# FIRST-ORDER METHODS FOR NUCLEAR NORM MINIMIZATION AND ITS APPLICATIONS

by

Hua Zheng

B.A., University of Science and Technology of China, 2007

A Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Mathematics
in the Department
of
Mathematics


SIMON FRASER UNIVERSITY
Fall 2009

## APPROVAL

| | |
|---|---|
| **Name:** | Hua Zheng |
| **Degree:** | Master of Mathematics |
| **Title of Thesis:** | First-Order Methods for Nuclear Norm Minimization and its Applications |
| **Examining Committee:** | Dr. Tamon Stephen<br>Chair |
| | Dr. Zhaosong Lu,<br>Senior Supervisor |
| | Dr. Abraham Punnen,<br>Supervisor |
| | Dr. Jiguo Cao,<br>SFU Examiner |
| **Date Approved:** | December 3, 2009 |

# Declaration of
# Partial Copyright Licence

# Abstract

We consider a nuclear norm minimization problem that can be viewed as convex relaxation of rank minimization problem arising in many fields in engineering and applied science. Though this problem can be reformulated as semidefinite programming (SDP) problem, it is computationally challenging for general SDP solvers due to the size of this problem. In this thesis, we study first-order methods for nuclear norm minimization problem. In particular, we first propose several reformulations for this problem. Then we apply two suitabe first-order methods, namely, fast iterative shrinkage algorithm (FISTA) and nonmonotone gradient method (NGM) for solving these reformulations. Finally, we compare the performance of these approaches on randomly generated instances and report some promising computational results.

**Key words:** nuclear norm minimization, first-order methods, fast iterative shrinkage algorithm, nonmonotone gradient methods, $l_1$-norm projection, matrix shrinkage threshold problem

*To whoever reads this!*

*"Don't worry, Gromit. Everything's under control!"*

— *The Wrong Trousers*, Aardman Animations, *1993*

# Acknowledgments

I would like to thank my supervisor Dr. Zhaosong Lu for his help all along the way. Thanks for the guidance, support, dinners and everything. Most what I've learned in the past two years is a direct result of his patience. I have learned a lot from Zhaosong including the attitude toward research, scientific writing and coding etc.

I would also like to thank Dr. Abraham Punnen and Dr. Tamon Stephen for their help.

I want to thank the members of our math lab in Surrey for being a supporting company and their friendships. Thank you Annie, Arman, Brad, John, Sareh and Yong.

Most importantly, I want to thank my family.

# Contents

# List of Tables

# Chapter 1

# Introduction

In recent years, there is a rapidly growing interest in the recovery of an unknown low-rank or approximately low-rank matrix from very limited information. This problem has many applications in machine learning [2, 3, 4], Euclidean embedding [28], dimension reduction and coefficient estimation in multivariate linear regression [61, 37], matrix completion [15, 17], control [41], computer vision [56] and collaborative filtering [52]. For instance, in computer vision, inferring scene geometry and camera motion from a sequence of images, known as the structure-from-motion problem, is a lower rank approximation problem since the images form a matrix which has very low rank. Similarly, in statistics, it is of great importance to be able to determine the number of factors for the linear factor model. A more accurate estimate is expected for a smaller number of factors since there are fewer free parameters. Given that a low-rank coefficient matrix of the model usually ensures a small number of factors, it is sufficient to find a coefficient matrix of a low rank.

The matrix rank minimization problem aims to find a matrix of minimal rank subject to a system of linear of equality constraints. Mathematically, it can be formulated as follows:

$$\min_{X \in \Re^{p \times q}} \{\text{rank}(X) : \mathcal{A}(X) = b\}, \tag{1.1}$$

where $\mathcal{A} : \Re^{p \times q} \to \Re^m$ is a linear map, $b \in \Re^m$, and $\text{rank}(X)$ denotes the rank of $X$.

It is well-known that the rank of a matrix equals the cardinality of its singular values. Due to this combinational nature, problem (1.1) is generally NP-hard. Recently, Fazel et al. [27] proposed a convex programming relaxation approach for (1.1). In particular, they replaced the rank function by the trace norm function in (1.1) and the resulting problem

becomes the following convex programming problem:

$$\min\{\|X\|_* : \mathcal{A}(X) = b\}, \tag{1.2}$$

where $\|X\|_*$ denotes the nuclear or trace norm of $X$, that is, the sum of singular values of $X$. One natural concern regarding the above relaxation is: when the relaxation is tight? Very recently, Candès and Tao [16] and Recht et al. [50] provided some sufficient conditions ensuring the above convex relaxation is tight. In other words, under those conditions, problems (1.1) and (1.2) are equivalent in the sense that they have exactly the same unique solutions.

It is interesting to observe that when the matrix $X$ is restricted to be diagonal, problem (1.2) becomes the well known basis pursuit (BP) problem:

$$\min_{x\in\Re^n}\{\|x\|_1 : Ax = b\}. \tag{1.3}$$

It has been shown in [23] that the BP problem (1.3) is capable of identifying sparse solutions to the underdetermined system $Ax = b$ under some suitable assumption. This problem has attracted substantial much interest in compressed sensing [19, 20, 21, 24, 25] and has a wide range of applications in statistics and signal/image processing. In addition, two closely related formualtions to (1.3) were also proposed in literature [23, 29, 12] and they are:

$$\min_{x\in\Re^n}\frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1, \tag{1.4}$$

$$\min_{x\in\Re^n}\{\|Ax - b\|_2^2 : \|x\|_1 \le \tau\}, \tag{1.5}$$

where $\lambda$ and $\tau$ are some constants. The first formulation above is frequently referred to as the basis pursuit denoising (BPDN) while the second one is commonly known as least absolute shrinkage and selection operator (LASSO). It is not hard to show that problems (1.4) and (1.5) are all equivalent to (1.3) for some suitable prameters $\lambda$ and $\tau$. They have also been widely used in statistics and signal/image processing. Similar to (1.3), problem (1.2) can also be solved as the following problems:

$$\min_{X\in\Re^{p\times q}}\frac{1}{2}\|\mathcal{A}(X) - b\|_F^2 + \lambda\|X\|_*, \tag{1.6}$$

$$\min_{X\in\Re^{p\times q}}\{\|\mathcal{A}(X) - b\|_F^2 : \|X\|_* \le \tau\}. \tag{1.7}$$

for some suitable parameters $\lambda$ and $\tau$.

Problems (1.2), (1.6) and (1.7) can be reformulated and solved as semidefinite programming (SDP) problem. Indeed, as shown in [27, 52], problem (1.2) can be reformulated as the following SDP problem:

$$\begin{array}{ll} \min\limits_{U_1,U_2,X} & \frac{1}{2}(\mathrm{Tr}(U_1)+\mathrm{Tr}(U_2)) \\ \text{s.t.} & \begin{bmatrix} U_1 & X \\ X^T & U_2 \end{bmatrix} \succeq 0, \\ & \mathcal{A}(X) = b. \end{array} \tag{1.8}$$

Though problem (1.8) is polynomially solvable, it is computationally challenging for existing SDP solvers when the size of problems is relatively large. For instance, the widely used SDP solvers SDPT3 [59] and SeDuMi [53] are based on interior-point methods. They need assemble and solve typically dense system of linear equations to compute the Newton direction each iterate. When the size of the matrix is relatively large, this system becomes prohibitively huge for storing and computing. Indeed, it is extremely difficult for these SDP solvers to handle the instances of (1.8) when $p$ and $q$ are both larger than 1000 or $m$ is larger than 5000.

Recently various first-order methods have been proposed to solve (1.2) and (1.6), especially for large-scale instances. These methods enjoy good theoretical convergence and have very promising practical performance. In particular, Lu et al. [37] studied a variant of Nestrov's smooth method [44] for solving problem (1.6) in which $\mathcal{A}$ is specialized to $\mathcal{A}(X) = AX\ \forall X \in \Re^{p\times q}$ for some $A \in \Re^{n\times p}$, namely,

$$\min_{X\in\Re^{p\times q}} \frac{1}{2}\|AX-B\|_F^2 + \lambda\|X\|_*, \tag{1.9}$$

where $B \in \Re^{n\times q}$. They showed that the dual of (1.9) is smooth and can be suitably solved by a variant of Nestrov's smooth method [44]. Recently, (accelerate) proximal gradient methods were applied to solve (1.9) in [34, 49]. In addition, Ma et al. [39] studied a fixed-point type method for solving problems (1.2) and (1.6), which can be viewed as an extension of the method proposed by Hale et al. [31] for (1.4). Also, Toh and Yun [55] applied the proximal gradient method proposed in [40] for solving (1.2) and (1.6). Lately, Liu et al. [35] studied inexact proximal point methods for solving (1.2) in a more general setting where second-order cone constraints are also allowed.

In this thesis, we mainly consider problems (1.9) and its closely related problem

$$\min_{X\in\Re^{p\times q}} \{\|AX-B\|_F^2 : \|X\|_* \le \tau\}, \tag{1.10}$$

where $A \in \Re^{n\times p}$ and $B \in \Re^{n\times q}$. We first show that they can be reduced to the case where $A$ has full column rank. Then we derived the dual counterpart of (1.9) and proposed a cone-constrained reformulation for (1.9). Subsequently, we applied the accelerate proximal method FISTA [11] and nonmonotone gradient method [38] for solving the aforementioned formulations, and compare their computational performance. Our numerical results show that nonmonotone gradient methods generally outperform FISTA substantially even though the latter method has a more attractive theoretical iteration complexity. In addition, we provided a simpler proof for matrix shrinkage operator than the one given in [39], which is a major operator used in most of first-order methods for nuclear norm minimization problems.

This thesis is organized as follows. In Chapter 2, we first review two types of first order methods for solving general nonsmooth minimization problem. Then we discuss how to solve $\ell_1$-norm projection and matrix shrinkage-threshold problems that will be used in Chapter 3. In Chapter 3, we discuss in details how to apply the first-order methods to nuclear norm minimization problems (1.9) and (1.10). In Chapter 4, we present the computational results comparing the performance of FISTA and nonmonotone gradient methods applied to the problems (1.9) and (1.10). Finally we present some concluding remarks in Chapter 5.

## 1.1 Notation

Throughout the thesis, all vector spaces are assumed to be finite dimensional. The symbols $\Re^p$ and $\Re^p_+$ (resp., $\Re^p_-$) denote the $p$-dimensional Euclidean space and the nonnegative (resp., nonpositive) orthant of $\Re^p$, respectively, and $\Re_{++}$ denotes the set of positive real numbers. For any $z \in \Re^p$, $\mathrm{Diag}(z)$ denotes the $p \times p$ diagonal matrix whose $i$th diagonal element is $z_i$ for $i = 1, \ldots, p$. Additionally, $\mathcal{D}^p$ denotes the space of $p \times p$ diagonal matrices. The Euclidean norm is denoted by $\|\cdot\|$. The space of all $p \times q$ matrices with real entries is denoted by $\Re^{p\times q}$. Given matrices $X$ and $Y$ in $\Re^{p\times q}$, the standard inner product is defined by $\langle X, Y\rangle = \mathrm{Tr}(X^T Y)$, where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. We denote by $\|X\|_F$ the Frobenius norm of a matrix $X$, that is, $\|X\|_F = \sqrt{\langle X, X\rangle}$. In addition, $\sigma_i(X)$ denotes the $i$th largest singular value of $X$. For a real symmetric matrix $X$, $\lambda_i(X)$ denotes its $i$th largest eigenvalue. The rank of $X$ is denoted by $\mathrm{rank}(X)$.

Let $\mathscr{X}$ be a normed vector space whose norm is denoted by $\|\cdot\|_{\mathscr{X}}$. The dual space of $\mathscr{X}$, denoted by $\mathscr{X}^*$, is the normed vector space consisting of all linear functionals of

$u^* : \mathscr{X} \to \Re$, endowed with the dual norm $\|\cdot\|^*_{\mathscr{X}}$ defined as

$$\|X^*\|^*_{\mathscr{X}} = \max_X \{\langle X^*, X\rangle : \|X\|_{\mathscr{X}} \leq 1\} \quad \forall X^* \in \mathscr{X}^*,$$

where $\langle X^*, X\rangle := X^*(X)$ is the value of the linear functional $X^*$ at $X$.

