1}W^T\alpha^*.$$

In view of the above two relations, one can immediately conclude that $\{(u^*, \alpha^*)\}$ is a fixed point of (5.13). ∎

In the view of Theorems 5.2.1, 5.2.2 and under some suitable assumptions, we can easily observe the following convergence of the BCD method.

Theorem 5.2.3 *Assume that $\mathcal{Y} = \mathbb{R}^n$ and $A^T A \succ 0$. Then, the sequence $\{(u^q, \alpha^q)\}$ defined by the BCD method in* (5.13) *is bounded. Furthermore, any limit point of the sequence $\{(u^q, \alpha^q)\}$ is a local minimizer of* (5.12).

For the PD method itself, similar arguments as in the proof of [89, Theorem 3.2] will lead to that every accumulation point of the sequence $\{(u^k, \alpha^k)\}$ is a feasible point of (5.6). Although it is not clear whether the accumulation point is a local minimizer of (5.6), our numerical results show that the solutions obtained by the PD method are superior than those obtained by the balanced approach and the analysis based approach.

5.3 Numerical results

In this subsection, we conduct numerical experiments to test the performance of the PD method for problem (5.6) presented in Section 5.2 and compare the results with the balanced approach (5.2) and the analysis based approach (5.3). We use the accelerated proximal gradient (APG) algorithm [118] (see also [6]) to solve the balanced approach; and we use the split Bregman algorithm [64, 19] to solve the analysis based approach.

For APG algorithm that solves balanced approach (5.2), we shall adopt the following stopping criteria:

$$\min\left\{\frac{\|\alpha^k - \alpha^{k-1}\|_2}{\max\{1, \|\alpha^k\|_2\}}, \frac{\|AW^T\alpha^k - f\|_D}{\|f\|_2}\right\} \leq \epsilon_P.$$

For split Bregman algorithm that solves the analysis based approach (5.3), we shall use the following stopping criteria:

$$\frac{\|Wu^{k+1} - \alpha^{k+1}\|_2}{\|f\|_2} \leq \epsilon_S.$$

Throughout this subsection, the codes of all the algorithms are written in MATLAB and all computations below are performed on a workstation with Intel Xeon E5410 CPU (2.33GHz) and 8GB RAM running Red Hat Enterprise Linux (kernel 2.6.18). If not specified, the piecewise linear B-spline framelets constructed by [111] are used in all the numerical

experiments. We also take $D = I$ for all three methods for simplicity. For the PD method, we choose $\epsilon_I = 10^{-4}$ and $\epsilon_O = 10^{-3}$ and set $\alpha^{0,0}$, α^{feas} and u^{feas} to be zero vectors. In addition, we choose[9, Algorithm 2.2] and set $M = 20$, $\epsilon_D = 5 \times 10^{-5}$ and $\epsilon_F = 10^{-4}$ (if necessary) for the projected gradient method applied to the subproblem arising in step 1a) of the BCD method.

5.3.1 Experiments on CT image reconstruction

In this subsection, we apply the PD method stated in Subsection 5.2 to solve problem (5.6) on CT images and compare the results with the balanced approach (5.2) and the analysis based approach (5.3). The matrix A in (5.1) is taken to be a projection matrix based on fan-beam scanning geometry using Siddon's algorithm [120], and η is generated from a zero mean Gaussian distribution with variance $\sigma = 0.01\|f\|_\infty$. In addition, we pick level of framelet decomposition to be 4 for the best quality of the reconstructed images. For balanced approach, we set $\kappa = 2$ and take $\epsilon_P = 1.5 \times 10^{-2}$ for the stopping criteria of the APG algorithm. We set $\epsilon_S = 10^{-5}$ for the stopping criteria of the split Bregman algorithm when solving the analysis based approach. Moreover, we take $\mathcal{Y} = \{x \in \mathbb{R}^n : x_i \geq 0 \; \forall i = 1, \ldots, n\}$ for model (5.6), and take $\delta = 10$ and $\varrho_0 = 10$ for the PD method. To measure quality of the restored image, we use the PSNR value defined by

$$\text{PSNR} := -20 \log_{10} \frac{\|u - \tilde{u}\|_2}{n},$$

where u and $\tilde{u}$ are the original and restored images respectively, and n is total number of pixels in u.

