SMOOTHING APPROXIMATIONS FOR TWO CLASSES OF CONVEX EIGENVALUE OPTIMIZATION PROBLEMS

YU QI

(B.Sc.(Hons.), BUAA)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF MATHEMATICS

NATIONAL UNIVERSITY OF SINGAPORE

2005

Acknowledgements

I am very grateful to my supervisor, Dr Sun Defeng, for his abundant knowledge and kindness. In the past two years, Dr Sun has always helped me when I was in trouble, encouraged me when I lost confidence. This thesis would not have been finished without the invaluable suggestions and patient guidance from him.

I would also like to thank Miss. Shi Shengyuan, who has helped me so much in my study. Thanks all the staff and friends at National University of Singapore for their help and support.

Yu Qi

August 2005

Summary

In order to solve a convex non-differentiable optimization problem, one may introduce a smoothing convex function to approximate the non-differentiable objective function and solve the smoothing convex optimization problem to get an approximate solution. Nesterov showed that if the gradient of the smoothing function is Lipschitz continuous, one may get an ϵ-approximation solution with the number of iterations bounded by $O(\frac{1}{\epsilon})$ [9]. In [11], Nesterov discussed the problem of minimizing the maximal eigenvalue and the problem of minimizing the spectral radius. Recently, Shi [12] presented a smoothing function for the sum of the κ largest components of a vector. This smoothing function is highly advantageous because the composition of this function and eigenvalue functions allows one to compute smoothing functions to approximate all eigenvalue functions of a real symmetric matrix.

In this thesis, we further study the properties of the smoothing functions to approximate the eigenvalue functions. In particular, we obtain an estimation of the Lipschitz constant of the gradient of these smoothing functions.

We then consider the problem of minimizing the sum of the κ largest eigenvalues by applying Nesterov's smoothing method [9]. Finally, we extend this algorithm to solve the problem of minimizing the sum of the κ largest absolute values of eigenvalues. These two problems are general forms of the problem of minimizing the maximal eigenvalue and the problem of minimizing the spectral radius, considered respectively in [11]. We also report some numerical results for both problems.

The organization of this thesis is as follows. In Chapter 1 we first describe the problems discussed in [11] by Nesterov and then extend them to general cases. In Chapters 2 and 3 we discuss some important properties of smoothing functions for approximating the sum of the κ largest eigenvalues and the sum of the κth largest absolute values of eigenvalues of a parametric affine operator, respectively. The smoothing algorithm and computational results are given in Chapter 4.

List of Notation

- $A, B, \ldots$ denote matrices; $M_{n,m}$ denotes the n-by-m matrix.
- $\mathcal{S}_m$ is the set of all $m \times m$ real symmetric matrices; $\mathcal{O}_m$ is the set of all $m \times m$ orthogonal matrices.
- A superscript "T" represents the transpose of matrices and vectors.
- For a matrix M, $M_{i\cdot}$ and $M_{\cdot j}$ represent the ith row and jth column of M, respectively. M_{ij} denotes the (i, j)th entry of M.
- A diagonal matrix is written as $\mathrm{Diag}(\beta_1, \ldots, \beta_n)$.
- We use $\circ$ to denote the Hadamard product between matrices, i.e.

$$X \circ Y = [X_{ij} Y_{ij}]_{i,j=1}^{m}.$$

- Let $A_1, \ldots, A_n \in \mathcal{S}_m$ be given, we define the linear operator $\mathcal{A} : \mathbb{R}^n \to \mathcal{S}_m$ by

$$\mathcal{A}(x) := \sum_{i=1}^{n} x_i A_i, \ \forall x \in \mathbb{R}^n. \tag{1}$$

- Let $\mathcal{A}^* : \mathcal{S}_m \to \mathbb{R}^n$ be the adjoint of the linear operator $\mathcal{A} : \mathbb{R}^n \to \mathcal{S}^m$ defined by (1):

$$\langle d, \mathcal{A}^* D\rangle = \langle D, \mathcal{A}d\rangle,\ \forall (d, D) \in \mathbb{R}^n \times \mathcal{S}_m.$$

  Hence, for all $D \in \mathcal{S}_m$,

$$\mathcal{A}^* D = (\langle A_1, D\rangle, \ldots, \langle A_n, D\rangle)^T.$$

  Denote $G \in S_m$ as $(G)_{ij} = \langle A_i, A_j\rangle$.

- The eigenvalues of $X \in \mathcal{S}_m$ are designated by $\lambda_i(X)$, $i = 1, \ldots, m$, and

$$\lambda_1(X) \geq \lambda_2(X) \geq \cdots \geq \lambda_m(X).$$

- We write $X = O(\alpha)$ (respectively, $o(\alpha)$) if $\|X\|/|\alpha|$ is uniformly bounded (respectively, tends to zero) as $\alpha \to 0$.

- For $B \in \mathcal{M}_{n,m}$, the operator $\Xi : \mathcal{M}_{n,m} \to \mathcal{S}_{m+n}$ is defined as

$$\Xi(B) = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}.$$

- We define the linear operator $\Gamma : \mathbb{R}_n \to \mathcal{S}_{2m}$ as

$$\Gamma(x) = \Xi(\mathcal{A}(x)). \tag{2}$$

- Let $\Gamma^* : \mathcal{S}_{2m} \to \mathbb{R}_n$ be the adjoint operator of the linear operator Γ, for any $Y \in \mathcal{S}_{2m}$,

$$\Gamma^*(Y) = (2\langle A_1, Y_2\rangle, \cdots, 2\langle A_n, Y_2\rangle)^T,$$

  where

$$Y = \begin{bmatrix} Y_1 & Y_2 \\ Y_2^T & Y_3 \end{bmatrix}, \text{ and } Y_1, Y_2, Y_3 \in \mathcal{S}_m.$$

Contents

Chapter 1

Introduction

Let $\mathcal{S}_m$ be the set of real m-by-m symmetric matrices. For $X \in \mathcal{S}_m$, let $\{\lambda_i(X)\}_{i=1}^m$ be the eigenvalues of X which are sorted in nonincreasing order, i.e.

$$\lambda_1(X) \geq \lambda_2(X) \geq \cdots \geq \lambda_\kappa(X) \geq \cdots \geq \lambda_m(X).$$