If $\mathscr{U}$ denotes another normed vector space with norm $\|\cdot\|_{\mathscr{U}}$ and $\mathcal{E} : \mathscr{X} \to \mathscr{U}^*$ is a linear opeartor, the operator norm of $\mathcal{E}$ is defined as

$$\|\mathcal{E}\|_{\mathscr{X},\mathscr{U}} = \max_X \{\|\mathcal{E}X\|^*_{\mathscr{U}} : \|X\|_{\mathscr{X}\leq 1}\}. \tag{1.11}$$

A function $f : \mathcal{X} \subseteq \mathscr{X} \to \Re$ is said to be $L$-Lipschitz-differentiable with respect to $\|\cdot\|_{\mathscr{X}}$ if it is differentiable and

$$\|\nabla f(x) - \nabla f(\bar{x})\|^*_{\mathscr{X}} \leq L\|x - \bar{x}\|_{\mathscr{X}} \quad \forall x, \bar{x} \in \mathcal{X}.$$

# Chapter 2

# Preliminaries

In this chapter, we first introduce a class of nonsmooth minimization problems and establish the first-order optimality conditions for them. Then we review two types of first-order methods, that is, fast iterative shrinkage-thresholding algorithm and nonmonotone gradient methods for these problems and state their convergence results. Finally we discuss how to solve $\ell_1$-norm projection and matrix shrinkage-threshold problems that will be used in Chapter 3.

## 2.1 Nonsmooth minimization problems

In this section we are interested in a class of nonsmooth minimization problems as follows:

$$\min_{x\in\mathcal{X}} \{F(x) := f(x) + \Psi(x)\}. \tag{2.1}$$

We now make some assumptions regarding problem (2.1):

- $\mathcal{X}$ is a closed convex set in $\Re^n$.
- $f$ is continuously differentiable in $\Re^n$ and moreover, $\nabla f$ is Lipschitz continuous on $\mathcal{X}$ with Lipschitz constant $L$ with respect to $\|\cdot\|$, that is,

  $$\|\nabla f(x) - \nabla f(y)\|^* \leq L\|x - y\| \quad \forall x, y \in \mathcal{X},$$

  where $\|\cdot\|$ is an arbitrary norm and $\|\cdot\|^*$ is its dual norm.

- $\Psi$ is a closed convex (possibly nonsmooth) function in $\Re^n$. A special and important case is the one in which $\Psi$ is separable into the sum of individual components of its argument, that is,
$$\Psi(x) = \sum_{i=1}^{n} \Psi_i(x_i).$$
- $\Psi(\cdot)$ is simple enough so that one can find a closed-form solution for minimizing the sum of $\Psi(\cdot)$ with some proximal functions (e.g., separable quadratic function).
- Problem (2.1) is solvable, i.e., $\operatorname{Arg\,min}\{F(x) : \ x \in \mathcal{X}\} \neq \emptyset$.

We next study the first-order optimality conditions for problem (2.1). Before proceeding, we introduce two terminologies, namely, the tangent and normal cones to the set $\mathcal{X}$ at a feasible point.

**Definition 1** *A vector $d$ is said to be a tangent to $\mathcal{X}$ at a point $x \in \mathcal{X}$ if there exist a feasible $\{x^k\}$ approaching $x$ and a sequence of positive scalars $\{t_k\}$ with $t_k \to 0$ such that*

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

*The set of all tangents to $\mathcal{X}$ at $x$ is called the tangent cone and is denoted by $T_{\mathcal{X}}(x)$.*

**Definition 2** *The normal cone to the set $\mathcal{X}$ at the point $x \in \mathcal{X}$ is defined as*

$$N_{\mathcal{X}}(x) = \{v \ | v^T w \leq 0 \quad \forall w \in T_{\mathcal{X}}(x)\}. \tag{2.2}$$

**Definition 3** *A vector $s$ is called a subgradient of the function $\Psi$ at $x \in \mathcal{X}$ if for all $y$,*

$$\Psi(y) - \Psi(x) \geq \langle s, y - x\rangle.$$

*The set of all subgradients of $\Psi$ at a point $x$ is called the subdifferential of $\Psi$ at $x$, denoted by $\partial\Psi(x)$.*

**Definition 4** *Let $x \in \mathcal{X}$ and $d$ be a nonzero vector such that $x + td \in \mathcal{X}$ for sufficiently small $t > 0$. The directional derivative of $\Psi$ at $x$ along $d$, denoted by $\Psi'(x; d)$, is given by the following limit if it exists :*

$$\Psi'(x; d) = \lim_{t\to 0^+} \frac{\Psi(x + td) - \Psi(x)}{t}.$$

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

**Theorem 2.1.1** *If $x^* \in \mathcal{X}$ is a local optimal solution of problem (2.1), then the following statement holds:*

$$0 \in \nabla f(x^*) + \partial\Psi(x^*) + N_{\mathcal{X}}(x^*). \tag{2.3}$$

*In addition, if $f$ is convex and (2.3) holds at $x^* \in \mathcal{X}$, then $x^*$ is an optimal solution of problem (2.1).*

*Proof.* We first show that

$$d^T\nabla f(x^*) + \Psi'(x^*; d) \geq 0 \quad \forall d \in T_{\mathcal{X}}(x^*). \tag{2.4}$$

Let $d \in T_{\mathcal{X}}(x^*)$ be arbitrarily chosen. Then, there exist sequences $\{x^k\}_{k=1}^{\infty} \subseteq \mathcal{X}$ 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 $\Psi$ is convex in $\Re^n$, we can have

$$f(x^* + t_k d) - f(x^k) = o(t_k), \quad \Psi(x^* + t_k d) - \Psi(x^k) = o(t_k), \tag{2.5}$$

where the first equality follows from the Mean Value Theorem while the second one comes from Theorem 10.4 of [51]. Clearly, $x^k \to x^*$. This together with the assumption that $x^*$ is a local minimizer of (2.1), implies that

$$f(x^k) + \Psi(x^k) \geq f(x^*) + \Psi(x^*) \tag{2.6}$$

when $k$ is sufficiently large. In view of (2.5) and (2.6), we obtain that

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

and hence (2.4) holds.

For simplicity of notations, let $S = -\nabla f(x^*) - \partial\Psi(x^*)$. We next show that $S \cap N_{\mathcal{X}}(x) \neq \emptyset$. Suppose for contradiction that $S \cap N_{\mathcal{X}}(x) = \emptyset$. This together with the fact that $S$ and $N_{\mathcal{X}}(x)$ 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 $y \in N_{\mathcal{X}}(x)$, and $d^T y \geq 1$ for any $y \in S$. Clearly, we see that $d \in T_{\mathcal{X}}(x^*)$, and

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

which contradicts (2.4). Hence, we have $S \cap N_{\mathcal{X}}(x) \neq \emptyset$. Using this relation, (2.2) and the definitions of $S$, we easily see that (2.3) holds.

We now prove the second part of this theorem. Given any $x \in \mathcal{X}$, let $d = x - x^*$. Clearly, $d \in T_{\mathcal{X}}(x^*)$. It together with (2.3) implies that

$$d^T(\nabla f(x^*) + s) \geq 0 \quad \forall s \in \partial\Psi(x^*).$$

Due to convexity of $f$ and $\Psi$, we easily know that

$$F(x) - F(x^*) = F(x^* + d) - F(x^*) \geq d^T(\nabla f(x^*) + s) \quad \forall s \in \partial\Psi(x^*).$$

It follows from the above inequalities that $F(x) - F(x^*) \geq 0$. Due to the arbitrarity of $x$, we conclude that $x^*$ is an optimal solution of problem (2.1).

■

**Definition 5** *$x \in \Re^n$ is called a stationary point of problem (2.1) if $x \in \mathcal{X}$ and it satisfies (2.3).*

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

$$d_H(x) := \arg\min_d \left\{ \nabla f(x)^T d + \frac{1}{2} d^T H d + \Psi(x + d) : \ x + d \in \mathcal{X} \right\}. \tag{2.7}$$

The following proposition provides an alternative characterization of stationarity that will be used subsequently.

**Proposition 2.1.2** *For any $H \succ 0$, $x \in \mathcal{X}$ is a stationary point of problem (2.1) if and only if $d_H(x) = 0$.*

*Proof.* We first observe that (2.7) is a convex problem, and moreover its objective function is strictly convex. The conclusion of this proposition immediately follows from this observation and the first-order optimality condition of (2.7). ∎

## 2.2 First-order methods

In this section we review two types of first-order methods for problem (2.1) and state their convergence results. In particular, in Subsection 2.2.1 we review a fast iterative shrinkage-thresholding algorithm (FISTA) proposed by Beck and Teboulle [10]. In Subsection 2.2.2, we review two nonmonotone gradient methods proposed by Lu and Zhang [38] and Wright et al. [60], respectively.

### 2.2.1 FISTA method

Recently Beck and Teboulle [10] proposed a first-order method called FISTA for solving problem (2.1) in the case when $f$ is convex and Lipschitz differentiable. In their method, a "quadratic" approximation of $F(X) = f(x) + \Psi(x)$ at a given point $y \in \mathcal{X}$ is introduced as follows:

$$Q_L(x, y) := f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2 + \Psi(x). \tag{2.8}$$

Let $T_L(y)$ denote a unique minimizer of $Q_L(x, y)$ over $\mathcal{X}$, that is,

$$T_L(y) = \arg\min_{x \in \mathcal{X}} Q_L(x, y). \tag{2.9}$$

We are now ready to describe the algorithm framework of FISTA [10] for solving problem (2.1).

**Algorithm 1**: FISTA

Let $L$ be the Lipschitz constant of $\nabla f$. Choose an arbitrary $x^0 \in \mathcal{X}$. Set $y^1 = x^0$, $\theta_1 = 1$ and $k = 1$.

**begin**

$x^k = T_L(y^k)$

$\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$

$y^{k+1} = x^k + (\frac{\theta_k - 1}{\theta_{k+1}})(x^k - x^{k-1})$

$k \leftarrow k + 1$

**end**

The convergence result of FISTA for solving problem (2.1) is stated as follows. Its proof can be found in [10].

**Theorem 2.2.1** *Let the sequences $\{x^k\}$ and $\{y^k\}$ be generated by FISTA when applied to problem (2.1) for which $f$ is convex and Lipschitz differentiable. Then, for any $k \geq 1$,*

$$F(x^k) - F(x^*) \leq \frac{2L\|x^0 - x^*\|^2}{(k+1)^2},$$

*where $x^*$ is an arbitrary optimal solution of problem (2.1).*

*Remark.* The main computational effort of FISTA lies in evaluating the operator $T_L(\cdot)$. Generally, the computation of $T_L(\cdot)$ can be intensive. Nevertheless, when $\mathcal{X}$ and $\Psi$ are simple, $T_L(\cdot)$ typically has a closed-form expression. ∎

### 2.2.2 Nonmonotone gradient methods

In this section we review two nonmonotone gradient methods proposed by Lu and Zhang [38] and Wright et al. [60], respectively, for solving problem (2.1).

We first describe the nonmonotone gradient method proposed by Lu and Zhang [38] for solving problem (2.1).

**Algorithm 2**: Nonmonotone gradient method I (NGM I)

*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 $x^0 \in \mathcal{X}$.*

**begin**

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

Solve (2.7) with $x = x^k$ and $H = H_k$ to obtain $d^k = d_H(x)$.

Compute $\Delta_k = \nabla f(x^k)^T d^k + \Psi(x^k + d^k) - \Psi(x^k)$.

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

$$F(x^k + \alpha_k d^k) \ \leq \ \max_{[k-M]^+ \leq i \leq k} F(x^i) + \sigma \alpha_k \Delta_k.$$

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

**end**

The convergence result of the method NGM I for solving problem (2.1) is stated as

follows. Its proof can be found in [38].