Table 5.1 summarizes the results of all three models when applying to the CT image restoration problem and the corresponding images and their zoom-in views are shown in Figure 5.1 and Figure 5.2. In Table 5.1, the CPU time (in seconds) and PSNR values of all three methods are given in the first and second row, respectively. In order to fairly compare the results, we have tuned the parameter λ to achieve the best quality of the restoration images for each individual method. We observe that based on the PSNR values listed in Table 5.1 the analysis based approach and the PD method obviously achieve better restoration results than the balanced approach. Nevertheless, the APG algorithm for the balanced approach is the fastest algorithm in this experiment. In addition, the PD method is faster and achieves larger PSNR than the split Bergman algorithm for the analysis based

Table 5.1: Comparisons: CT image reconstruction

	Balanced approach	Analysis based approach	PD method
Time	56.0	204.8	147.6
PSNR	56.06	59.90	60.22

approach. Moreover, we can observe from Figure 5.2 that the edges are recovered better by the PD method and the balanced approach.

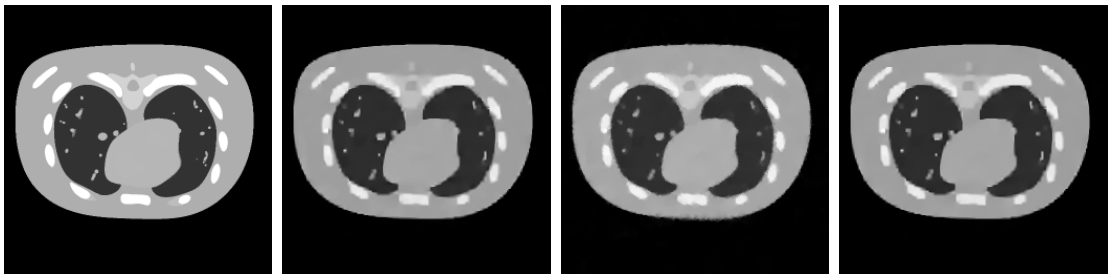


Figure 5.1: CT image reconstruction. Images from left to right are: original CT image, reconstructed image by balanced approach, reconstructed image by analysis based approach and reconstructed image by PD method.

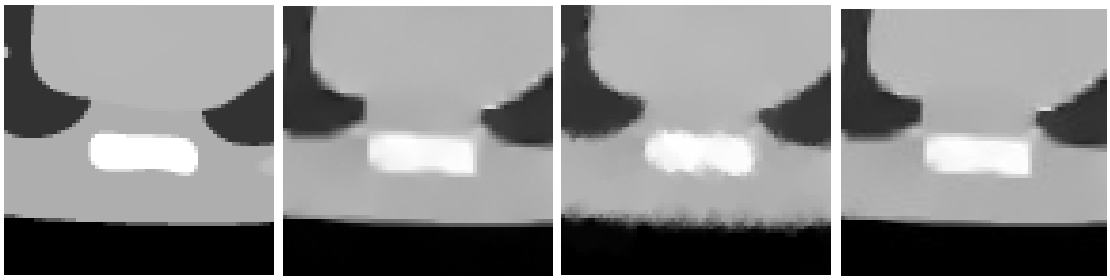


Figure 5.2: Zoom-in views of the CT image reconstruction. Images from left to right are: original CT image, reconstructed image by balanced approach, reconstructed image by analysis based approach and reconstructed image by PD method.