Let $A_1, \ldots, A_n \in \mathcal{S}_m$ be given. Define the operator $\mathcal{A} : \mathbb{R}^n \to \mathcal{S}_m$ by

$$\mathcal{A}(x) = \sum_{i=1}^n x_i A_i, \quad x \in \mathbb{R}^n. \tag{1.1}$$

Let $\mathcal{Q}$ be a bounded closed convex set in $\mathbb{R}^n$ and $C \in \mathcal{S}_m$. In [11], Nesterov considered the following nonsmooth problem:

$$\min_{x \in \mathcal{Q}} \lambda_1(C + \mathcal{A}(x)). \tag{1.2}$$

Nesterov's approach is to replace the above nonsmooth function $\lambda_1(C + \mathcal{A}(x))$ by the following smooth function:

$$S_\mu(C + \mathcal{A}(x)), \tag{1.3}$$

where the tolerance parameter $\mu > 0$, and $S_\mu(X)$ is the product of the entropy function and μ, i.e.

$$S_\mu(X) = \mu \ln[\sum_{i=1}^n e^{\lambda_i(X)/\mu}]. \tag{1.4}$$

The function $S_\mu(X)$ approximates $\lambda_1(X)$ as $\mu \downarrow 0$. Now, let us consider the following smoothed optimization problem:

$$\min_{x \in \mathcal{Q}} \{S_\mu(C + \mathcal{A}(x))\}. \tag{1.5}$$

Note that for any $\mu > 0$, the gradient mapping $\nabla S_\mu(\cdot)$ is globally Lipschitz continuous. Nesterov suggested to use a gradient based numerical method [9] to solve problem (1.5).

For a given matrix $X \in \mathcal{S}_m$, we define its spectral radius by:

$$\rho(X) := \max_{1 \le i \le m} |\lambda_i(X)| = \max\{\lambda_1(X), -\lambda_m(X)\}. \tag{1.6}$$

Let

$$\varphi(x) := \rho(C + \mathcal{A}(x)).$$

Another problem discussed in Nesterov's paper [11] is

$$\min_{x \in \mathcal{Q}} \varphi(x) \tag{1.7}$$

with $C \equiv 0$. Nesterov constructed the smoothing function as

$$\varphi_p(x) = F_p(\mathcal{A}(x)), \ \forall x \in \mathcal{Q},$$

where $F_p(\cdot)$ is:

$$F_p(X) = \frac{1}{2}\langle X^{2p}, I_n \rangle^{\frac{1}{p}}. \tag{1.8}$$

Nesterov considered the smoothing problem

$$\min_{x \in \mathcal{Q}} \{\varphi_p(x)\} \tag{1.9}$$

and used method [8] to solve the smoothing problem (1.9).

In this thesis, we shall extend Nesterov's approach to the following two problems. Let $\phi(x)$ be the sum of κth largest eigenvalues of $C + \mathcal{A}(x)$, i.e.

$$\phi(x) = \sum_{i=1}^{\kappa} \lambda_i(C + \mathcal{A}(x)). \tag{1.10}$$

We shall first consider in this thesis the following nonsmooth problem:

$$\min_{x\in\mathcal{Q}} \phi(x). \tag{1.11}$$

Clearly, if $\kappa = 1$, (1.11) turns to be problem (1.2).

Let $|\lambda|^{[\kappa]}(X)$ be the κth largest absolute value of eigenvalues of X, sorted in the nonincreasing order, i.e.

$$|\lambda|^{[1]}(X) \geq |\lambda|^{[2]}(X) \geq \cdots \geq |\lambda|^{[\kappa]}(X) \geq \cdots \geq |\lambda|^{[n]}(X).$$

We define the following function:

$$\psi(x) = \sum_{i=1}^{\kappa} |\lambda|^{[i]}(C + \mathcal{A}(x)). \tag{1.12}$$

The second problem that we shall consider in this thesis is

$$\min_{x\in\mathcal{Q}} \psi(x), \tag{1.13}$$

which is a general case of (1.7).

We shall construct smoothing functions for problem (1.11) and (1.13) respectively. The gradients of the smoothing functions must satisfy the global Lipschitz condition, which makes it possible for us to apply Nesterov's algorithm [9] to solve the smoothing problems.

Nesterov's method is to solve the optimization problem:

$$\min_{x\in Q} \theta(x), \tag{1.14}$$

where $\theta(\cdot)$ is a nonsmooth convex function. Our goal is to find an $\epsilon-$solution $\bar{x} \in \mathcal{Q}$, i.e.

$$\theta(\bar{x}) - \theta^* \leq \epsilon, \tag{1.15}$$

where $\theta^* = \min\limits_{x\in Q} \theta(x)$. We denote $\theta_\mu(\cdot)$ as a smoothing function for $\theta(\cdot)$. The smoothing function $\theta_\mu(\cdot)$ satisfies the following inequality:

$$\theta_\mu(x) \leq \theta(x) \leq \theta_\mu(x) + \mu R, \tag{1.16}$$

where R is a constant for a specified smoothing function $\theta_\mu(x)$ (We will give its definition in Chapter 2). Nesterov proved the following inequality [9, Theorem 2]

$$\theta(\bar{x}) - \theta^* \leq \theta_\mu(\bar{x}) - {\theta_\mu}^* + \mu R,$$

where $\theta_\mu^* = \min_{x \in Q} \theta_\mu(x)$. Let $\mu = \mu(\epsilon) = \frac{\varepsilon}{2R}$ and

$$\theta_\mu(\bar{x}) - \theta_\mu^* \leq \frac{1}{2}\epsilon. \tag{1.17}$$

We have

$$\theta(\bar{x}) - \theta^* \leq \epsilon.$$

Nesterov's algorithm can improve the bound on the number of iterations to $O(\frac{1}{\epsilon})$, while the traditional algorithms need $O(\frac{1}{\epsilon^2})$ iterations.

The remaing part of this thesis is as follows. We discuss the properties of smoothing functions for approximating the sum of the κ largest eigenvalues of a parametric affine operator in Chapter 2 and the sum of the κ largest absolute values in Chapter 3. Computational results are given in Chapter 4.