**Theorem 2.2.2** *Suppose that $F$ is bounded below in $\mathcal{X}$ and uniformly continuous in the level set $\mathcal{L} := \{x \in \mathcal{X} : F(x) \leq F(x^0)\}$. Then, the method NGM I is well defined and moreover, any accumulation point of the sequence $\{x^k\}$ generated by NGM I is a stationary point of (2.1).*

*Remark.* Recall that the method FISTA can only be applied to problem (2.1) for which $f$ is convex and Lipschitz differentiable. Nevertheless, the method NGM I can be applied to (2.1) with any smooth function $f$. ∎

We next describe another nonmontone gradient method proposed by Wright et al. [60] for solving problem (2.1).

**Algorithm 3**: Nonmonotone gradient method II (NGM II)

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

**begin**

Choose $\alpha_k \in [\underline{\alpha}, \bar{\alpha}]$.

**repeat**

$x^{k+1} = \arg\min_{x \in \mathcal{X}} \{\nabla f(x^k)^T(x - x^k) + \frac{\alpha_k}{2}\|x - x^k\|_2^2 + \Psi(x)\}$,

$\alpha_k \leftarrow \eta\alpha_k$

**until**

$$F(x^{k+1}) \leq \max_{[k-M]^+ \leq i \leq k} F(x^i) - \frac{\sigma}{2}\alpha_k\|x^{k+1} - x^k\|_2^2.$$

$k \leftarrow k + 1$.

**end**

The convergence result of the method NGM II for solving problem (2.1) is stated as follows. Its proof can be found in [60].

**Theorem 2.2.3** *Suppose that the method NGM II is applied to (2.1), where $f$ is Lipschitz continuously differentiable, $\Psi(x)$ is convex and finite-valued and $F$ is bounded below. Then all accumulation points are stationary points of (2.1).*

*Remark.* We remark that for global convergence, the method NGM II [60] needs a strong assumption that the function $f$ is *Lipschitz* continuously differentiable, which is not required for the method NGM I [38]. In addition, the method NGM I can be viewed as an extension of the second projected gradient method (namely, SPG2) studied in [13] for smooth problems, but the method NGM II cannot. ∎

## 2.3 $l_1$-norm projection

In this section we discuss efficient approaches to solving the $l_1$-norm projection problem that appears in the first-order methods for problems (1.4) and (1.5).

Given $c \in \Re^n$ and $\tau > 0$, consider the $l_1$-norm projection problem in the form of

$$\min_x \{\|c - x\|_2 : \ \|x\|_1 \leq \tau\}. \tag{2.10}$$

Let $x^*$ be the unique optimal solution of (2.10). Clearly, if $\|c\|_1 \leq \tau$, we have $x^* = c$. Otherwise, $\|x^*\|_1 = \tau$ and moreover, there exists a $\lambda \geq 0$ such that $x^*$ is also the optimal solution of the problem:

$$\min_x \frac{1}{2}\|c - x\|_2^2 + \lambda\|x\|_1. \tag{2.11}$$

It is well-known that the optimal solution of (2.11) is given by the soft thresholding operator

$$x_c(\lambda) = \text{sign}(c) \cdot \max\{0, |c| - \lambda\}, \tag{2.12}$$

where sign : $\Re^n \to \Re^n$ is defined as follows:

$$[\text{sign}(c)]_i = \begin{cases} 1 & \text{if } c_i \geq 0, \\ -1 & \text{otherwise,} \end{cases} \quad \text{for } i = 1, \ldots, n.$$

Thus, the solution of (2.10) can be reduced to find the root of the equation

$$\varphi(\lambda) - \tau = 0, \tag{2.13}$$

where $\varphi(\lambda) := \|x_c(\lambda)\|_1$.

Recently two efficient approaches were proposed in literature for finding the root of (2.13). In particular, Candès and Romberg [18] and Daubechies et al. [22] independently developed an $\mathcal{O}(n \log n)$ approach. More recently, van der Berg et al. [11] and Duchi et al. [26] independently developed a linear time (i.e., $\mathcal{O}(n)$) approach. Although the latter

approach has a more attractive theoretical complexity, we observed from implementation that the former approach typically performs better.

We now present the two approaches mentioned above for solving $l_1$-norm projection problem (2.10).

**Algorithm 4**: $\mathcal{O}(n \log n)$ method

**if** $\|c\|_1 \leq \tau$ **then** return $x^* = c$

Initialize $p := [|c|;\ 0],\ s = 0$

$y = -\text{sort}(-\text{p})$

**for** $i = 1$ **to** size(y) **do**

  $s = s + y_i$

  $\varphi(y_i) = s - i \cdot y_i$

**end**

**for** $i = 1$ **to** size(y) **do**

  **if** $\varphi(y_i) > \tau$ **then**

    break

  **end**

**end**

$k \leftarrow i - 1$

$\delta = (\tau - \varphi(y_k))/k$

$\lambda = y_k - \delta$

return $x^* = \text{sign}(c) \cdot \max\{0, |c| - \lambda\}$

**Algorithm 5**: $\mathcal{O}(n)$ method

**if** $\|c\|_1 \leq \tau$ **then** return $x^* = c$

Initialize $p = [|c|;\ 0],\ s = 0$

**while** 1 **do**

$k = \lceil \text{length(p)}/2 \rceil$

$y_k = \text{upper} - \text{median(p)}$

$p_{\text{high}} = \text{find}(\text{p} > \text{y}_\text{k})$

$p_{\text{low}} = \text{find}(\text{p} < \text{y}_\text{k})$

$s_{\text{high}} = \|p_{\text{high}}\|_1 + y_k$

$\varphi(y_k) = s + s_{high} - k \cdot y_k$

**if** $\varphi(y_k) > \tau$ **then**

$p = p_{\text{high}}$

**else**

$y_{k+1} = \max(p_{\text{low}})$

$\varphi(y_{k+1}) = s + s_{\text{high}} - (k+1) \cdot y_{k+1}$

**if** $\varphi(y_{k+1}) > \tau$ **then**

break

**end**

$p = p_{\text{low}}$

$s = s + s_{\text{high}}$

$\lambda = y_k - (\tau - \varphi(y_k))/k$

**end**

return $x^* = \text{sign}(c) \cdot \max\{0, |c| - \lambda\}$

**end**

It shall be mentioned that the above two algorithms can be easily modified to solve the following problem:

$$\min_x \{\|c - x\|_2 :\ \|x\|_1 = \tau\}. \tag{2.14}$$

## 2.4 Matrix shrinkage threshold problem

In this section we consider the following matrix shrinkage threshold problem:

$$\min_{X \in \Re^{p \times q}} \frac{1}{2}\|C - X\|_F^2 + \lambda \|X\|_* \tag{2.15}$$

for some $C \in \Re^{p\times q}$ and $\lambda \geq 0$. Clearly, problem (2.15) can be viewed as a generalization of problem (2.11).

For convenience, we assume throughout this section that $p \leq q$. Let $C = U[\mathrm{Diag}(c)\ 0]V^T$ be the singular value decomposition of $C$, where $c \in \Re^p_+$ consists of the singular values of $C$, $U \in \Re^{p\times p}$ and $V \in \Re^{q\times q}$ are orthogonal matrices. We next show that the optimal solution of (2.15) is given by

$$X_C(\lambda) := U[\mathrm{Diag}(x_c(\lambda))\ 0]V^T, \tag{2.16}$$

where $x_c(\lambda)$ is defined in (2.12).

Before proceeding, we introduce two lemmas as follows. The proof of the first lemma can be found in [33].

**Lemma 2.4.1** *Let $A, B \in \Re^{p\times p}$ be symmetric matrices. Assume that $B$ is positive semidefinite. Then*

$$\lambda_i(A) \leq \lambda_i(A+B) \quad \textit{for all} \quad i = 1, 2, \ldots, p. \tag{2.17}$$

**Lemma 2.4.2** *Let $X = [X_1\ X_2]$, where $X_1 \in \Re^{p\times q_1}$ and $X_2 \in \Re^{p\times q_2}$. Then $\|X\|_* \geq \|X_1\|_*$.*

*Proof.* We first observe that

$$XX^T = X_1X_1^T + X_2X_2^T.$$

It follows from Lemma 2.4.1 that $\lambda_i(XX^T) \geq \lambda_i(X_1X_1^T)$ for all $i$, which together with the fact that $\lambda_i(YY^T) = \sigma_i^2(Y)$ for any matrix $Y$, implies that $\sigma_i(X) \geq \sigma_i(X_1)$ for all $i$. Using this inequality and the definition of $\|\cdot\|_*$, we see that the conclusion holds. ∎

**Theorem 2.4.3** *Given any $C \in \Re^{p\times q}$, the optimal solution of problem (2.15) is $X_C(\lambda)$ that is defined in (2.16).*

*Proof.* Let $U$, $V$ and $c$ be defined above, and let $D = \mathrm{Diag}(c)$. Upon performing change of variable by letting $X = U\tilde{X}V^T$ and using the fact that $C = U[D\ 0]V^T$, we see that problem (2.15) is equivalent to

$$\min_{\tilde{X}} \frac{1}{2}\|\tilde{X} - [D\ \ 0]\|_F^2 + \lambda\|\tilde{X}\|_* \tag{2.18}$$

We rewrite $\tilde{X}$ as $\tilde{X} = [\tilde{X}_1\ \tilde{X}_2]$, where $\tilde{X}_1 \in \Re^{p\times p}$ and $\tilde{X}_2 \in \Re^{p\times(q-p)}$. In view of Lemma 2.4.2, we observe that the optimal solution of (2.18) is $\tilde{X}^* = [\tilde{X}_1^*\ 0]$, where $\tilde{X}_1^*$ solves the following problem

$$\min_{\tilde{X}_1} \frac{1}{2}\|\tilde{X}_1 - D\|_F^2 + \lambda\|\tilde{X}_1\|_*. \tag{2.19}$$

We now claim that $\tilde{X}_1^*$ is a diagonal matrix. Given any index $i$, let $P \in \Re^{p\times p}$ be the diagonal matrix with all diagonal entries are 1 except that the $i$th one is $-1$. Let $\hat{X}_1^* = P\tilde{X}_1^*P$. It is easy to observe that

$$\|\tilde{X}_1^* - D\|_F = \|\hat{X}_1^* - D\|_F, \quad \|\hat{X}_1^*\|_* = \|\tilde{X}_1^*\|_*.$$

It implies that $\hat{X}_1^*$ is also the optimal solution of (2.18). Since problem (2.18) is strictly convex, it has a unique optimal solution. Hence, $\tilde{X}_1^* = \hat{X}_1^*$, which together with the definition of $\hat{X}_1^*$ implies that $\tilde{X}_1^*$ is diagonal. It then follows from (2.18), (2.11) and the relation $D = \mathrm{Diag}(c)$ that $\tilde{X}_1^* = \mathrm{Diag}(x_c(\lambda))$. Using this result, and the relations $X^* = U\tilde{X}^*V^T$ and $\tilde{X}^* = [\tilde{X}_1^*\ 0]$, we see that the conclusion holds. ∎

Before ending this section we now discuss a closely related problem to (2.15) as follows:

$$\min_{X\in\Re^{p\times q}} \|C - X\|_F :\ \|X\|_* \le \tau\} \tag{2.20}$$

for some $C \in \Re^{p\times q}$ and $\tau \ge 0$.

Let $C = U[\mathrm{Diag}(c)\ 0]V^T$ be the singular value decomposition of $C$, where $c \in \Re^p_+$ consists of the singular values of $C$, $U \in \Re^{p\times p}$ and $V \in \Re^{q\times q}$ are orthogonal matrices. Using the similar arguments as used in the proof Theorem 2.4.3, we can show that the optimal solution $X^*$ of problem (2.20) is given by

$$X^* := U[\mathrm{Diag}(x^*)\ 0]V^T,$$

where $x^*$ is the optimal solution of (2.10) that can be efficiently computed by the methods discussed in Section 2.3.

In addition, using similar arguments as above, we can see that the solution of the problem

$$\min_{X\in\Re^{p\times q}} \{\|C - X\|_F :\ \|X\|_* = \tau\} \tag{2.21}$$

is given by

$$X^* := U[\mathrm{Diag}(x^*)\ 0]V^T,$$

where $x^*$ is the optimal solution of (2.14) that can be efficiently computed.