5.3.2 Experiments on image deconvolution

In this subsection, we apply the PD method stated in Subsection 5.2 to solve problem (5.6) on image deblurring problems and compare the results with the balanced approach (5.2) and the analysis based approach (5.3). The matrix A in (5.6) is taken to be a convolution matrix with corresponding kernel a Gaussian function (generated in MATLAB by

"fspecial('gaussian',9,1.5);") and η is generated from a zero mean Gaussian distribution with variance $\sigma = 3$ if not specified. In addition, we pick level of framelet decomposition to be 4 for the best quality of the reconstructed images. We set $\kappa = 1$ for balanced approach and choose both ϵ_P and ϵ_S to be 10^{-4} for the stopping criteria of both APG algorithm and the split Bregman algorithm. Moreover, we set $\mathcal{Y} = \{x \in \mathbb{R}^n : 0 \leq x_i \leq 255 \ \forall i = 1, \ldots, n\}$ for model (5.6), and take $\delta = 10$ and $\varrho_0 = 10^{-3}$ for the PD method. To measure quality of restored image, we use the PSNR value defined by

$$\text{PSNR} := -20 \log_{10} \frac{\|u - \tilde{u}\|_2}{255n}.$$

We first test all three methods on twelve different images by using piecewise linear wavelet and summarize the results in Table 5.2. The names and sizes of images are listed in the first two columns. The CPU time (in seconds) and PSNR values of all three methods are given in the rest six columns. In addition, the zoom-in views of original images, observed images and recovered images are shown in Figure 5.3-5.4. In order to fairly compare the results, we have tuned the parameter λ to achieve the best quality of the restoration images for each individual method and each given image.

We first observe that in Table 5.2, the PSNR values obtained by the PD method are generally better than those obtained by other two approaches. Although for some of the images (i.e. "Downhill", "Bridge", "Duck" and "Barbara"), the PSNR values obtained by the PD methods are comparable to those of balanced and analysis based approaches, the quality of the restored images can not only be judged by their PSNR values. Indeed, the zoom-in views of the recovered images in Figure 5.3 and Figure 5.4 show that for all tested images, the PD method produces visually superior results than the other two approaches in terms of both sharpness of edges and smoothness of regions away from edges. Takeing the image "Barbara" as an example, the PSNR value of the PD method is only slightly greater than that obtained by the other two approaches. However, the zoom-in views of "Barbara" in Figure 5.4 show that the face of Barbara and the textures on her scarf are better recovered by the PD method than the other two approaches. This confirms the observation that penalizing l_0-"norm" of Wu should provide good balance between sharpness of features and smoothness of the reconstructed images. We finally note that the PD method is slower than other two approaches in these experiments but the processing time of the PD method is still acceptable.

We next compare all three methods on "portrait I" image by using three different tight

wavelet frame systems, i.e., Haar framelets, piecewise linear framelets and piecewise cubic framelets constructed by [111]. We summarize the results in Table 5.3. The names of three wavelets are listed in the first column. The CPU time (in seconds) and PSNR values of all three methods are given in the rest six columns. In Table 5.3, we can see that the quality of the restored images by using the piecewise linear framelets and the piecewise cubic framelets is better than that by using the Haar framelets. In addition, all three methods are generally faster when using Haar framelets and slower when using piecewise cubic framelets. Overall, all three approaches when using the piecewise linear have balanced performance in terms of time and quality (i.e., the PSNR value). Finally, we observe that the PD method consistently achieves the best quality of restored images among all the approaches for all three different tight wavelet frame systems.

Finally, we test how the different noise levels effect the restored images obtained from all three methods. We choose three different noise levels (i.e., $\sigma = 3, 5, 7$) for "portrait I" image and test all the methods by using piecewise linear framelets. We summarize the results in Table 5.4. The variances of noises are listed in the first column. The CPU time (in seconds) and PSNR values of all three methods are given in the rest six columns. In Table 5.4, we can see that the quality of the restored images by all three methods is decreased when the noise level is increased. Nevertheless, the quality of the recovered by the PD method is still significantly better than other two methods for all three noise levels. We also observe that the PD method is slower than other two approaches in these experiments but the processing time of the PD method is still acceptable.