Chapter 2

Properties of the Smoothing Function for the Sum of the κ Largest Eigenvalues

2.1 The Smoothing Function for the Sum of κ Largest Components

In her thesis [12], Shi discussed the smoothing function for the κ largest components of a vector. For $x \in \mathbb{R}^n$ we denote by $x^{[\kappa]}$ the κth largest component of x, i.e.,

$$x^{[1]} \geq x^{[2]} \geq \cdots \geq x^{[\kappa]} \geq \cdots \geq x^{[n]}$$

sorted in the nonincreasing order. Define

$$f_\kappa(x) = \sum_{i=1}^{\kappa} x^{[i]}. \tag{2.1}$$

Denote by $\mathcal{Q}_\kappa$ the convex set in $\mathbb{R}^n$:

$$\mathcal{Q}_\kappa = \{v \in \mathbb{R}^n : \sum_{i=1}^{n} v_i = \kappa, 0 \leq v_i \leq 1,\ i = 1, 2, \ldots, n\}, \tag{2.2}$$

and

$$p(z) = \begin{cases} z \ln z, & z \in (0, 1], \\ 0, & z = 0. \end{cases} \tag{2.3}$$

Let

$$r(v) = \sum_{i=1}^{n} p(v_i) + \sum_{i=1}^{n} p(1 - v_i) + R, \forall v \in \mathcal{Q}_\kappa, \tag{2.4}$$

where

$$R := n \ln n - \kappa \ln \kappa - (n - \kappa) \ln(n - \kappa), \tag{2.5}$$

which is the maximal value of $r(v)$.

The smoothing function for $f_\kappa(\cdot)$ is $f_\kappa^\mu(\cdot) : \mathbb{R}^n \to \mathbb{R}$, which is defined by:

$$\begin{aligned} f_\kappa^\mu(x) := \quad & \max \quad x^T v - \mu r(v) \\ & \text{s.t.} \quad \sum_{i=1}^{n} v_i = \kappa \\ & \qquad\;\; 0 \leq v_i \leq 1, \; i = 1, \ldots, n. \end{aligned} \tag{2.6}$$

Shi has also provided the optimal solution to $f_\kappa^\mu(\cdot)$ in (2.6):

$$v_i(\mu, x) = \frac{1}{1 + e^{\frac{\alpha(\mu,x) - x_i}{\mu}}} \tag{2.7}$$

where α satisfies

$$\sum_{i=1}^{n} \frac{1}{1 + e^{\frac{\alpha(\mu,x) - x_i}{\mu}}} = \kappa. \tag{2.8}$$

In order to introduce Lemma 2.1 in [12], we need the definition of the γ function:

$$\gamma_i(\mu, x) := \frac{e^{\frac{\alpha(\mu,x) - x_i}{\mu}}}{\mu(1 + e^{\frac{\alpha(\mu,x) - x_i}{\mu}})^2}, \quad i = 1, \cdots, n. \tag{2.9}$$

Without causing any confusion, let $\gamma_i := \gamma_i(\mu, x)$, for $i = 1, \ldots, n$, and $\gamma = (\gamma_1, \ldots, \gamma_n)^T$.

Lemma 2.1. *The optimal solution to problem (2.6), $v(\mu, x)$, is continuously differentiable on $\mathbb{R}_{++} \times \mathbb{R}^n$, with*

$$\nabla_x v(\mu, x) = \begin{pmatrix} \gamma_1 & & \\ & \ddots & \\ & & \gamma_n \end{pmatrix} - \frac{1}{\sum\limits_{k=1}^{n}(\gamma_k)} \gamma(\gamma)^T. \tag{2.10}$$

Proof. From (2.7), for each $i = 1, \cdots, n$,

$$v_i(\mu, x)(1 + e^{\frac{\alpha(\mu,x)-x_i}{\mu}}) = 1. \tag{2.11}$$

Taking derivatives of x on both side of (2.11), we have

$$\begin{aligned} (\nabla_x v(\mu, x))_i &= \frac{1}{\mu(1 + e^{\frac{\alpha(\mu,x)-x_i}{\mu}})^2} e^{\frac{\alpha(\mu,x)-x_i}{\mu}} (e_i - \nabla_x \alpha(\mu, x)) \\ &= \gamma_i(e_i - \nabla_x \alpha(\mu, x)). \end{aligned} \tag{2.12}$$

From (2.8), we have

$$\begin{aligned} \nabla_x \alpha(\mu, x) &= \frac{1}{\sum\limits_{k=1}^{n}(\gamma_k)} \sum_{i=1}^{n} (\gamma_i e_i) \\ &= \frac{1}{\sum\limits_{k=1}^{n}(\gamma_k)} (\gamma_1, \cdots, \gamma_n)^T, \end{aligned} \tag{2.13}$$

where $e_i \in \mathbb{R}^n$, its ith entry is 1 and others are all zeros. From (2.12) and (2.13), we obtain:

$$(\nabla_x v(\mu, x))_i = e_i \gamma_i - \frac{\gamma_i}{\sum\limits_{k=1}^{n}(\gamma_k)} (\gamma_1, \cdots, \gamma_n)^T. \tag{2.14}$$

□

Lemma 2.2. *For γ_i, $i = 1, \ldots, n$ given by (2.9), has bounds:*

$$0 \leq \gamma_i \leq \frac{1}{4\mu}. \tag{2.15}$$

Proof. Denote

$$p_i(\mu, x) = e^{\frac{\alpha(\mu,x)-x_i}{\mu}}, \ i = 1, \ldots, n.$$

Thus we have

$$\begin{aligned}
\gamma_i &= \frac{1}{\mu} \cdot \frac{p_i}{(1+p_i)^2} \\
&= \frac{1}{\mu}\left(\frac{1}{1+p_i} - \frac{1}{(1+p_i)^2}\right) \\
&= \frac{1}{\mu}\left(-\left(\frac{1}{1+p_i} - \frac{1}{2}\right)^2 + \frac{1}{4}\right) \\
&\leq \frac{1}{4\mu}.
\end{aligned} \tag{2.16}$$

□

The following theorem from [12] describes some properties of $f_\kappa^\mu(x)$.

Theorem 2.3. *For $\mu > 0, \forall x \in \mathbb{R}^n$, the function $f_\kappa^\mu(x)$ has the following properties:*

1. *$f_\kappa^\mu(x)$ is convex;*
2. *$f_\kappa^\mu(x)$ is continuously differentiable;*
3. *$f_\kappa^\mu(x) \leq f_\kappa(x) \leq f_\kappa^\mu(x) + \mu R$.*

From the above theorem, for each $\mu > 0$, $f_\kappa^\mu(x)$ is continuously differentiable. According to [12], the gradient of $f_\kappa^\mu(x)$ is the optimal solution to problem (2.6), i.e., $v(\mu, x)$. Therefore the gradient and Hessian of $f_\kappa^\mu(x)$ for any $\mu > 0$ are given by

$$\nabla f_\kappa^\mu(x) = v(\mu, x); \tag{2.17}$$

$$\nabla^2 f_\kappa^\mu(x) = \nabla_x v(\mu, x). \tag{2.18}$$

Up to now we have reviewed some of Shi's results. Based on these results, we provide an estimate to $\nabla^2 f_\kappa^\mu(x)$ in the following theorem.