# Chapter 3

# First-order methods for nuclear norm problems

In this chapter, we discuss in details how to apply the first-order methods reviewed in Chapter 2 to nuclear norm minimization problems (1.9). In Section 3.1 we first show that problem (1.9) is reducible to the case where $A$ has full column rank and also provide simplification for problem (1.10). In Section 3.2 we discuss the implementation details of FISTA and the NGM methods for solving problem (3.3). More specifically, in Subsection 3.2.1, we present the implementation details of first-order methods for solving problem (3.3) in primal form. In Subsection 3.2.2, we present the implementation details of first-order methods for solving problem (3.9) in dual form. In Section 3.3 we discuss the implementation details of FISTA and the NGM methods for solving problem (3.8).

## 3.1 Problem simplification

In Lu et al. [37], $A$ is assumed to be full column rank for problem (1.9). In this section we show that for the case where $A$ lacks full column rank, problem (1.9) can be reduced to the one with full column rank.

Suppose $A$ lacks full column rank, namely, $r := \text{rank}(A) < p$. Let

$$A = U \begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix} V^T,$$

where $U \in \Re^{n\times n}$, $V \in \Re^{p\times p}$ and $T \in \mathcal{D}^r_{++}$. We easily observe that

$$\|AX - B\|_F = \left\| U \begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix} V^T X - B \right\|_F = \left\| \begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix} V^T X - U^T B \right\|_F .$$

Let

$$\hat{X} = V^T X \qquad U^T B = \begin{bmatrix} H \\ \breve{H} \end{bmatrix}, \tag{3.1}$$

where $\hat{X} \in \Re^{p\times q}$, $H \in \Re^{r\times q}$ and $\breve{H} \in \Re^{(n-r)\times q}$. We further have

$$\|AX - B\|_F^2 = \|[T\ 0]\hat{X} - H\|_F^2 + \|\breve{H}\|_F^2, \quad \|\hat{X}\|_* = \|X\|_*.$$

It is easy to see that problem (1.9) is equivalent to the following problem:

$$\min_{\hat{X}\in\Re^{p\times q}} \left\{ \frac{1}{2}\|[T\ 0]\hat{X} - H\|_F^2 + \lambda\|\hat{X}\|_* \right\}. \tag{3.2}$$

Let $\hat{X}$ be partitioned into

$$\hat{X} = \begin{bmatrix} \tilde{X} \\ \breve{X} \end{bmatrix},$$

where $\tilde{X} \in \Re^{r\times q}$ and $\breve{X} \in \Re^{(p-r)\times q}$.

We next show that problem (3.2) can be reduced to the following problem:

$$\min_{\tilde{X}\in\Re^{r\times q}} \left\{ \frac{1}{2}\|T\tilde{X} - H\|_F^2 + \lambda\|\tilde{X}\|_* \right\}, \tag{3.3}$$

where $T$, $\tilde{X}$ and $H$ are defined above.

**Theorem 3.1.1** *$\tilde{X}^* \in \Re^{r\times q}$ optimizes problem (3.3) if and only if $\hat{X}^* = \begin{bmatrix} \tilde{X}^* \\ 0 \end{bmatrix} \in \Re^{p\times q}$ is an optimal solution of problem (3.2).*

*Proof.* We first show that problems (3.2) and (3.3) have the same optimal value. Clearly, the objective function of (3.3) minorizes that of (3.2). Thus, the optimal value of (3.3) is bounded above by that of (3.2). On the other hand, suppose $\tilde{X}^* \in \Re^{r\times q}$ is an optimal solution of (3.3). Let

$$\hat{X}^* = \begin{bmatrix} \tilde{X}^* \\ 0 \end{bmatrix}.$$

We easily see that the objective function of (3.2) achieves the optimal value of (3.3) at $\hat{X}^*$. Thus, the optimal value of (3.2) is majorized by that of (3.3). It immediately follows that the conclusion of this theorem holds. ∎

Given that problems (1.9) and (3.2) are equivalent, we see from Theorem 3.1.1 and (3.1) that an optimal solution of problem (1.9) can be obtained as follows. Let $\tilde{X}^*$ be the optimal solution of (3.3). Then

$$X^* = V \begin{bmatrix} \tilde{X}^* \\ 0 \end{bmatrix}$$

is an optimal solution of (1.9). Thus, for the remainder of this chapter, we will focus on how to efficiently solve problem (3.3) instead.

We next provide some bound on the optimal solution of problem (3.3).

**Lemma 3.1.2** *For every $\lambda > 0$, the optimal solution $X_\lambda^*$ for problem (3.3) is unique and bounded. In addition,*

$$\|X_\lambda^*\|_F \leq \|X_\lambda^*\|_* \leq r_x = \min\left\{\frac{\|H\|_F^2}{2\lambda}, \|T^{-1}H\|_*\right\}. \tag{3.4}$$

*Proof.* Using the fact that $T$ is a $r \times r$ positive diagonal matrix, it is easy to see that $\frac{1}{2}\|TX - H\|_F^2$ is strongly convex function with respect to $X$. Together with the fact that the nuclear norm $\|\cdot\|_*$ is convex, it follows that the objective function of problem (3.3) is strongly convex, for which we conclude that the optimal solution $X_\lambda^*$ of (3.3) is unique.

Considering the objective values of (3.3) at $X = 0$ and $X = T^{-1}H$, respectively, we obtain that for the optimal solution $X_\lambda^*$,

$$\lambda\|X_\lambda^*\|_* \leq \frac{1}{2}\|TX_\lambda^* - H\|_F^2 + \lambda\|X_\lambda^*\|_* = \frac{1}{2}\|H\|_F^2, \tag{3.5}$$

$$\lambda\|X_\lambda^*\|_* \leq \frac{1}{2}\|TX_\lambda^* - H\|_F^2 + \lambda\|X_\lambda^*\|_* = \lambda\|T^{-1}H\|_*. \tag{3.6}$$

Moreover, using the fact that

$$\|X\|_F^2 = \sum_{i=1}^{\min(r,q)} \sigma_i^2(X) \quad \forall\, X \in \Re^{r\times q},$$

we can easily see that

$$\|X\|_F = \sqrt{\sum_{i=1}^{\min(r,q)} \sigma_i^2(X)} \leq \sum_{i=1}^{\min(r,q)} \sigma_i(X) = \|X\|_*. \tag{3.7}$$

The relation (3.4) then immediately follows from (3.5), (3.6) and the above inequality. ∎

Before ending this section we provide a simplification for problem (1.10) that is closely related to problem (1.9). Indeed, using similar arguments as above, we can show that problem (1.10) can be reduced to

$$\begin{array}{ll} \min\limits_{X\in\Re^{r\times q}} & \frac{1}{2}\|TX-H\|_F^2 \\ \text{s.t.} & \|X\|_* \le \tau, \end{array} \tag{3.8}$$

where $T$ and $H$ are given above.

## 3.2 Implementation details of first-order methods for (3.3)

In this section we discuss the implementation details of the first-order methods reviewed in Chapter 2 for solving problem (3.3). More specifically, in Subsection 3.2.1, we present the implementation details of first-order methods for solving problem (3.3) in primal form. In Subsection 3.2.2, we present the implementation details of first-order methods for solving problem (3.9) in dual form.

### 3.2.1 First-order methods for the primal of (3.3)

In view of Lemma 3.1.2, we can easily observe that problem (3.3) is equivalent to the following problem:

$$\min_{X\in\mathcal{X}} \left\{ \frac{1}{2}\|TX-H\|_F^2 + \lambda\|X\|_* \right\}, \tag{3.9}$$

where $\mathcal{X} = \{X \in \Re^{r\times q} : \|X\|_* \le r_x\}$ and $r_x$ is given in (3.4).

We next discuss the implementation details of the first-order methods reviewed in Chapter 2, namely, FISTA and NGM methods, for solving problem (3.9).

#### FISTA for the primal of (3.9)

The implementation details of FISTA [10] for problem (3.9) are addressed as follows.

Let $f(X) = \frac{1}{2}\|TX-H\|_F^2$ and $\Psi(X) = \lambda\|X\|_*$. It is not hard to verify that $f(X)$ is continuously differentiable in $\Re^{r\times q}$ and moreover,

$$\nabla f(X) = T(TX-H) \tag{3.10}$$

is Lipschitz continuous on closed convex set $\mathcal{X}$ with Lipschitz constant $L_f = \|T\|^2$. Here $\|\cdot\|$ denotes the operator norm of a matrix. Therefore, the objective function of problem (3.9) satisfies the assumptions we described in Section 2.1. It immediately implies that FISTA can be applied to solve problem (3.9).

We are now ready to discuss the implementation details of FISTA. It is not hard to see that the main effort of FISTA per iteration lies in solving the subproblem in the form of

$$X^*_{\mathcal{X}} = \arg\min_{X\in\mathcal{X}} \left\{ f(Y) + \langle \nabla f(Y), X - Y\rangle + \frac{L_f}{2}\|X - Y\|_F^2 + \Psi(X) \right\} \tag{3.11}$$

for some $Y \in \Re^{r\times q}$. Simple algebra shows that (by dropping constant terms in $Y$)

$$X^*_{\mathcal{X}} = \arg\min_{X\in\mathcal{X}} \left\{ \frac{1}{2}\|C - X\|_F^2 + \frac{\lambda}{L_f}\|X\|_* \right\}, \tag{3.12}$$

where

$$C = Y - \frac{\nabla f(Y)}{L_f}.$$

We next discuss how to solve problem (3.12). Let $\hat{X}^*$ be the optimal solution of the following unconstrained optimization problem:

$$\hat{X}^* = \arg\min_{X\in\Re^{r\times q}} \left\{ \frac{1}{2}\|C - X\|_F^2 + \frac{\lambda}{L_f}\|X\|_* \right\}. \tag{3.13}$$

If $\|\hat{X}^*\| \le \tau$, then $X^*_{\mathcal{X}} = \hat{X}^*$. Otherwise, $X^*_{\mathcal{X}}$ is also an optimal solution of the problem (2.21) with $\tau = r_x$, which can be efficiently computed by the method discussed in Section 2.4.

We now address initialization and termination criterion for FISTA method. In particular, we choose the initial point $Y^1 = X^0 = 0$. In addition, we set parameter $\theta_1$ to be 1. Given an approximate solution $X^k \in \mathcal{X}$, we observe that

$$F(X) \ge F_k(X) := f(X^k) + \langle \nabla f(X^k), X - X^k\rangle + \frac{1}{2}\sigma\|X - X^k\|_F^2 + \lambda\|X\|_*$$

for all $X \in \Re^{r\times q}$, where $\sigma = \lambda^2_{\min}(T)$ that is the strong convexity parameter of $f$. Let $F_k^*$ be defined as follows:

$$F_k^* = \min\{F_k(X) : X \in \mathcal{X}\}.$$

Then, $F_k^*$ is a lower bound of the optimal value of problem (3.9). Given an accuracy parameter $\epsilon > 0$, we terminate FISTA method once

$$F(X^k) - F_k^* \le \epsilon \tag{3.14}$$

is satisfied.

The iteration complexity of FISTA for finding an $\epsilon$-optimal solution of problem (3.9) is established in the following theorem.

**Theorem 3.2.1** *Let $\{X^k\}_{k=1}^{\infty}$ be generated by FISTA algorithm when applied to problem (3.9). For a given $\epsilon > 0$, FISTA finds an $\epsilon$-optimal solution of problem (3.9) in a number of iterations which does not exceed*

$$\left\lceil \frac{\sqrt{2} r_x \|T\|}{\sqrt{\epsilon}} - 1 \right\rceil.$$

*Proof.* The conlcusion of this theorem follows from Theorem 2.2.1 with $L_f = \|T\|^2$, $X^0 = 0$, $\|X^*_{\mathcal{X}}\|_* \leq r_x$. ∎

**NGM methods for the primal of (3.9)**

The implementation details of NGM methods that were reviewed in Chapter 2 for problem (3.9) are addressed as follows.