5.4 Conclusion

In this chapter, we proposed a wavelet frame based l_0 minimization model, which is motivated by the analysis based approach and balanced approach. The penalty decomposition (PD) method of [89] was used to solve the proposed optimization problem. Numerical results showed that the proposed model solved by the PD method can generate images with better quality than those obtained by either analysis based approach or balanced approach in terms of restoring sharp features like edges as well as maintaining smoothness of the recovered images. Convergence analysis of the sub-iterations in the PD method was also provided.

Table 5.2: Comparisons: image deconvolution

		Balanced approach		Analysis based approach		PD method	
Name	Size	Time	PSNR	Time	PSNR	Time	PSNR
Downhill	256	12.5	27.24	6.1	27.36	29.5	27.35
Cameraman	256	18.2	26.65	7.0	26.73	31.1	27.21
Bridge	256	14.5	25.40	5.1	25.46	33.0	25.44
Pepper	256	21.6	26.82	7.5	26.63	32.1	27.29
Clock	256	17.3	29.42	19.9	29.48	22.3	29.86
Portrait I	256	32.7	33.93	19.3	33.98	27.1	35.44
Duck	464	30.6	31.00	16.1	31.11	72.5	31.09
Barbara	512	38.8	24.62	12.3	24.62	77.4	24.69
Aircraft	512	55.9	30.75	35.1	30.81	67.5	31.29
Couple	512	91.4	28.40	41.5	28.14	139.1	29.32
Portrait II	512	45.2	30.23	22.1	30.20	48.9	30.90
Lena	516	89.3	12.91	31.0	12.51	67.0	13.45

Table 5.3: Comparisons among different wavelet representations

	Balanced approach		Analysis based approach		PD method	
Wavelets	Time	PSNR	Time	PSNR	Time	PSNR
Haar	17.9	33.63	20.2	33.80	24.3	34.68
Piecewise linear	32.7	33.93	22.3	33.98	27.1	35.44
Piecewise cubic	61.0	33.95	37.3	34.00	37.8	35.20

Table 5.4: Comparisons among different noise levels

	Balanced approach		Analysis based approach		PD method	
Variances of noises	Time	PSNR	Time	PSNR	Time	PSNR
$\sigma = 3$	32.7	33.93	22.3	33.98	27.1	35.44
$\sigma = 5$	23.7	32.84	19.4	32.89	27.2	34.48
$\sigma = 7$	19.6	32.11	25.0	32.14	29.7	33.69

Figure 5.3: Zoom-in to the texture part of "downhill", "cameraman", "bridge", "pepper", "clock", and "portrait I". Image from left to right are: original image, observed image, results of the balanced approach, results of the analysis based approach and results of the PD method.

Figure 5.4: Zoom-in to the texture part of "duck", "barbara", "aircraft", "couple", "portrait II" and "lena". Image from left to right are: original image, observed image, results of the balanced approach, results of the analysis based approach and results of the PD method.

Bibliography

[1] O. Alter, P. Brown, and D. Botstein. Singular value decomposition for Genome-Wide expression data Processing and Modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000. 37

[2] A. Antoniadis and J. Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–967, 2001. 104

[3] G. Aubert and P. Kornprobst. *Mathematical problems in image processing: partial differential equations and the calculus of variations*. Springer, 2006. 96

[4] O. Banerjee, L. E. Ghaoui, and A. D'Aspremont. Model selection through sparse maximum likelihood estimation. *The Journal of Machine Learning Research*, 9:485–516, 2008. 82

[5] J. Barzilai and J. M. Borwein. Two point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988. 49

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. 21, 110

[7] D. P. Bertsekas *Nonlinear Programming*. Athena Scientific, 1999. 17, 49

[8] J. A. Bilmes. Factored sparse inverse covariance matrices. *International Conference on Acoustics, Speech and Signal processing*, Washington, D.C., 1009–1012, 2000. 3

[9] E. G. Birgin, J. M. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10(4):1196–1211, 2000. 21, 24, 31, 49, 79, 101, 103, 111

[10] C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2007. 4

[11] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008. 6, 90, 91, 100, 104, 108

[12] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009. 6, 90, 91, 100

[13] S. Burer and R. D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003. 36