Theorem 2.4. *For $\mu > 0$, we have the following conclusions on $\nabla^2 f_\kappa^\mu(x)$:*

1. *For $h \in \mathbb{R}^n$,*

$$0 \le \langle h, \nabla^2(f_\kappa^\mu(x))h\rangle \le \frac{1}{4\mu}\|h\|_2^2. \tag{2.19}$$

2. *$((\nabla^2(f_\kappa^\mu(x)))_{ij}$ is the (i,j)th entry of $(\nabla^2(f_\kappa^\mu(x)))$, for $i \neq j$*

$$0 \le (\nabla^2(f_\kappa^\mu(x)))_{ii} \le \frac{1}{4\mu}; \tag{2.20}$$

$$-\frac{1}{4\mu} \le (\nabla^2(f_\kappa^\mu(x)))_{ij} \le 0. \tag{2.21}$$

Proof. First we prove part 1. Since $f_\kappa^\mu(x)$ is convex, its Hessian $\nabla^2 f_\kappa^\mu(x)$ is positive semidefinite, i.e.,

$$\langle h, \nabla_x f_\kappa^\mu(x)h\rangle \ge 0.$$

By Lemma 2.1,

$$\begin{aligned}
\langle h, \nabla^2 f_\kappa^\mu(x)h\rangle &= \langle h, \nabla_x v(\mu, x)h\rangle \\
&= \sum_{i=1}^n \gamma_i h_i^2 - \frac{1}{\sum\limits_{k=1}^n \gamma_k}(\gamma^T h)^2 \\
&\le \sum_{i=1}^n \gamma_i h_i^2 \\
&\le \sum_{i=1}^n \frac{1}{4\mu} h_i^2 \\
&= \frac{1}{4\mu}\|h\|_2^2.
\end{aligned} \tag{2.22}$$

Now let us prove part 2. For any $1 \le i \le n$, $1 \le j \le n$, $i \neq j$,

$$\begin{aligned}
(\nabla^2(f_\kappa^\mu(x)))_{ii} &= (\nabla_x v(\mu, x))_{ii} \\
&= \gamma_i - \frac{(\gamma_i)^2}{\sum\limits_{k=1}^n \gamma_k} \\
&= \gamma_i \left(1 - \frac{\gamma_i}{\sum\limits_{k=1}^n \gamma_k}\right).
\end{aligned} \tag{2.23}$$

According to Lemma 2.2, we have

$$0 \le \gamma_i \le \frac{1}{4\mu},$$

and

$$0 \le \left(1 - \frac{\gamma_i}{\sum_{k=1}^{n} \gamma_k}\right) \le 1.$$

Therefore,

$$0 \le (\nabla^2(f_\kappa^\mu(x)))_{ii} \le \frac{1}{4\mu}. \tag{2.24}$$

On the other hand,

$$\begin{aligned}(\nabla^2(f_\kappa^\mu(x)))_{ij} &= (\nabla_x v(\mu, x))_{ij} \\ &= -\frac{\gamma_i \gamma_j}{\sum_{k=1}^{n} \gamma_k} \\ &= -\gamma_i \frac{\gamma_j}{\sum_{j=1}^{n} \gamma_j}.\end{aligned} \tag{2.25}$$

Since

$$0 \le \frac{\gamma_j}{\sum_{j=1}^{n} \gamma_j} \le 1, \tag{2.26}$$

we have

$$-\frac{1}{4\mu} \le (\nabla^2(f_\kappa^\mu(x)))_{ij} \le 0, \quad i \ne j. \tag{2.27}$$

□

Remark. Nesterov proved in [9, Theorem 1] that for problem (1.14), if the function $\theta^\mu(\cdot)$ $(\mu > 0)$ is continuously differentiable, then its gradient

$$\nabla \theta^\mu(x) = \mathcal{A}^* v^\mu(x)$$

is Lipschitz continuous with its Lipschitz constant

$$L_\mu = \frac{1}{\mu \sigma_2} \|\mathcal{A}\|_{1,2}^2.$$

As to the smoothing function $f_\kappa^\mu(\cdot)$, $\sigma_2 = 4$ and $\mathcal{A}$ is an identity operator. So Theorem 2.4 may also be derived from [9, Theorem 1]. Here we provide a direct proof.

2.2 Spectral Functions

A function F on the space of m-by-m real symmetric matrices is called spectral if it depends only on the eigenvalues of its argument. Spectral functions are just symmetric functions of the eigenvalues. A symmetric function is a function that is unchanged by any permutation of its variables. In this thesis, we are interested in functions F of a symmetric matrix argument that are invariant under orthogonal similarity transformations [6]:

$$F(U^T AU) = F(A), \forall U \in \mathcal{O}, A \in \mathcal{S}_m, \tag{2.28}$$

where $\mathcal{O}$ is the set of orthogonal matrices. Every such function can be decomposed as $F(A) = (f \circ \lambda)(A)$, where λ is the map that gives the eigenvalues of the matrix A and f is a symmetric function. We call such functions F spectral functions because they depend only on the spectrum of the operator A. Therefore, we can regard a spectral function as a composition of a symmetric function and the eigenvalue function.

In order to show some preliminary results, we give the following definition. For each $X \in \mathcal{S}_m$, define the set of orthonormal eigenvectors of X by

$$\mathcal{O}_X := \{P \in \mathcal{O} : P^T XP = \mathrm{Diag}[\lambda(X)]\}.$$

Now we refer to the formula for the gradient of a differential spectral function [6].

Proposition 2.5. *Let f be a symmetric function from $\mathbb{R}^n$ to $\mathbb{R}$ and $X \in \mathcal{S}_n$. Then the following holds:*

(a) *$(f \circ \lambda)$ is differentiable at point X if and only if f is differentiable at point $\lambda(X)$. In the case the gradient of $(f \circ \lambda)$ at X is given by*

$$\nabla(f \circ \lambda)(X) = U\mathrm{Diag}[\nabla f(\lambda(X))]U^T, \ \forall U \in \mathcal{O}_X. \tag{2.29}$$

(b) *$(f \circ \lambda)$ is continuously differentiable at point X if and only if f is continuously differentiable at point $\lambda(X)$.*

Lewis and Sendov [7, Theorems 3.3 and 4.2] proved the following proposition, which gives the formula for calculating the Hessian of the spectral function.