Given an iterate $X^k$, the main effort of NGM methods lies in finding the search direction $d^k$. It can be computed by solving subproblem (2.7) with $H = \alpha_k^{-1} I$ for NGM methods, which becomes,

$$d^k := \arg\min_d \left\{ \langle \nabla f(X^k), d \rangle + \frac{1}{2\alpha_k} \|d\|_F^2 + \lambda \|X^k + d\|_* : \ X^k + d \in \mathcal{X} \right\}, \tag{3.15}$$

where $\nabla f(X^k)$ is computed by (3.10). Here, $\alpha_k > 0$ is chosen according to the scheme proposed by Barzilai and Borwein [7], which is also used by Birgin et al. [13] for studying a class of projected gradient methods. Let $0 < \alpha_{\min} < \alpha_{\max}$ be given. Initially, choose an arbitrary $\alpha_0 \in [\alpha_{\min}, \alpha_{\max}]$. Then, $\alpha_k$ is updated as follows:

$$\alpha_{k+1} = \begin{cases} \alpha_{\max}, & \text{if } b_k \leq 0; \\ \max\{\alpha_{\min}, \min\{\alpha_{\max}, a_k/b_k\}\}, & \text{otherwise,} \end{cases} \tag{3.16}$$

where $a_k = \|X^k - X^{k-1}\|_F^2$ and $b_k = \left\langle \nabla f(X^k) - \nabla f(X^{k-1}), X^k - X^{k-1} \right\rangle$. Upon performing change of variable by letting $X = X^k + d$, it is easy to see that problem (3.15) can be reduced to the one in the form of (3.11), and so (3.15) can be efficiently solved by the approach discussed above. In addition, the termination criterion for NGM method is the same as for FISTA method.

Note that when the NGM methods are applied to solve problem (3.9), multiple objective function evaluations are required to obtain a new iterate from a current iterate. Given that a nuclear norm needs to be computed for evaluating objective function, it is clear that multiple nuclear norm compuations are required between successive iterates. When problem size is relatively large, evaluation of nuclear norm can be very expensive. To avoid this drawback, we propose another reformualtion of problem (3.9) and then apply NGM methods to solve it.

It is easy to see that problem (3.9) is equivalent to

$$\begin{array}{ll} \min\limits_{t,X} & \frac{1}{2}\|TX - H\|_F^2 + \lambda t \\ \text{s.t.} & \|X\|_* \leq t \\ & X \in \mathcal{X}. \end{array} \tag{3.17}$$

Letting $Z = (X, t)$, $f(Z) = \frac{1}{2}\|TX - H\|_F^2$ , $\Psi(Z) = \lambda t$ and

$$\mathcal{Z} := \{(X, t) : \|X\|_* \leq t, X \in \mathcal{X}\},$$

we see that problem (3.17) can be rewritten as

$$\min_{Z \in \mathcal{Z}} \{f(Z) + \Psi(Z)\}. \tag{3.18}$$

It is clear that the NGM methods can be applied to solve problem (3.18). Moreover, the search direction $d^k$ used in these methods for (3.18) can be essentially computed by solving

$$\min_{X \in \mathcal{X}, \|X\|_* \leq t} \left\{ \langle \nabla f(X^k), X - X^k \rangle + \frac{1}{2\alpha_k}\|X - X^k\|_F^2 + \lambda t \right\}. \tag{3.19}$$

Using the similar arguments as used in Subsection 2.2.2, problem (3.19) is equivalent to the following problem:

$$\min_{X \in \mathcal{X}, \|X\|_* \leq t} \left\{ \frac{1}{2}(\|C^k - X\|_F^2 + \lambda \alpha_k t \right\}, \tag{3.20}$$

where

$$C^k = X^k - \alpha_k \nabla f(X^k).$$

It is not hard to see that the optimal solution of problem (3.19) can be given by $t^* = \|X_{\mathcal{X}}^*\|_*$ and $\|X_{\mathcal{X}}^*\|_*$ is the optimal solution of problem (3.11) that can be efficiently computed (see Chapter 2).

We next establish a convergence result for the NGM methods when applied to problem (3.9) or (3.17).

**Theorem 3.2.2** *Let $\{X^k\}_{k=0}^{\infty}$ be generated by the NGM methods when applied to problem (3.9) or (3.17). Then any accumulation point of $\{X^k\}_{k=0}^{\infty}$ is the optimal solution of problem (3.9).*

*Proof.* Recall that problem (3.17) is equivalent to (3.9). The conclusion of this theorem immediately follows from Theorems 2.2.2 and 2.2.3, and the fact that problems (3.9) and (3.17) are convex. ∎

### 3.2.2 First-order methods for the dual of (3.3)

In this subsection, we consider first-order methods for solving the dual of problem (3.3) and show that the latter problem can be simultaneously solved.

Before proceeding, we first introduce the following notations. Let

$$\mathcal{U} = \left\{U \in R^{r\times q} : \|U\| \le 1\right\} \tag{3.21}$$

and

$$\mathcal{B}_F^{r\times q}(t) := \left\{X \in \Re^{r\times q} : \|X\|_F \le t\right\}. \tag{3.22}$$

**Theorem 3.2.3** *Given any $\epsilon \ge 0$, let $X_\epsilon$ be an $\epsilon$-optimal solution of the following smooth saddle point problem*

$$\min_{X\in\mathcal{B}_F^{p\times q}(r_x)} \max_{U\in\mathcal{U}} \left\{\frac{1}{2}\|TX - H\|_F^2 + \lambda\langle X, U\rangle\right\}, \tag{3.23}$$

*where $\mathcal{U}$ and $r_x$ are defined in (3.21) and (3.4), respectively. Then, $X_\epsilon$ is an $\epsilon$-optimal solution of problem (3.3).*

*Proof.* First, note that the nuclear norm $\|\cdot\|_*$ is the dual norm of the operator norm $\|\cdot\|$, namely,

$$\|X\|_* = \max_{\|U\|\le 1} \langle X, U\rangle$$

(e.g., see Fazel et al. [27] and Boyd and Vandenberghe [9]). Then, the conclusion of this theorem immediately follows from the above relation and Lemma 3.1.2. ∎

From the above theorem, we see that problem (3.9) can be solved as problem (3.23). Clearly, the dual of the latter problem is as follows:

$$\max_{U\in\mathcal{U}} \min_{X\in\mathcal{B}_F^{p\times q}(r_x)} \left\{\frac{1}{2}\|TX - H\|_F^2 + \lambda\langle X, U\rangle\right\}. \tag{3.24}$$

For the purpose of our implementation, we reformulate problem (3.24) into the problem

$$\min_{U\in\mathcal{U}} \max_{X\in\mathcal{B}_F^{r\times q}(1)} \left\{ \phi(X,U) := -\frac{1}{2}\|r_x TX - H\|_F^2 - r_x\lambda\langle X, U\rangle \right\} \tag{3.25}$$

obtained by scaling the variable $X$ of (3.24) as $X \leftarrow X/r_x$, and multiplying the resulting formulation by $-1$. In the remaining part of this subsection, we will consider problem (3.25) rather than (3.24).

For convenience of notation, let $\Omega = \mathcal{B}_F^{r\times q}(1)$ and

$$\min_{U\in\mathcal{U}} \left\{ G(U) := \max_{X\in\Omega} \phi(X,U) \right\}. \tag{3.26}$$

Observing that $\phi(X,U) : \Omega \times \mathcal{U} \to \Re$ is a smooth function which is strictly concave in $X$ for every fixed $U \in \mathcal{U}$, and convex in $U$ for every fixed $X \in \Omega$, we can conclude that (i) problem (3.26) and its dual, that is,

$$\max_{X\in\Omega} \left\{ F(X) := \min_{U\in\mathcal{U}} \phi(X,U) \right\}, \tag{3.27}$$

are both solvable and have the same optimal values; and (ii) the function $G(\cdot)$ is convex differentiable and its gradient is given by

$$\nabla G(U) = \nabla_U \phi(X(U), U) \quad \forall U \in \mathcal{U},$$

where

$$X(U) := \arg\max_{X\in\Omega} \phi(X,U).$$

The following result shows that an approximate solution of problem (3.27) can be obtained by solving smooth convex minimization problem (3.26). Its proof is similar to that of Theorem 2.4 of Lu [37].

**Proposition 3.2.4** *Let $X^*$ be the unique optimal solution of problem (3.27), and let $F^*$ be the optimal value of problems (3.27) and (3.26). Suppose that the sequence $\{U^k\}_{k=0}^{\infty} \subseteq \mathcal{U}$ is such that $G(U^k) \to F^*$ as $k \to \infty$. Then, $X(U^k) \to X^*$ and $G(U^k) - F(X(U^k)) \to 0$ as $k \to \infty$.*

From the above proposition, we easily see that problem (3.3) can be solved simultaneously while solving problem (3.26). Indeed, suppose that $\{U^k\}_{k=0}^{\infty} \subseteq \mathcal{U}$ is a sequence of approximate solutions obtained by solving (3.26). It follows from the above proposition

that for any given $\epsilon$, there exists some iterate $U^k$ such that $G(U^k) - F(X(U^k)) \leq \epsilon$. Clearly, $X(U^k)$ is an $\epsilon$-optimal solution of (3.27) and hence (3.3).

We next discuss the first-order methods, that is, FISTA and NGM methods for solving problem (3.26). Before proceeding, we state a proposition that will be used to compute the Lipschitz constant of $\nabla G(U)$, where $G$ is defined in problem (3.26). Its proof can be found in Theorem 1 of [44].

**Proposition 3.2.5** *Let $\mathscr{U}$ and $\mathscr{X}$ be normed vector spaces with the respective norms denoted by $\|\cdot\|_{\mathscr{U}}$ and $\|\cdot\|_{\mathscr{X}}$. Assume that $\phi(X,U) : \Omega \times \mathcal{U} \to \Re$ has the form*

$$\phi(X,U) = \langle U, \mathcal{E}X\rangle - h(X) + \theta(U), \quad \forall (X,U) \in \Omega \times \mathcal{U}, \tag{3.28}$$

*where $\mathcal{E} : \mathscr{X} \to \mathscr{U}^*$ is a linear map, $h : \Omega \to \Re$ is a differentiable strongly convex function with modulus $\sigma_h > 0$ with respect to $\|\cdot\|_{\mathscr{X}}$, and $\theta : \mathcal{U} \to \Re$ is $L_\theta$-Lipschitz-differentiable function with respect to $\|\cdot\|_{\mathscr{U}}$. Then, the function $G(U)$ defined by (3.26) is $(L_\theta + \|\mathcal{E}\|^2_{\mathscr{X},\mathscr{U}}/\sigma_h)$-Lipschitz-differentiable in $\mathcal{U}$ with respect to $\|\cdot\|_{\mathscr{U}}$.*

We next show how to compute the gradient and its associated Lipschitz constant of $G(U)$ defined in (3.26). Indeed, the function $\phi$ of problem (3.25) is of the form (3.28) with $\theta \equiv 0$ and the functions $\mathcal{E}$ and $h$ given by

$$\begin{array}{rl} \mathcal{E}X & = -\lambda r_x X, \quad \forall X \in \Omega, \\ h(X) & = \frac{1}{2}\|r_x TX - H\|_F^2, \ \forall X \in \Omega. \end{array} \tag{3.29}$$

Moreover, in the context of problem (3.26), $\mathscr{X}$ and $\mathscr{U}$ are specified as $\Re^{r\times q}$, and their associated norms $\|\cdot\|_{\mathscr{X}}$ and $\|\cdot\|_{\mathscr{U}}$ are Frobenius norm. Clearly, for any $X \in \Omega$, we have

$$\nabla^2 h[\Delta X, \Delta X] \ = \ r_x^2 \|T\Delta X\|_F^2 \ \geq \ r_x^2 \|\Delta X\|_F^2 / \|T^{-1}\|^2 \quad \forall \Delta X \in \Re^{r\times q},$$

which immediately implies that $h$ is strongly convex with modulus $\sigma_h = r_x^2/\|T^{-1}\|^2$ with respect to $\|\cdot\|_{\mathscr{X}}$. Now, using (1.11), we obtain

$$\begin{array}{rl} \|\mathcal{E}\|_{\mathscr{X},\mathscr{U}} & = \max\left\{\|\lambda r_x X\|_{\mathscr{U}}^* : X \in \mathscr{X}, \|X\|_{\mathscr{X}} \leq 1\right\}, \\ & = \lambda r_x \max\left\{\|X\|_F : X \in \mathscr{X}, \|X\|_F \leq 1\right\}, \\ & = \lambda r_x. \end{array} \tag{3.30}$$