[14] J. Cadima and I. Jolliffe. Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, 22(2):203–214, 1995. 8, 39

[15] J. F. Cai, R. H. Chan, L. Shen, and Z. Shen. Convergence analysis of tight framelet approach for missing data recovery. *Advances in Computational Mathematics*, 31(1):87–113, 2009. 94, 97

[16] J. F. Cai, R. H. Chan, and Z. Shen. Simultaneous cartoon and texture inpainting. *Inverse Problems and Imaging*, 4(3):379–395, 2010. 97

[17] J. F. Cai, B. Dong, S. Osher, and Z. Shen. Image restorations: total variation, wavelet frames and beyond. *Journal of the American Mathematical Society*, 25(4):1033–1089, 2012. 96, 98, 99

[18] J. F. Cai, S. Osher, and Z. Shen. Linearized Bregman iterations for frame-based image deblurring. *SIAM Journal on Imaging Sciences*, 2(1):226–252, 2009. 96

[19] J. F. Cai, S. Osher, and Z. Shen. Split Bregman methods and frame based image restoration. *Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal*, 8(2):337–369, 2009. 94, 96, 98, 110

[20] E. J. Candés. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589-592, 2008. 5

[21] E. J. Candés, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2011. 98, 99

[22] E. J. Candés and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. 89

[23] E. J. Candés, J. Romberg and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006. 5, 98

[24] E. J. Candés and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005. 98

[25] E. J. Candés and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006. 98

[26] A. Chai and Z. Shen. Deconvolution: A wavelet frame approach. *Numerische Mathematik*, 106(4):529–587, 2007. 98

[27] A. Chambolle and P. L. Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997. 96

[28] R. H. Chan, T. F. Chan, L. Shen, and Z. Shen. Wavelet algorithms for high-resolution image reconstruction. *SIAM Journal on Scientific Computing*, 24(4):1408–1432, 2003. 94, 96, 97

[29] R. H. Chan, S. D. Riemenschneider, L. Shen, and Z. Shen. Tight frame: an efficient way for high-resolution image reconstruction. *Applied and Computational Harmonic Analysis*, 17(1):91–115, 2004. 97

[30] R. H. Chan, L. Shen, and Z. Shen. A framelet-based approach for image inpainting. Technical Report, Department of Mathematics, The Chinese University of Hong Kong, 2005. 96

[31] R. H. Chan, Z. Shen, and T. Xia. A framelet algorithm for enhancing video stills. *Applied and Computational Harmonic Analysis*, 23(2):153–170, 2007. 98

[32] T. Chan, S. Esedoglu, F. Park, and A. Yip. Total variation image restoration: Overview and recent developments. *Handbook of Mathematical Models in Computer Vision*, 17–31, 2006. 96

[33] T. F. Chan and J. Shen. *Image Processing and Analysis: variational, PDE, wavelet, and stochastic methods*. Society for Industrial and Applied Mathematics, 2005. 96

[34] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007. 6

[35] S. Chen, D. Donoho and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998. 2, 5, 87

[36] X. Chen, F. Xu and Y. Ye. Lower bound theory of nonzero entries in solutions of l_2-l_p Minimization. *SIAM Journal on Scientific Computing*, 32(5):2832–2852, 2010. 6

[37] X. Chen and W. Zhou. Convergence of reweighted l_1 minimization algorithms and unique solution of truncated l_p minimization. Technical report, Department of Applied Mathematics, The Hong Kong Polytechnic University, 2010. 6

[38] K. Chin, S. Devries, J. Fridlyand, P. Spellman, R. Roydasgupta, W.-L. Kuo, A. Lapuk, R. Neve, Z. Qian, T. Ryder, and others. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541, 2006. 8, 37, 40, 57

[39] J. Claerbout and F. Muir. Robust modelling of erratic data. *Geophysics*, 38(5):826–844, 1973. 2

[40] A. d'Aspremont, F. R. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294, 2008. 8, 39

[41] A. d'Aspremont, O. Banerjee and L. E. Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008. 5, 82, 84, 85