Proposition 2.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be symmetric. Then for any $X \in \mathcal{S}_n$, it holds that $(f \circ \lambda)$ is twice (continuously) differentiable at X if and only if f is twice (continuously) differentiable at $\lambda(X)$. Moreover, in this case the Hessian of the spectral function at X is*

$$\nabla^2(f \circ \lambda)(X)[H] = U(\mathrm{Diag}[\nabla^2 f(\lambda(X))\mathrm{diag}[\widetilde{H}]] + \mathcal{C}(\lambda(X)) \circ \widetilde{H})U^T, \quad \forall H \in \mathcal{S}_n, \tag{2.30}$$

where U is any orthogonal matrix in $\mathcal{O}_X$ and $\widetilde{H} = U^T H U$.

The matrix $\mathcal{C}$ in Proposition 2.6 is defined as follows: $\mathcal{C}(\omega) \in \mathbb{R}^{n\times n}$:

$$(\mathcal{C}(\omega))_{ij} := \begin{cases} 0, & \text{if } i = j \\ (\nabla^2 f(\omega))_{ii} - (\nabla^2 f(\omega))_{ij}\,, & \text{if } i \neq j \text{ and } \omega_i = \omega_j \\ \dfrac{(\nabla f(\omega))_i - (\nabla f(\omega))_j}{\omega_i - \omega_j}\,, & \text{else.} \end{cases} \tag{2.31}$$

2.3 Smoothing Functions for $\phi(x)$

Consider the $\phi(x)$ in (1.10), clearly, it is a composition function:

$$\phi(x) = (f_\kappa \circ \lambda)(C + \mathcal{A}(x)), \tag{2.32}$$

where $f_\kappa(x)$ is given in (2.1). In [12], Shi investigated the smoothing function $f_\kappa^\mu(x)$ as a approximation for $f_\kappa(x)$. So it is natural for us to think of the below composition function:

$$\phi^\mu(x) = (f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x)), \tag{2.33}$$

According to the properties of $f_\kappa^\mu(\cdot)$, for $\mu > 0$,

$$\phi^\mu(x) \le \phi(x) \le \phi^\mu(x) + \mu R, \quad \forall x \in \mathcal{Q}. \tag{2.34}$$

Since the function $f_\kappa^\mu(\cdot)$ is a symmetric function, $(f_\kappa^\mu \circ \lambda)$ is a composition function of a symmetric function $f_\kappa^\mu(\cdot) : \mathbb{R}_m \to \mathbb{R}$ and the eigenvalue function $\lambda(\cdot) : \mathcal{S}_m \to \mathbb{R}_m$. Hence the function $(f_\kappa^\mu \circ \lambda)$ is a spectral function. Consequently, $(f_\kappa^\mu \circ \lambda)$ is twice continuously differentiable. We will prove that the composition $\phi^\mu(x)$ is twice continuously differentiable and provide an estimation of the Lipschitz constant of the gradient of $\phi^\mu(x)$ as follows.

First, we will derive the gradient of $\phi^\mu(x)$. For $\mu > 0$, for any $h \in \mathbb{R}^n$ and $h \to 0$,

$$\begin{aligned} \phi^\mu(x+h) - \phi^\mu(x) &= f_\kappa^\mu(\lambda(C + \mathcal{A}(x+h))) - f_\kappa^\mu(\lambda(C + \mathcal{A}(x))) \\ &= f_\kappa^\mu(\lambda(C + \mathcal{A}(x) + \mathcal{A}(h))) - f_\kappa^\mu(\lambda(C + \mathcal{A}(x))) \\ &= \langle \nabla(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x)), \mathcal{A}(h)\rangle + O(\|h\|^2) \\ &= \langle \mathcal{A}^*(\nabla(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x))), h\rangle + O(\|h\|^2), \end{aligned} \tag{2.35}$$

where $\mathcal{A}^* D = (\langle A_1, D\rangle, \ldots, \langle A_m, D\rangle)^T$.

Thus we have the following proposition.

Proposition 2.7. *For $\mu > 0$, $\phi^\mu(\cdot)$ is continuously differentiable with its gradient given by:*

$$\nabla\phi^\mu(x) = \mathcal{A}^*(\nabla(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x))). \tag{2.36}$$

Next, we shall consider the Hessian of $\phi^\mu(x)$. First we define $\mathcal{C}^\mu(\omega) \in \mathbb{R}^{n\times n}$ as

$$(\mathcal{C}^\mu(\omega))_{ij} := \begin{cases} 0, & \text{if } i = j \\ (\nabla^2 f_\kappa^\mu(\omega))_{ii} - (\nabla^2 f_\kappa^\mu(\omega))_{ij}\,, & \text{if } i \neq j \text{ and } \omega_i = \omega_j \\ \dfrac{(\nabla f_\kappa^\mu(\omega))_i - (\nabla f_\kappa^\mu(\omega))_j}{\omega_i - \omega_j}, & \text{else.} \end{cases} \tag{2.37}$$

Proposition 2.8. *For $\mu > 0$, for any $h \in \mathbb{R}^n$ and $h \to 0$, $\phi(x)$ is twice continuously differentiable with its Hessian given by:*

$$\nabla^2\phi^\mu(x)[h] = \mathcal{A}^*(\nabla^2(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x))[H]), \tag{2.38}$$

where $H := \mathcal{A}(h)$.

Proof. For $\mu > 0$, for any $h \in \mathbb{R}^n$ and $h \to 0$,,

$$\begin{aligned}
&\langle \nabla\phi^\mu(x+h) - \nabla\phi^\mu(x), h\rangle \\
&= \langle \mathcal{A}^*(\nabla(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x+h))) - \mathcal{A}^*(\nabla(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x))), h\rangle \\
&= \langle \mathcal{A}^*(\nabla(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x+h)) - \nabla(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x))), h\rangle \\
&= \langle \nabla(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x+h)) - \nabla(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x)), \mathcal{A}(h)\rangle \\
&= \langle \nabla(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x) + \mathcal{A}(h)) - \nabla(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x)), \mathcal{A}(h)\rangle \\
&= \langle \nabla^2(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x))\mathcal{A}(h), \mathcal{A}(h)\rangle + O(\|h\|^2) \\
&= \langle \mathcal{A}^*(\nabla^2(f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x)))H, h\rangle + O(\|h\|^2),
\end{aligned} \tag{2.39}$$

which shows that (2.50) holds. □

In order to estimate $\nabla^2\phi^\mu(x)$, we need the estimation of $\mathcal{C}^\mu(\omega)$ given in (2.31). The following lemma is motivated by Lewis and Sendov [7].