Hence, by Proposition 3.2.5, we conclude that

$$L = \|\mathcal{E}\|^2_{\mathcal{U},\Omega}/\sigma_h = \lambda^2 \|T^{-1}\|^2.$$

The gradient of $G$ is given by

$$\nabla G(U) = \nabla_U \phi(X(U), U) = -\lambda r_x X(U), \tag{3.31}$$

where

$$X(U) = \arg\max_{X \in \Omega} \left\{ -\frac{1}{2}\|r_x TX - H\|_F^2 - \lambda r_x \langle X, U \rangle \right\}. \tag{3.32}$$

We can easily see that $X(U)$ is the optimal solution of the following equivalent problem:

$$\min_{X \in \Omega} \left\{ \frac{1}{2}\|r_x TX - H\|_F^2 + \lambda r_x \langle X, U \rangle \right\}. \tag{3.33}$$

We now discuss some of the computational technicalities to find the optimal solution $X(U)$ of problem (3.33). For any $\rho \geq 0$, let

$$\begin{array}{ll} v(\rho) & = (r_x^2 T^2 + 2\rho I)^{-1}(r_x TH - r_x W), \\ \varphi(\rho) & = \|v(\rho)\|_F^2 - 1. \end{array} \tag{3.34}$$

If $\varphi(0) \leq 0$, then clearly $v(0)$ is the optimal solution of problem (3.33). Otherwise, the optimal solution of problem (3.33) is equal to $v(\rho^*)$ where $\rho^*$ is the root of the equation $\varphi(\rho) = 0$. The latter can be found by the well-known root finding schemes specially tailored for solving the above equations. For example, Newton's method can be suitably applied to find the aforementioned root by choosing $\rho = 0$ as the initial guess and generating a sequence of $\rho_k$ according to

$$\rho_{k+1} = \rho_k - \frac{\varphi(\rho_k)}{\varphi'(\rho_k)} \quad k \geq 0.$$

In addition, when applied to problem (3.26), the FISTA and NGM methods require solving the subproblem in the form of

$$\min_{U \in \mathcal{U}} \frac{1}{2}\|C - U\|_F^2 \tag{3.35}$$

for some $C \in \Re^{r \times q}$, where $\mathcal{U}$ is given in (3.21). Without loss of generality, suppose that the singular value decomposition of $C$ is given by

$$C = \hat{U} \left[ \mathrm{Diag}(c) \; 0 \right] \hat{V}^T,$$

where $\hat{U} \in \Re^{r \times r}$, $\hat{V} \in \Re^{q \times q}$ and $c \in \Re^r_+$. Thus the optimal solution $U^*$ of problem (3.35) is given as follows:

$$U^* = \hat{U} \left[ \mathrm{Diag}(w) \; 0 \right] \hat{V}^T,$$

where $w \in \Re^r_+$ is defined by

$$w_i = \begin{cases} c_i & \text{if } c_i \in [0,1], \\ 1 & \text{otherwise,} \end{cases} \quad \forall i = 1, \dots r.$$

The convergence result for the FISTA method when applied to the dual of problem (3.3) is given as follows.

**Theorem 3.2.6** *Let $\{U^k\}_{k=0}^{\infty}$ be generated by FISTA algorithm with initial point $U^0 = 0$. For a given $\epsilon > 0$, FISTA finds an $\epsilon$-optimal solution of (3.26) and its dual (3.27), and hence of problem (3.3), in a number of iterations which does not exceed*

$$\left\lceil \frac{\sqrt{2m}\lambda\|T^{-1}\|}{\sqrt{\epsilon}} - 1 \right\rceil, \tag{3.36}$$

*where $m := \min\{r, q\}$.*

*Proof.* In view of Theorem 2.2.1, we see that $G(U^k) - G(U^*) \leq \epsilon$ when

$$k \geq \sqrt{\frac{2L}{\epsilon}}\|U^0 - U^*\|_F - 1.$$

Since $U^* \in \mathcal{U}$, we see from (3.21) that $\|U^*\| \leq 1$. Thus, we have

$$\|U^*\|_F = \sqrt{\mathrm{Tr}(U^{*T}U^*)} \leq \sqrt{\min\{r,q\}\lambda_{\max}(U^{*T}U^*)} = \sqrt{m}\|U^*\| \leq \sqrt{m}.$$

The above observation together with the fact that $\|U^0\|_F = 0$ and $L = \lambda^2\|T^{-1}\|^2$ implies that the conclusion of this theorem holds. ∎

*Remark.* The practical termination criterion for FISTA is based on the dulity gap $G(U^k) - F(X(U^k))$. In addition, it shall be mentioned that a variant of Nestrov's smooth method (VNSM) was applied to solve another dual formulation of problem (3.3) by Lu et al. [37]. They showed that an $\epsilon$-optimal solution of (3.3) can be found in at most

$$\left\lceil \frac{2\lambda\sqrt{2m\log(n/m)}\|T^{-1}\|}{\sqrt{\epsilon}} - 1 \right\rceil \tag{3.37}$$

number of iterations, where $n := r + q$. Thus, we improve their iteration complexity by a factor of $\sqrt{\log(n/m)}$ while retaining a similar iteration cost. ∎

We next establish a convergence result for the NGM methods when applied to solve problems (3.26) and (3.27).

**Theorem 3.2.7** *Let $\epsilon > 0$ be given. The NGM methods generate a pair of $\epsilon$-optimal solutions $(U^k, X(U^k))$ to problems (3.26) and (3.27) in finite number of iterations.*

*Proof.* Suppose by contradiction that the NGM methods do not terminate. Then they generate a sequence $\{U^k\}_{k=0}^{\infty} \subseteq \mathcal{U}$ satisfying $G(U^k) - F(X(U^k)) > \epsilon$. Note that $G(\cdot)$ is convex, which together with Theorem 2.2.2 (2.2.3) implies that any accumulation point of $\{U^k\}_{k=0}^{\infty}$ is an optimal solution of problem (3.26). By the continuity of $G(\cdot)$, it further implies that any accumulation point of $\{G(U^k)\}_{k=0}^{\infty}$ is the optimal value $F^*$ of (3.26). Using this observation and the fact that $\{G(U^k)\}_{k=0}^{\infty}$ is bounded, we conclude that $G(U^k) \to F^*$ as $k \to \infty$. Further, in view of Proposition 3.2.4, we have $G(U^k) - F(X(U^k)) \to 0$ as $k \to \infty$, and arrive at a contradiction. Therefore, the conclusion of this theorem holds. ∎

## 3.3 Implementation details of first-order methods for (3.8)

In this section, we discuss the implementation details of first-order methods for solving problem (3.8).

Let $f(X) = \frac{1}{2}\|TX - H\|_F^2$ and $\Psi(X) = 0$. As discussed in Subsection 3.2.1, $f(X)$ is continuously differentiable in $\Re^{r\times q}$ and moreover, $\nabla f(X)$ is Lipschitz continuous on the closed convex set $\{X \in \Re^{r\times q} : \|X\|_* \le \tau\}$ with Lipschitz constant $L_f = \|T\|^2$. Therefore, problem (3.8) satisfies the assumptions imposed on problem (2.1) described in Section (2.1). It implies that the FISTA and NGM methods can be suitably applied to solve problem (3.8). In addition, when applied to problem (3.8), the FISTA and NGM methods require solving the subproblem in the form of

$$\begin{array}{ll} \min & \frac{1}{2}\|C - X\|_F^2 \\ \text{s.t.} & \|X\|_* \le \tau \end{array} \tag{3.38}$$

for some $C \in \Re^{r\times q}$ and $\tau \ge 0$, which can be efficiently solved by the approach described in Section 2.4.

The convergence result for the FISTA method when applied to problem (3.8) is given as follows.

**Theorem 3.3.1** *Let $\{X^k\}_{k=0}^{\infty}$ be generated by FISTA algorithm when applied to (3.8) with initial point $X^0 = 0$. For a given $\epsilon$, FISTA finds an $\epsilon$-optimal solution of problem (3.8) in*

*a number of iterations which does not exceed*

$$\left\lceil \frac{\sqrt{2}\|T\|\tau}{\sqrt{\epsilon}} - 1 \right\rceil. \tag{3.39}$$

*Proof.* In view of Theorem 2.2.1, we see that $F(X^k) - F(X^*) \leq \epsilon$ when

$$k \geq \sqrt{\frac{2L}{\epsilon}}\|X^0 - X^*\|_F - 1.$$

Since $\|X^*\|_* \leq \tau$, we see from (3.7) that $\|X^*\|_F \leq \tau$. The above observation together with the fact that $\|X^0\|_F = 0$ and $L = \|T\|^2$ implies that the conclusion of this theorem holds. ∎

We next establish a convergence result for the NGM methods when applied to problem (3.8).

**Theorem 3.3.2** *Let $\{X^k\}_{k=0}^{\infty}$ be generated by the NGM methods when applied to problem (3.8). Then any accumulation point of $\{X^k\}_{k=0}^{\infty}$ is the optimal solution of problem (3.8).*

*Proof.* The conclusion of this theorem immediately follows from Theorems 2.2.2 and 2.2.3, and the fact that problem (3.8) is convex. ∎

# Chapter 4

# Numerical results

In this chapter we report the results of our experiment which compare the performance of FISTA method with NGM methods discussed in Sections 3.2 and 3.3 for solving problems (3.3) and (3.8) on a set of randomly generated instances. In Section 4.1, we compare the performance of NGM I method based on formulation (3.17) with FISTA and NGM methods based on formulation (3.9). In Section 4.2 we compare the performance of NGM I method with FISTA method for solving (3.26). In Section 4.3 we compare the performance of NGM I method and FISTA method for solving problem (3.8).

The random instances for problems (3.3) and (3.8) used in our experiment were generated as follows. We first randomly generated matrices $A \in \Re^{n\times p}$ and $B \in \Re^{n\times q}$, where $p = 2q$ and $n = 10q$, with entries uniformly distributed in $[0, 1]$ for different values of $q$. We then computed $T$ and $H$ for (3.3) and (3.8) according to the procedures described in section (3.1). In addition, all computations were performed on an Intel Xeon 5320 CPU (1.86GHz) and 12GB RAM running Red Hat Enterprise Linux 4 (kernel 2.6.9) and by Matlab V7.0 (R14). We use the default random seed in Matlab to generate all datasets and the CPU time reported in this section does not include the time for inputing the data and outputing the results.

## 4.1 Numerical results on the primal of problem (3.3)

In this section, we first evaluate the performance of NGM I, NGM II for solving problem (3.9) and NGM I for solving problem (3.17). Then we compare the performance of FISTA method for solving problem (3.9) and NGM I method for solving problem (3.17).

We choose initial point $X^0 = 0$ for FISTA and NGM methods. And for NGM methods we set $\alpha^0 = 1$, $\gamma = 1e-4$, $M = 2$, $\alpha_{\max} = 1$, $\alpha_{\min} = 10^{-15}$. We use (3.14) as our termination criterion with $\epsilon = 10^{-1}$.