[42] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007. 4, 8, 39, 50, 52, 53, 56

[43] J. Dahl, L. Vandenberghe and V. Roychowdhury. Covariance selection for nonchordal graphs via chordal embedding. *Optimization Methods & Software*, 23(4):501–520, 2008. 82

[44] I. Daubechies. *Ten lectures on wavelets.* Philadelphia: Society for industrial and applied mathematics, 1992. 96, 97

[45] I. Daubechies, B. Han, A. Ron, and Z. Shen. Framelets: MRA-based constructions of wavelet frames. *Applied and Computational Harmonic Analysis*, 14(1):1–46, 2003. 96

[46] I. Daubechies, G. Teschke, and L. Vese. Iteratively solving linear inverse problems under general convex constraints. *Inverse Problems and Imaging*, 1(1):29–46, 2007. 98

[47] A. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1978.

[48] A. Dobra. Dependency networks for genome-wide data. Biostatistics, 8(1):1–28, 2007. 86

[49] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao and M. West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212, 2004. 3

[50] B. Dong, A. Chien, and Z. Shen. Frame based segmentation for medical images. *Communications in Mathematical Sciences*, 9(2):551–559, 2010. 96

[51] B. Dong and Z. Shen. MRA based wavelet frames and applications. *IAS Lecture Notes Series, Summer Program on "The Mathematics of Image Processing"*, Park City Mathematics Institute, 2010. 96, 97, 99

[52] B. Dong and Z. Shen. Wavelet frame based surface reconstruction from unorganized points. *Journal of Computational Physics*, 230(22):8247–8255, 2011. 96

[53] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. 98

[54] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. 78

[55] M. Elad, J.L. Starck, P. Querre, and D. L. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Applied and Computational Harmonic Analysis*, 19(3):340–358, 2005. 94, 98

[56] M. J. Fadili and J. L. Starck. Sparse representations and bayesian image inpainting. *Proceedings of SPARS*, 2005. 98

[57] M. J. Fadili, J. L. Starck, and F. Murtagh. Inpainting and zooming using sparse representations. *The Computer Journal*, 52(1):64, 2009. 98

[58] K. Fan. On a theorem of Weyl concerning the eigenvalues of linear transformations. *Proceedings of the National Academy of the Sciences of U.S.A.*, 35: 652–655 (1949). 42

[59] M. A. T. Figueiredo and R. D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906–916, 2003. 98

[60] M. A. T. Figueiredo and R. D. Nowak. A bound optimization approach to wavelet-based image deconvolution. *IEEE International Conference on Image Processing*, 2005. 98

[61] M. A. T. Figueiredo, R. D. Nowak and S. J. Wright. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–598, 2007. 90, 91

[62] J. Friedman, T. Hastie and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432-441, 2008. 3, 82

[63] D. Geman and C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing*, 4(7):932–946, 1995. 96

[64] T. Goldstein and S. Osher. The split Bregman algorithm for L1 regularized problems. *SIAM Journal on Imaging Sciences*, 2(2):323–343, 2009. 110

[65] 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. 4

[66] G. Golub and C. Van Loan. *Matrix Computations*. John Hopkins University Press, 2012. 79

[67] 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(5439):531–537, 1999. 86

[68] E. T. Hale, W. Yin and Y. Zhang. Fixed-point continuation applied to compressed sensing: Implementation and numerical experiments. *Journal of Computational Mathematics* , 28(2):170–194, 2010. 90

[69] P. Hancock, A. Burton, and V. Bruce. Face processing: Human perception and principal components analysis. *Memory and Cognition*, 24(1):26–40, 1996. 37

[70] T. Hastie, R. Tibshirani, M. Eisen, P. Brown, D. Ross, U. Scherf, J. Weinstein, A. Alizadeh, L. Staudt, and D. Botstein. ́gene Shavingás a method for identifying distinct sets of genes with similar Expression Patterns. *Genome Biology*, 1(2):1–21, 2000. 37

[71] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning; Data mining, Inference and Prediction*. New York: Springer Verlag. 37