Lemma 2.9. *For $\omega \in \mathbb{R}^n$ with $\omega_i \neq \omega_j$, $i \neq j$, $i, j = 1, \cdots, n$, there exists $\xi, \eta \in \mathbb{R}^n$, such that*

$$\frac{(\nabla f_\kappa^\mu(\omega))_i - (\nabla f_\kappa^\mu(\omega))_j}{\omega_i - \omega_j} = (\nabla^2 f_\kappa^\mu(\xi))_{ii} - (\nabla^2 f_\kappa^\mu(\eta))_{ij}. \tag{2.40}$$

Proof. For each $\omega \in \mathbb{R}^n$, we define the two vectors $\dot{\omega}$ and $\ddot{\omega} \in \mathbb{R}^n$ coordinatewise as follows:

$$\dot{\omega} = \begin{cases} \omega_p, & p \neq i, \\ \omega_j, & p = i. \end{cases} \qquad \ddot{\omega} = \begin{cases} \omega_p, & p \neq i, j, \\ \omega_j, & p = i. \\ \omega_i, & p = j. \end{cases} \tag{2.41}$$

By the mean value theorem,

$$\begin{aligned} & \frac{(\nabla f_\kappa^\mu(\omega))_i - (\nabla f_\kappa^\mu(\omega))_j}{\omega_i - \omega_j} \\ =& \frac{(\nabla f_\kappa^\mu(\omega))_i - (\nabla f_\kappa^\mu(\dot{\omega}))_i + (\nabla f_\kappa^\mu(\dot{\omega}))_i - (\nabla f_\kappa^\mu(\omega))_j}{\omega_i - \omega_j} \\ =& \frac{(\omega_i - \omega_j)(\nabla^2 f_\kappa^\mu(\xi))_{ii} + (\nabla f_\kappa^\mu(\dot{\omega}))_i - (\nabla f_\kappa^\mu(\omega))_j}{\omega_i - \omega_j} \\ =& (\nabla^2 f_\kappa^\mu(\xi))_{ii} + \frac{(\nabla f_\kappa^\mu(\dot{\omega}))_i - (\nabla f_\kappa^\mu(\ddot{\omega}))_i + (\nabla f_\kappa^\mu(\ddot{\omega}))_i - (\nabla f_\kappa^\mu(\omega))_j}{\omega_i - \omega_j} \\ =& (\nabla^2 f_\kappa^\mu(\xi))_{ii} + \frac{(\omega_j - \omega_i)(\nabla^2 f_\kappa^\mu(\eta))_{ij} + (\nabla f_\kappa^\mu(\ddot{\omega}))_i - (\nabla f_\kappa^\mu(\omega))_j}{\omega_i - \omega_j} \\ =& (\nabla^2 f_\kappa^\mu(\xi))_{ii} - (\nabla^2 f_\kappa^\mu(\eta))_{ij} + \frac{(\nabla f_\kappa^\mu(\ddot{\omega}))_i - (\nabla f_\kappa^\mu(\omega))_j}{\omega_i - \omega_j}, \end{aligned} \tag{2.42}$$

where ξ is a vector between ω and $\dot{\omega}$, and η is a vector between $\dot{\omega}$ and $\ddot{\omega}$. We next consider the term $\frac{(\nabla f_\kappa^\mu(\ddot{\omega}))_i - (\nabla f_\kappa^\mu(\omega))_j}{\omega_i - \omega_j}$. By the definitions, we know

$$(\nabla f_\kappa^\mu(\omega))_j = \frac{\partial}{\partial \omega_j} f_\kappa^\mu(\omega) \tag{2.43}$$

and

$$(\nabla f_\kappa^\mu(\ddot{\omega}))_i = \frac{\partial}{\partial \ddot{\omega}_i} f_\kappa^\mu(\ddot{\omega}) = \frac{\partial}{\partial \omega_j} f_\kappa^\mu(\ddot{\omega}). \tag{2.44}$$

Since $f_\kappa^\mu(\cdot)$ is a symmetric function,

$$(\nabla f_\kappa^\mu(\ddot{\omega}))_i = \frac{\partial}{\partial \omega_j} f_\kappa^\mu(\omega) = (\nabla f_\kappa^\mu(\omega))_j.$$

Consequently,

$$\frac{(\nabla f_\kappa^\mu(\ddot{\omega}))_i - (\nabla f_\kappa^\mu(\omega))_j}{\omega_i - \omega_j} = 0.$$

Therefore (2.40) holds. $\square$

Lemma 2.10. *For any* $1 \le i,j \le m$, *each entry of* $(\mathcal{C}^\mu(\omega))_{ij}$ *has the following bound:*

$$0 \le (\mathcal{C}^\mu(\omega))_{ij} \le \frac{1}{2\mu}. \tag{2.45}$$

Proof.

$$(\mathcal{C}^\mu(\omega))_{ij} = \begin{cases} (\nabla^2 f_\kappa^\mu(\omega))_{ii} - (\nabla^2 f_\kappa^\mu(\omega))_{ij}\,, & i = j \\ (\nabla^2 f_\kappa^\mu(\xi))_{ii} - (\nabla^2 f_\kappa^\mu(\eta))_{ij}\,, & i \neq j. \end{cases} \tag{2.46}$$

For any $x, y \in \mathbb{R}^n$, by Theorem 2.4,

$$(\nabla^2 f_\kappa^\mu(x))_{ii} - (\nabla^2 f_\kappa^\mu(y))_{ij} \le \max(\nabla^2 f_\kappa^\mu(x))_{ii} - \min(\nabla^2 f_\kappa^\mu(y))_{ij} = \frac{1}{2\mu}; \tag{2.47}$$

$$(\nabla^2 f_\kappa^\mu(x))_{ii} - (\nabla^2 f_\kappa^\mu(y))_{ij} \ge \min(\nabla^2 f_\kappa^\mu(x))_{ii} - \max(\nabla^2 f_\kappa^\mu(y))_{ij} = 0. \tag{2.48}$$