Table 4.1: Comparison of NGM methods for $\lambda = 1$

| Problem | Iter | | | Time | | | Nf | | |
|---|---|---|---|---|---|---|---|---|---|
| (p,q) | NGMC | NGM I | NGM II | NGMC | NGM I | NGM II | NGMC | NGM I | NGM II |
| (200,100) | 50 | 50 | 50 | 2.7 | 4.1 | 5.3 | 56 | 105 | 164 |
| (400,200) | 60 | 60 | 60 | 24.4 | 35.0 | 51.1 | 67 | 126 | 203 |
| (600,300) | 50 | 50 | 60 | 77.5 | 108.0 | 205.2 | 56 | 105 | 211 |
| (800,400) | 70 | 70 | 60 | 255.7 | 361.2 | 511.7 | 78 | 147 | 218 |
| (1000,500) | 60 | 60 | 60 | 422.5 | 597.1 | 1005.2.7 | 67 | 126 | 221 |

Table 4.2: Comparison of NGM methods for $\lambda = 10$

| Problem | Iter | | | Time | | | Nf | | |
|---|---|---|---|---|---|---|---|---|---|
| (p,q) | NGMC | NGM I | NGM II | NGMC | NGM I | NGM II | NGMC | NGM I | NGM II |
| (200,100) | 60 | 60 | 40 | 3.4 | 5.3 | 5.4 | 67 | 126 | 160 |
| (400,200) | 80 | 80 | 80 | 32.6 | 47.6 | 73.5 | 89 | 168 | 290 |
| (600,300) | 80 | 80 | 80 | 122.3 | 172.2 | 281.0 | 89 | 168 | 291 |
| (800,400) | 90 | 90 | 90 | 322.1 | 461.2 | 755.8 | 100 | 189 | 327 |
| (1000,500) | 90 | 90 | 90 | 625.9 | 891.1 | 1471.4 | 100 | 189 | 328 |

Table 4.3: Comparison of NGM methods for $\lambda = 100$

| Problem | Iter | | | Time | | | Nf | | |
|---|---|---|---|---|---|---|---|---|---|
| (p,q) | NGMC | NGM I | NGM II | NGMC | NGM I | NGM II | NGMC | NGM I | NGM II |
| (200,100) | 50 | 60 | 30 | 2.8 | 5.1 | 4.1 | 56 | 126 | 118 |
| (400,200) | 190 | 160 | 190 | 69.0 | 88.6 | 156.1 | 210 | 336 | 692 |
| (600,300) | 300 | 300 | 320 | 352.8 | 535.2 | 844.9 | 331 | 630 | 1158 |
| (800,400) | 540 | 490 | 330 | 1491.6 | 2085.6 | 2094.3 | 595 | 1029 | 1211 |
| (1000,500) | 380 | 800 | 650 | 2066.6 | 6672.9 | 8017.6 | 419 | 1680 | 2337 |

For convenience of presentation, we use NGMC to label NGM I method for solving problem (3.17). The performance of NGM methods for the randomly generated instances with parameters $\lambda = 1$, 10 and 100 is presented in Tables 4.1-4.3, respectively. The problem size $(p, q)$ is given in column one. The number of iterations of NGM I, NGM II and NGMC are given in columns two, three and four. CPU times (in seconds) are given in columns four, five and six. The number of function evaluations (labeled as Nf) is given in the last three columns, respectively.

From Tables 4.1-4.3, we can see that the number of iterations of NGM methods varies greatly with the parameter $\lambda$. In addition, NGMC generally outperforms NGM I and NGM II. We also see that NGM I substantially outperforms NGM II in general. This behavior

is consistent with the fact that NGM I has less function evaluations (mainly, nuclear norm evaluations) than NGM II. Thus, we will only focus on NGM I method rather than NGM II method in the remaining numerical tests.

Since NGMC method generally outperforms NGM I and NGM II, we next aim to compare NGMC with FISTA for the primal of problem (3.3). We apply the same termination criterion (3.14) and set an upper bound of 5 hours computation time (or 18,000 seconds) per instance for both methods. The performance of NGMC and FISTA for our randomly generated instances with parameters $\lambda = 1$ and $\lambda = 500$ are presented in Tables 4.4-4.5, respectively. As seen from the tables, NGMC can solve all instances for $\lambda = 1$ within 5 hours while FISTA is only able to solve the instances with $q \leq 400$. On the other hand, when $\lambda = 500$ and $q \geq 900$, both methods cannot terminate within 5 hours. In addition, NGMC substantially outperforms FISTA.

Table 4.4: Comparison of NGMC and FISTA for $\lambda = 1$

| Problem | Iter | | Time | | Obj | |
|---|---|---|---|---|---|---|
| (p,q) | NGMC | FISTA | NGMC | FISTA | NGMC | FISTA |
| (200,100) | 80 | 590 | 4.5 | 29.0 | 45.27 | 45.31 |
| (400,200) | 80 | 1270 | 32.8 | 475.0 | 91.28 | 91.33 |
| (600,300) | 90 | 1970 | 138.0 | 2804.2 | 137.45 | 137.49 |
| (800,400) | 80 | 2370 | 292.5 | 9280.9 | 183.63 | 183.68 |
| (1000,500) | 100 | 2749 | 701.4 | 18001.6 | 229.19 | 229.29 |
| (1200,600) | 90 | 1620 | 1076.5 | 18004.5 | 275.24 | 276.67 |
| (1400,700) | 90 | 1041 | 1677.7 | 18012.1 | 321.62 | 332.31 |
| (1600,800) | 90 | 702 | 2465.4 | 18009.3 | 367.38 | 428.82 |
| (1800,900) | 100 | 497 | 3841.3 | 18025.7 | 413.62 | 688.86 |
| (2000,1000) | 120 | 366 | 6305.8 | 18037.1 | 459.36 | 1608.14 |

## 4.2 Numerical results on the dual of problem (3.3)

In this section we report the computational results of NGM I and FISTA for solving the dual of problem (3.3), namely, problem (3.26). We also compare their performance when applied to the primal and dual of problem (3.3). The parameters for FISTA and NGM I are the same as those given in Section 4.1. In addition, we use $G(U^k) - F(X(U^k)) \leq \epsilon$ as the termination criterion for NGM I and FISTA with $\epsilon = 10^{-1}$.

For convenience of presentation, we use NGMD to label NGM I method for solving

Table 4.5: Comparison of NGMC and FISTA for $\lambda = 500$

| Problem | Iter | | Time | | Obj | |
|---|---|---|---|---|---|---|
| (p,q) | NGMC | FISTA | NGMC | FISTA | NGMC | FISTA |
| (200,100) | 20 | 30 | 1.2 | 1.5 | 1173.52 | 1173.51 |
| (400,200) | 30 | 50 | 13.5 | 19.7 | 3656.33 | 3656.37 |
| (600,300) | 70 | 80 | 109.5 | 115.8 | 7851.82 | 7851.80 |
| (800,400) | 110 | 100 | 408.8 | 348.0 | 13676.02 | 13676.09 |
| (1000,500) | 180 | 220 | 1276.4 | 1454.3 | 21120.07 | 21120.12 |
| (1200,600) | 180 | 270 | 2146.4 | 3006.5 | 30302.61 | 30302.63 |
| (1400,700) | 200 | 410 | 3763.8 | 7078.3 | 41195.78 | 41195.78 |
| (1600,800) | 350 | 540 | 9422.7 | 13554.5 | 53603.97 | 53603.97 |
| (1800,900) | 510 | 547 | 18030.8 | 18030.3 | 67694.38 | 67694.52 |
| (2000,1000) | 400 | 439 | 18015.9 | 18005.5 | 83038.61 | 83037.75 |

Table 4.6: Comparison of NGMD and FISTA for $\lambda = 1$

| Problem | Iter | | Time | | Obj | |
|---|---|---|---|---|---|---|
| (p,q) | NGMD | FISTA | NGMD | FISTA | NGMD | FISTA |
| (200,100) | 10 | 10 | 0.8 | 0.8 | 45.26 | 45.26 |
| (400,200) | 10 | 10 | 4.5 | 5.1 | 91.28 | 91.28 |
| (600,300) | 10 | 10 | 16.1 | 18.1 | 137.44 | 137.44 |
| (800,400) | 10 | 10 | 38.7 | 42.2 | 183.63 | 183.63 |
| (1000,500) | 10 | 10 | 100.3 | 80.9 | 229.18 | 229.18 |
| (1200,600) | 10 | 10 | 124.7 | 136.3 | 275.23 | 275.23 |
| (1400,700) | 10 | 10 | 231.2 | 213.3 | 321.62 | 321.62 |
| (1600,800) | 10 | 10 | 284.7 | 314.9 | 367.38 | 367.38 |
| (1800,900) | 10 | 10 | 394.6 | 435.3 | 413.62 | 413.62 |
| (2000,1000) | 10 | 10 | 534.4 | 592.6 | 459.36 | 459.36 |

problem (3.26). The performance of NGMD and FISTA for our randomly generated instances with parameters $\lambda = 1$ and $\lambda = 500$ is presented in Tables 4.6 and 4.7, respectively. As observed from these tables, NGMD generally outperforms FISTA except for the case $q = 500$ and 700 when $\lambda = 1$. In addition, NGMD is nearly two times faster than FISTA when $\lambda = 500$. We also observe that a large $\lambda$ often leads to a slower rate of convergence for FISTA and NGMD. Finally, in view of Tables 4.4-4.7, NGMD and FISTA generally outperform the corresponding ones that are tailored for solving the primal of problem (3.3).

Table 4.7: Comparison of NGMD and FISTA for $\lambda = 500$

| Problem | Iter | | Time | | Obj | |
|---|---|---|---|---|---|---|
| (p,q) | NGMD | FISTA | NGMD | FISTA | NGMD | FISTA |
| (200,100) | 40 | 70 | 2.5 | 4.4 | 1173.52 | 1173.54 |
| (400,200) | 40 | 70 | 16.7 | 30.3 | 3656.33 | 3656.40 |
| (600,300) | 30 | 70 | 44.2 | 108.5 | 7851.83 | 7851.88 |
| (800,400) | 30 | 80 | 103.8 | 291.4 | 13676.03 | 13676.06 |
| (1000,500) | 30 | 80 | 201.5 | 556.8 | 21120.07 | 21120.10 |
| (1200,600) | 30 | 80 | 333.5 | 935.7 | 30302.57 | 30302.63 |
| (1400,700) | 20 | 80 | 358.7 | 1453.8 | 41195.82 | 41195.78 |
| (1600,800) | 30 | 80 | 770.4 | 2165.5 | 53603.93 | 53603.98 |
| (1800,900) | 30 | 80 | 1094.7 | 3056.6 | 67694.10 | 67694.15 |
| (2000,1000) | 30 | 80 | 1446.9 | 4115.0 | 83035.63 | 83035.68 |

## 4.3 Numerical results on problem (3.8)

In this section we report the computational results of NGM I and FISTA for solving problem (3.8). We choose initial point $X^0 = 0$ for FISTA and NGM I. And for NGM I we set $\alpha^0 = 1$, $\gamma = 1e-4$, $M = 2$, $\alpha_{\max} = 1$, $\alpha_{\min} = 10^{-15}$. We use (3.14) as our termination criterion with $\epsilon = 10^{-1}$.