[72] 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. 6, 100

[73] J. B. Hiriart–Urruty and C. Lemaréchal. *Convex Analysis and Minimization algorithms I.* Comprehensive Study in Mathematics, Vol. 305, Springer-Verlag, New York, 1993. 19

[74] J. Jeffers. Two Case Studies in the Application of Principal Component. *Applied Statistics*, 16:225–236, 1967. 8, 37, 40, 52

[75] X. Jia, Y. Lou, B. Dong, and S. Jiang. GPU-based iterative cone beam CT reconstruction using tight frame regularization. *arXiv preprint*, arXiv:1008.2042, 2010. 98

[76] I. Jolliffe. Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22(1):29–35, 1995. 8, 39

[77] M. Journée, Yu. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research*, 11:517–553, 2010. 4, 8, 39, 50, 52, 53, 56, 57

[78] I. T. Jolliffe, N. T. Trendafilov, and M. L. Uddin. A modified principal component technique based on the Lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003. 8, 39, 40, 53

[79] 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 Journal of Selected Topics in Signal Processing*, 1(4):606–617, December 2007. 90

[80] K. Koh, S. J. Kim and S. Boyd. An interior-point method for large-scale l_1-regularized logistic regression. *The Journal of Machine Learning Research*, 8:1519–1555, 2007. 78, 79, 81

[81] Linear Algebra PACKage. Available at http://www.netlib.org/lapack/index.html. 84

[82] S. Lee, H. Lee, P. Abbeel and A. Ng. Efficient l_1-regularized logistic regression. In *21th National Conference on Artificial Intelligence (AAAI)*, 2006. 78

[83] 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(9):1235–1243, 1981. 2

[84] 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. 4

[85] 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. 78, 79

[86] L. Li and K. C. Toh. An inexact interior point method for l_1-regularized sparse covariance selection. *Mathematical Programming Computation*, 2(3-4):291–315, 2010. 82, 84, 86

[87] Z. Lu. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization*, 19(4):1807–1827, 2009. 82, 84

[88] Z. Lu. Adaptive first-order methods for general sparse inverse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 31(4):2000–2016, 2010. 82, 84, 85

[89] Z. Lu and Y. Zhang. Penalty decomposition methods for l_0-norm minimization. Technical Report, Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada, 2010. 94, 99, 100, 101, 104, 110, 114

[90] Z. Lu and Y. Zhang. An Augmented Lagrangian Approach for Sparse Principal Component Analysis. *Mathematical Programming*, 135(1-2):149–193, 2012. 12, 37

[91] Z. Lu and Y. Zhang. Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization*, 23(4):2448-2478, 2013 61

[92] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE T. Image Process.*, 41(12):3397–3415, 1993. 1, 3, 6

[93] Y. Meyer. *Oscillating patterns in image processing and nonlinear evolution equations: the fifteenth Dean Jacqueline B. Lewis memorial lectures*, volume 22. Amer Mathematical Society, 2001. 95, 96

[94] A. Miller. *Subset Selection in Regression.* Chapman and Hall, London, 2002. 2

[95] B. Moghaddam, Y. Weiss, and S. Avidan. Spectral bounds for sparse PCA: Exact and greedy algorithms. *Advances in Neural Information Processing Systems*, 18:915–922, 2006. 8, 39

[96] R. D. C. Monteiro. *Private Communication*, 2009. 36

[97] Y. E. Nesterov. Gradient methods for minimizing composite objective functions. CORE Discussion paper 200/76, Catholic University of Louvain, Belgium, September 2007. 21

[98] 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. 79, 81

[99] 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. 4, 5

[100] S. Osher and R.P. Fedkiw. *Level set methods and dynamic implicit surfaces*. Springer, 2003. 95, 96

[101] M. L. Overton and R. S. Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 62(1-3):321–357, 1993. 41

[102] M. Y. Park and T. Hastie. *Regularization path algorithms for detecting gene interactions*. Technical Report, Department of Statistics, Stanford University, 2006. 78