Let $i = j, \quad x = \omega, \quad y = \omega$, then

$$0 \le (\mathcal{C}^\mu(\omega))_{ij} \le \frac{1}{2\mu};$$

Let $i \neq j, \quad x = \xi, \quad y = \eta$, then

$$0 \le (\mathcal{C}^\mu(\omega))_{ij} \le \frac{1}{2\mu}.$$

□

Next, we estimate the Lipschitz constant of $\nabla \phi^\mu(x)$. For any $h \in \mathbb{R}^n$, we have

$$\begin{aligned} &\langle h, \nabla^2 \phi^\mu(x)[h]\rangle \\ =&\langle h, \mathcal{A}^*(\nabla^2 (f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x))[H])\rangle \\ =&\langle H, \nabla^2 (f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x))[H]\rangle, \end{aligned} \tag{2.49}$$

with $H = \mathcal{A}(h)$, and

$$\begin{aligned} &\nabla^2 (f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x))[H] \\ =&U(\mathrm{Diag}[\nabla^2 f_\kappa^\mu(\lambda(C + \mathcal{A}(x)))\mathrm{diag}[\widetilde{H}]] + \mathcal{C}^\mu(\lambda(C + \mathcal{A}(x))) \circ \widetilde{H})U^T, \end{aligned} \tag{2.50}$$

where $U \in \mathcal{O}_{C+\mathcal{A}(x)}$, and $\widetilde{H} = U^T H U$. Therefore

$$\begin{aligned}
\langle h, \nabla^2 \phi^\mu(x)[h] \rangle &= \sum_{i=1}^{n} (\nabla^2 (f_\kappa^\mu \circ \lambda)(C + \mathcal{A}(x)))_{ii} \widetilde{H}_{ii} + \sum_{i,j=1 \quad i \neq j}^{n} \mathcal{C}_{ij}^\mu (\lambda(C + \mathcal{A}(x))) \widetilde{H}_{ij} \\
&\leq \frac{1}{4\mu} \sum_{i=1}^{n} \widetilde{H}_{ii} + \frac{1}{2\mu} \sum_{i,j=1 \quad i \neq j}^{n} \widetilde{H}_{ij} \\
&\leq \frac{1}{2\mu} \langle \widetilde{H}, \widetilde{H} \rangle \\
&= \frac{1}{2\mu} \langle \mathcal{A}(h), \mathcal{A}(h) \rangle \\
&= \frac{1}{2\mu} \sum_{i,j=1}^{m} h_i h_j \langle A_i, A_j \rangle \\
&= \frac{1}{2\mu} h^T G h \\
&\leq \frac{1}{2\mu} \|G\| \|h\|^2,
\end{aligned} \tag{2.51}$$

where $G \in S_m$, $(G)_{ij} = \langle A_i, A_j \rangle$, and $\|G\| := \max_{1 \leq i \leq m} |\lambda_i(G)| = \max\{\lambda_1(G), -\lambda_m(G)\}$. Thus the Lipschitz constant for the gradient of the smoothing function $\phi^\mu(x)$ is:

$$L = \frac{1}{2\mu} \|G\|. \tag{2.52}$$

In particular, if we take $\mu := \mu(\varepsilon) = \frac{\epsilon}{2R}$, where $R = 2m \ln(2m) - \kappa \ln \kappa - (2m - \kappa) \ln(2m - \kappa)$, then we have

$$L = \frac{R}{\epsilon} \|G\|. \tag{2.53}$$

Chapter 3

Smoothing Functions for the Sum of the κ Largest Absolute Values of Eigenvalues

In order to solve problem (1.13), we need the concept of singular values. According to some properties of singular values, we can obtain a computable smoothing function for problem $\psi(x)$.

3.1 Preliminaries

Similar to the eigenvalue decomposition of a symmetric matrix, a non-symmetric matrix has singular value decomposition. Let $A \in \mathcal{M}_{n,m}$, and without generality we assume $n \leq m$. Then there exist orthogonal matrices $U \in \mathcal{M}_{n,n}$ and $V \in \mathcal{M}_{m,m}$ such that A has the following *singular value decomposition (SVD)*:

$$U^T AV = [\Sigma(A) \quad 0], \tag{3.1}$$

where $\Sigma(A) = \text{diag}(\sigma_1(A), \ldots, \sigma_n(A))$ and $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_n(A) \geq 0$ are the singular values of A [5].

The singular value has the following property:

$$\sqrt{AA^T} = \sqrt{U\Sigma^2(A)U^T} = U\text{diag}[\sigma_1(A), \ldots, \sigma_n(A)]U^T. \tag{3.2}$$

In particular, if A is symmetric, we have

$$AA^T = A^2 = P\text{diag}[\lambda_1(A)^2, \ldots, \lambda_n(A)^2]P^T, \tag{3.3}$$

where $\lambda_1(A), \ldots, \lambda_n(A)$ are the eigenvalues of A. Comparing (3.2) and (3.3), the singular values are the square root of respective eigenvalues of AA^T, which means $\sigma_i(A) = |\lambda_i(A)|$, $i = 1, \ldots, n$ for all symmetric matrix A.

For any $W \in \mathcal{S}_{n+m}$, we define:

$$\Lambda(W) = \text{diag}(\lambda^{(1)}(W), \ldots, \lambda^{(n)}(W), \lambda^{(n+m)}(W), \ldots, \lambda^{(n+1)}(W)), \tag{3.4}$$

where $\{\lambda_i(W) : i = 1, \cdots, n+m\}$ are the eigenvalues of W arrange in decreasing order. Noted that the first n diagonal entries of $\Lambda(W)$ are just the n largest eigenvalues of W, arranged in deceasing order, while the last m diagonal entries of $\Lambda(W)$ are the m smallest eigenvalues of W, arranged in increasing order. This arrangement will show convenience shortly afterwards.

Define the linear operator: $\Xi : \mathcal{M}_{n,m} \to \mathcal{S}_{m+n}$ by

$$\Xi(B) = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}, \quad B \in \mathcal{M}_{n,m}. \tag{3.5}$$

For $A \in \mathcal{S}_m$, let it have the following eigenvalue decomposition:

$$V^T AV = \Sigma(A), \quad V \in \mathcal{O}_A. \tag{3.6}$$

The following result is derived from Golub and Van Loan [5, Section 8.6].

Proposition 3.1. *Suppose that $A \in \mathcal{S}_m$ has the eigenvalue decomposition (3.6). Then the matrix $\Xi(A)$ has the following spectral decomposition:*

$$\Xi(A) = Q(\Lambda(\Xi(A)))Q^T = Q \begin{bmatrix} \Sigma(A) & 0 \\ 0 & -\Sigma(A) \end{bmatrix} Q^T, \tag{3.7}$$

where $Q \in \mathcal{O}_{2m,2m}$ is defined by:

$$Q = \frac{1}{\sqrt{2}} \begin{bmatrix} V & V \\ V & -V \end{bmatrix}, \tag{3.8}$$

i.e. the eigenvalues of $\Xi(A)$ are $\pm\sigma_i(A)$.