Table 4.8: Comparison of FISTA and NGM I for $\tau = 1$

| Problem | Iter | | Obj | | Time | |
|---|---|---|---|---|---|---|
| (p,q) | FISTA | NGM I | FISTA | NGM I | FISTA | NGM I |
| (200,100) | 70 | 100 | 801.60 | 801.61 | 3.1 | 5.2 |
| (400,200) | 170 | 140 | 3260.56 | 3260.56 | 54.7 | 52.9 |
| (600,300) | 210 | 270 | 7435.46 | 7435.43 | 251.9 | 379.9 |
| (800,400) | 250 | 230 | 13240.67 | 13240.64 | 750.0 | 786.4 |
| (1000,500) | 350 | 380 | 20666.34 | 20666.34 | 2011.7 | 2527.3 |
| (1200,600) | 390 | 370 | 29832.24 | 29832.22 | 3783.4 | 2449.7 |
| (1400,700) | 450 | 360 | 40708.45 | 40708.44 | 7126.0 | 6800.8 |
| (1600,800) | 490 | 520 | 53104.42 | 53104.40 | 11032.5 | 13744.6 |
| (1800,900) | 540 | 380 | 68287.06 | 68287.06 | 17636.6 | 14382.2 |
| (2000,1000) | 570 | 470 | 83049.24 | 83049.24 | 24827.1 | 23272.1 |

The results of NGM I and FISTA for (3.8) with $\tau = 1$ and 10 are presented in Tables 4.8 and 4.9, respectively. We observe from these tables that NGM I produces smaller function values and has better CPU time than FISTA. Thus, we conclude that the method NGM I generally outperforms FISTA when applied to problem (3.8). We also see that as $\tau$ increases,

Table 4.9: Comparison of FISTA and NGM I for $\tau = 10$

| Problem | Iter | | Obj | | Time | |
|---|---|---|---|---|---|---|
| (p,q) | FISTA | NGM I | FISTA | NGM I | FISTA | NGM I |
| (200,100) | 230 | 110 | 453.21 | 453.22 | 8.63 | 5.36 |
| (400,200) | 330 | 130 | 2448.80 | 2448.80 | 87.1 | 43.9 |
| (600,300) | 490 | 200 | 6126.44 | 6126.43 | 469.4 | 247.2 |
| (800,400) | 580 | 430 | 11422.86 | 11422.84 | 1410.7 | 1259.2 |
| (1000,500) | 650 | 420 | 18322.86 | 18322.85 | 3086.8 | 2449.1 |
| (1200,600) | 710 | 490 | 26973.12 | 26973.10 | 5789.5 | 4944.5 |
| (1400,700) | 770 | 610 | 37315.98 | 37315.98 | 10466.4 | 10184.8 |
| (1600,800) | 820 | 520 | 49183.59 | 49183.58 | 16069.7 | 11948.2 |
| (1800,900) | 880 | 500 | 62828.01 | 62828.01 | 24714.8 | 16957.1 |
| (2000,1000) | 1101 | 780 | 78048.23 | 78048.23 | 40860.0 | 34126.1 |

the performance of NGM I and FISTA become worse.

# Chapter 5

# Concluding Remarks

In this thesis, we first provide some preliminaries for solving nuclear norm minimization problems (1.9) and (1.10). Further, we provide problem simplification and apply first-order methods for the resulting problem and its closely related problem. The performance of these methods is then compared using a set of randomly generated instances. We show that first-order methods typically have better performance when applied to solve the dual counterpart of the problems. Moreover, NGM I method [38] generally outperforms the FISTA method [10]. Finally, we shall mention that the methods discussed in this thesis can be extended to solve more general problems in the form of (1.6) and (1.7).

# Bibliography


[1] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. *A new approach to collaborative filtering: Operator estimation with spectral regularization.* Journal of Machine Learning Research, 10:803-826, 2009.

[2] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. *Low-rank matrix factorization with attributes.* Technical Report N24/06/MM, Ecole des Mines de Paris, 2006.

[3] Y. Amit, M. Fink, N. Srebro, and S. Ullman. *Uncovering shared structures in multiclass classification.* In Proceedings of the Twenty-fourth International Conference on Machine Learning, 2007.

[4] A. Argyrious and M. Pontil. *Multi-task feature learning.* In Neural Information Proceeding Systems, 2007.

[5] A. Argyrious, T. Evgeniou, and M. Pontil. *Convex multi-task feature learning.* Machine Learning, 73(3):243-272, 2008.

[6] F. R. Bach. *Consistency of trace norm minimization.* Journal of Machine Learning Research, 8:1019-1048, 2008.

[7] J. Barzilai and J. M. Borwein. *Two point stepsize gradient methods.* IMA J. Numer. Anal., 8:141-148, 1988.

[8] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization.* Springer-Verlag, 2003.

[9] S. Boyd and L. Vandenbergh. *Convex Optimization.* Cambridge University Press, 2004.

[10] A. Beck and M. Teboulle. *A fast iterative shrinkage-thresholding algorithm for linear inverse problems.* To appear in SIAM Journal on Imaging Sciences, 2008.

[11] E. V. D. Berg, M. Schmidt, M. P. Frielander, and K. Murphy. *Group sparsity via linear-time projection.* Technical Report TR-2008-09, Department of Computer Science, University of British Columbia, June 2008.

[12] E. V. D. Berg and M. P. Friedlander. *In pursuit of a root.* Technical Report TR-2007-19, Department of Computer Science, University of British Columbia, June 2007.

[13] E. G. Birgin, J. M. Martines, and M. Raydan. *Nonmontone spectral projected gradients methods on convex sets.* SIAM Journal on Optimization, 10:1196-1211, 2000.

[14] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, New York, Second Edition, 1999.

[15] J. F. Cai, E. J. Candès and Z. Shen. *A singular value thresholding algorithm for matrix completion.* Technical Report 08-77, UCLA Computational and Applied Math, 2008.

[16] E. J. Candès and B. Recht. *The power of convex relaxation: Near-optimal matrix completion.* Preprint, 2009.

[17] E. J. Candès and B. Recht. *Exact matrix completion via convex optimization.* Technical Report 08-77, UCLA Computational and Applied Math, 2008.

[18] E. J. Candès and J. Romberg. *Practiacal signal recovery from random projections.* In Computational Imaging III, Proc. SPIE Conf., 5914:76-86, March 2005.

[19] E. J. Candès and J. Romberg. *Quantitative robust uncertainty principles and optimally sparse decompositions.* Found. Comput. Math., 6(2):227-254, June 2006.

[20] E. J. Candès, J. Romberg, and T. Tao *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information.* IEEE Trans. Info. Theory, 52:280-509, 2006.

[21] E. J. Candès and T. Tao. *Nearly optimal signal recovery from random projections: Universal encoding strategies.* IEEE Trans. Info. Theory, 52:5406-5425, 2006.

[22] I. Daubechies, M. Fonasier, and I. Loris. *Accelerated projected gradient method for linear inverse problems with sparsity constraints.* J. Fourier Anal. Appl., to appear, 2007.

[23] S. Chen, D. Donoho, and M. Saunders. *Atomic decomposition by basis pursuit.* SIAM on Scientific Computing, 20(1):31-61, 1998.

[24] D. Donoho. *Compressed sensing.* IEEE Trans. Info. Theory, 52:1289-1306, 2006.

[25] D. Donoho and J. Tanner. *Neighborliness of randomly projected simplices in high dimensions.* Proc. Natl. Acad. Sci. USA, 102:9452-9457, 2005

[26] J. Douchi, S. van de Geer, and P. Buhlmann *The group lasso for logistic regression.* Journal of the Royal Statistical Society: Series B, 70:53-71, 2008.

[27] M. Fazel, H. Hindi, and S. P. Boyd *A rank minimization heuristic with application to minimum order system approximation.* In proceedings American Control Conference, 6:4734-4739, 2001.

[28] M. Fazel, H. Hindi, and S. P. Boyd. *Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices.* In proceedings American Control Conference, pp. 2156-2162, 2003.

[29] M. Figueiredo, R. Nowak, and S. J. Wright. *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems.* IEEE journal on Selected Topics in Signal Processing, 1:586-598, 2007.

[30] L. Grippo, F. Lampariello, and S. Lucidi. *A nonmonotone linesearch technique for Newton's method.* SIAM Journal on Numerical Analysis, 23(4):707-716, 1986.

[31] E. T. Hale, W. Yin, and Y. Zhang. *A fixed-point continuation method for $\ell_1$ regularized minimization with applications to compressed sensing.* CAAM Technical Report TR07-07, Department of Computational and Applied Mathematics, Rice University, July 2007.

[32] E. T. Hale, W. Yin, and Y. Zhang. *Fixed-point continuation for l1-minimization: Methodology and convergence.* SIAM Journal on Optimization, 19(3):1107-1130, 2008a.

[33] R. A .Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1990.

[34] S. Ji and J. Ye. *An accelerated gradient method for trace norm minimization.* In ICML, 2009.

[35] Y. J. Liu, D. F. Sun, and K.-C. Toh *An implementable proximal point algorithmic framework for nuclear norm minimization.* Preprint, National University of Singapore, July 2009.

[36] Z. Lu. *Smooth optimization approach for covariance selection.* SIAM Journal on Optimization, 19(4):1807-1827, 2009.

[37] Z. Lu, R. D. C. Monteiro, and M. Yuan. *Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression.* Submitted to Mathematical Programming, January 2008 (revised : March 2009).

[38] Z. Lu and Y. Zhang. *An augmented lagrangian approach for sparse principal component analysis.* Submitted, July 2009.

[39] S. Ma, D. Goldfarb, and L. Chen. *Fixed point and Bregman iterative methods for matrix rank minimization.* Technical Report 08-78, UCLA Computational and Applied Math, 2008.

[40] B. Martinet. *Régularisation d'inéquations variationnelles par approximations successives.* Rev. Francaise Informat. Recherche Opérationnelle, 4:154-158, 1970.

[41] M. Mesbahi, G. P. Papavassilopoulos. *On the rank minimization problem over a positive semidefinite linear matrix inequality.* IEEE transactions on Automatic Control, 42(2):239-243, 1997.

[42] Y. E. Nesterov. *A method for unconstrained convex minimization problem with the rate of convergence* $O(1/k^2)$. Doklady AN SSSR, 269:543-547, 1983. Translated as Soviet Math. Docl.

[43] Y. E. Nesterov. *Excessive gap technique in nonsmooth convex minimization.* SIAM Journal on Optimization, 16:235–249, 2005.

[44] Y. E. Nesterov. *Smooth minimization of nonsmooth functions.* Mathematical Programming, 103:127-152, 2005.

[45] Y. E. Nesterov. *Gradient methods for minimizing composite objective function.* Technical Report 2007/76, CORE, Université catholique de Louvain, 2007.

[46] Y. E. Nesterov and A. S. Nemirovski. *Interior Point Polynomial Algorithms in Convex Programming: Theory and Applications.* Society for Industrial and Applied Mathematics, Philadelphia, 1994.

[47] J. Nocedal and S. J. Wright. *Numerical Optimization, second edition.* Springer Series in Operations Research, Springer, New York, 2006.

[48] B. T. Polyak. *Introduction to Optimization.* Translations Series in Mathematics and Engineering, Optimization Software, Publications Division, New York, 1987.

[49] T. K. Pong, T. Paul, S. W. Ji, and J. P. Ye. *Trace norm regularization: Reformulations, algorithms and multi-task learning.* Submitted to SIAM Journal on Optimization, June 2009.

[50] B. Recht, M. Fazel, and P. A. Parrilo. *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization.* In Technical Report arXiv:0706.4138v1, arXiv, 2007.

[51] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[52] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. *Maximum-margin Matrix factorization.* In Advances in Neural Information Processing Systems 17, 2005.

[53] J. F. Sturm. *Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones.* Optimization Methods and Software 11-12, 625-653, 1999.

[54] R. Tibshirani. *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B, 58:267-288, 1996.

[55] T.-C. Toh and S. Yun. *An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems.* Technical Report, Department of Mathematics, National University of Singapore, Singapore, 2009.

[56] C. Tomasi and T. Kanade. *Shape and motion from image streams under orthography: a factorization method.* International Journal of Computer Vision, 9(2):137-154, 1992.

[57] P. Tseng. *On accelerated proximal gradient algorithm for convex-concave optimization.* Submitted to SIAM Journal on Optimization, 2008.

[58] P. Tseng and S. Yun. *A coordinate gradient descent method for nonsmooth separable minimization.* Mathematical Programming: Series B, 117:387-423, 2009.

[59] R. H. Tütüncü, K.-C. Toh, and M. J. Todd. *Solving semidefinite-quadratic-linear programs using SDPT3.* Mathematical Programming: Series B, 95:189-217, 2003.

[60] S. J. Wright, R. Nowak, and M. P. Figueiredo. *Sparse reconstruction by separable approximation.* Report, Computer Science Department, University of Wisconsin, Madison, October 2007.

[61] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro. *Dimension reduction and coefficient estimation in multivariate linear regression.* Journal of the Royal Statistical Society: Series B, 69(3):329-346, 2007.

[62] S. Yun and K.-C. Toh. *A coordinate gradient descent method for $l_1$-regularized convex minimization.* Comput. Optim. Appl., accepted, October 2009.