[103] 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. 4

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

[105] 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. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8431–8436, 2004. 86

[106] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2007. 89

[107] S. M. Robinson. Stability theory for systems of inequalities, Part 2: Differentiable nonlinear systems. *SIAM Journal on Numerical Analysis*, 13(4):497–513, 1976. 13

[108] S. M. Robinson. *Local structure of feasible sets in nonlinear programming, Part I: Regularity.* Springer-Verlag, Berlin, 1983. 13

[109] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970. 15, 20, 24, 27, 28, 33

[110] R. T. Rockafellar and J. B. W. Roger. *Variational Analysis*. Vol. 317, Springer, 2004. 104

[111] A. Ron and Z. Shen. Affine Systems in $L_2(\mathbb{R}^d)$: The Analysis of the Analysis Operator. *Journal of Functional Analysis*, 148(2):408–447, 1997. 96, 97, 110, 114

[112] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physics D: Nonlinear Phenomena*, 60(1):259–268, 1992. 95

[113] A. Ruszczyński. *Nonlinear Optimization*. Princeton University Press, 2006. 13, 14, 16, 17, 20, 24, 27, 28, 33, 62, 63

[114] F. Santosa and W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986. 2

[115] G. Sapiro. *Geometric partial differential equations and image analysis*. Cambridge University Press, 2001. 95

[116] Z. Shen. Wavelet frames and image restorations. *Proceedings of the International Congress of Mathematicians, Hyderabad, India*, 2010. 96

[117] H. Shen and J. Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034, 2008. 4, 8, 39, 50, 52, 53, 56

[118] Z. Shen, K. C. Toh, and S. Yun. An accelerated proximal gradient algorithm for frame based image restorations via the balanced approach. *SIAM Journal on Imaging Sciences*, 4(2):573–596, 2011. 98, 99, 110

[119] J. Shi, W. Yin, S. Osher and P. Sajda. A fast hybrid algorithm for large-scale l_1-regularized logistic regression. *The Journal of Machine Learning Research*, 11:713–741, 2010. 78

[120] R. L. Siddon. Fast calculation of the exact radiological path for a three-dimensional CT array. *Medical Physics*, 12(2):252–255, 1985. 111

[121] J. L. Starck, M. Elad, and D. L. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Transactions on Image Processing*, 14(10):1570–1582, 2005. 94, 98

[122] G. Steidl, J. Weickert, T. Brox, P. Mrázek, and M. Welk. On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and sides. *SIAM Journal on Numerical Analysis*, pages 686–713, 2005. 96

[123] H. Taylor, S. Bank and J. McCoy. Deconvolution with the l_1-norm. *Geophysics*, 44(1):39–52, 1979. 2

[124] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996. 4, 5

[125] J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004. 6

[126] J. Tropp. Just relax: Convex programming methods for identifying sparse signals. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006. 2

[127] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001. 99

[128] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009. 8, 21, 22, 23, 28, 29, 31, 34, 35

[129] 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. 4

[130] E. Van Den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2)890-912, 2008. 87, 89

[131] E. Van Den Berg, M. P. Friedlander, G. Hennenfent, F. Herrmann, R. Saab and O. Yilmaz. Algorithm 890: Sparco: a testing framework for sparse reconstruction. *ACM Transactions on Mathematical Software*, 35(4):1–16, 2009. 89

[132] C. Wang, D. Sun and K. C. Toh. Solving log-determinant optimization problems by a Newton-CG proximal point algorithm. *SIAM Journal on Optimization*, 20(6):2994–3013, 2010. 82, 83, 84

[133] Y. Wang, J. Yang, W. Yin, and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 1(3):248–272, 2008. 96

[134] S J. Wright, R. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(3):2479–2493, 2009. 21, 23, 24

[135] 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. 86

[136] Y. Zhang, B. Dong and Z. Lu. l_0 minimization for wavelet frame based image restoration. *Mathematics of Computation*, 82(282):995–1015, 2013. 94

[137] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006. 4, 8, 37, 38, 39, 40, 50, 51, 52, 53, 56