3.2 Smoothing Functions for $\psi(x)$

From Proposition 3.1, we know that

$$\psi(x) = \sum_{i=1}^{\kappa} \lambda_i(\Xi(C + \mathcal{A}(x))). \tag{3.9}$$

Similar to Chapter 2, we define the smoothing function for $\psi(x)$ by

$$\psi^{\mu}(x) = (f_{\kappa}^{\mu} \circ \lambda)(\Xi(C + \mathcal{A}(x))), \tag{3.10}$$

where the $f_{\kappa}^{\mu}(\cdot)$ is the smoothing function defined in Chapter 2 and $\kappa \leq m$.

For $x \in \mathbb{R}_n$, we define a linear operator $\Gamma : \mathbb{R}_n \to \mathcal{S}_{2m}$ as follows:

$$\Gamma(x) = \Xi(\mathcal{A}(x)). \tag{3.11}$$

Let us consider the adjoint of $\Gamma(x)$. For $Y \in \mathcal{S}_{2m}$,

$$\begin{aligned} \langle \Gamma(x), Y \rangle &= \langle \Xi(\mathcal{A}(x)), Y \rangle \\ &= \left\langle \begin{pmatrix} 0 & \mathcal{A}(x) \\ (\mathcal{A}(x))^T & 0 \end{pmatrix}, \begin{pmatrix} Y_1 & Y_2 \\ Y_2^T & Y_3 \end{pmatrix} \right\rangle \\ &= \langle \mathcal{A}(x), Y_2 \rangle + \langle (\mathcal{A}(x))^T, Y_2^T \rangle \\ &= 2\langle \mathcal{A}(x), Y_2 \rangle \\ &= \sum_{i=1}^{n} 2x_i \langle A_i, Y_2 \rangle. \end{aligned} \tag{3.12}$$

Thus, $\Gamma^*(Y) = (2\langle A_1, Y_2 \rangle, \cdots, 2\langle A_n, Y_2 \rangle)^T$.

The smoothing function $\psi^\mu(x)$ takes the following form:

$$\psi^\mu(x) = (f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x))). \tag{3.13}$$

Since we have already known that, for any $\mu > 0$, $(f_\kappa^\mu \circ \lambda)$ is twice continuously differentiable, $\psi^\mu(\cdot)$ is also twice continuously differentiable. Now we discuss its gradient and Hessian.

Proposition 3.2. *For $\mu > 0$, $\psi^\mu(\cdot)$ is continuously differentiable with its gradient given by*

$$\nabla\psi^\mu(x) = \Gamma^*(\nabla(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x))). \tag{3.14}$$

Proof. For $\mu > 0$, for any $h \in \mathbb{R}^n$ and $h \to 0$,

$$\begin{aligned} \psi^\mu(x+h) - \psi^\mu(x) &= f_\kappa^\mu(\lambda(\Xi(C) + \Gamma(x+h))) - f_\kappa^\mu(\lambda(\Xi(C) + \Gamma(x))) \\ &= f_\kappa^\mu(\lambda(\Xi(C) + \Gamma(x) + \Gamma(h))) - f_\kappa^\mu(\lambda(\Xi(C) + \Gamma(x))) \\ &= \langle \nabla(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x)), \Gamma(h)\rangle + O(\|h\|^2) \\ &= \langle \Gamma^*(\nabla(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x))), h\rangle + O(\|h\|^2), \end{aligned} \tag{3.15}$$

which agrees with (3.14). □

Proposition 3.3. *For $\mu > 0$, $\psi^\mu(\cdot)$ is twice continuously differentiable with its Hessian given by*

$$\nabla^2\psi^\mu(x)[h] = \Gamma^*(\nabla^2(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x))[H]), \quad h \in \mathbb{R}^n, \tag{3.16}$$

where $H := \Gamma(h)$.

Proof. For $\mu > 0$, for any $h \in \mathbb{R}^n$ and $h \to 0$,

$$\begin{aligned}
&\langle \nabla\psi^\mu(x+h) - \nabla\psi^\mu(x), h\rangle \\
&= \langle \Gamma^*(\nabla(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x+h))) - \Gamma^*(\nabla(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x))), h\rangle \\
&= \langle \Gamma^*(\nabla(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x+h)) - \nabla(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x))), h\rangle \\
&= \langle \nabla(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x+h)) - \nabla(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x)), \Gamma(h)\rangle \qquad (3.17) \\
&= \langle \nabla(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x) + \Gamma(h)) - \nabla(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x)), \Gamma(h)\rangle \\
&= \langle \nabla^2(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x))\Gamma(h), \Gamma(h)\rangle + O(\|h\|^2) \\
&= \langle \Gamma^*(\nabla^2(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x)))H, h\rangle + O(\|h\|^2),
\end{aligned}$$

which shows that (3.16) holds. □

Next, we estimate the Lipschitz constant of $\nabla\psi^\mu(x)$. For any $h \in \mathbb{R}^n$, we have

$$\begin{aligned}
\langle \nabla h, \nabla^2\psi^\mu(x)[h]\rangle =& \langle h, \Gamma^*(\nabla^2(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x))[H])\rangle \\
=& \langle H, (\nabla^2 f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x))H\rangle,
\end{aligned} \qquad (3.18)$$

with

$$\begin{aligned}
&\nabla^2(f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x))[H] \\
=& U(\mathrm{Diag}[\nabla^2 f_\kappa^\mu(\lambda(\Xi(C) + \Gamma(x)))\mathrm{diag}[\widetilde{H}]] + \mathcal{C}^\mu(\lambda(\Xi(C) + \Gamma(x))) \circ \widetilde{H})U^T,
\end{aligned} \qquad (3.19)$$

where $U \in \mathcal{O}_{\Xi(C)+\Gamma(x)}$ and $\widetilde{H} = U^T H U$. Therefore

$$\begin{aligned}
&\langle h, \nabla^2 \psi^\mu(x)[h] \rangle \\
=&\sum_{i=1}^{n} (\nabla^2 (f_\kappa^\mu \circ \lambda)(\Xi(C) + \Gamma(x)))_{ii} \widetilde{H}_{ii}^2 + \sum_{i,j=1 \quad i \neq j}^{n} \mathcal{C}_{ij}^\mu (\lambda(\Xi(C) + \Gamma(x))) \widetilde{H}_{ij}^2 \\
\leq&\frac{1}{4\mu} \sum_{i=1}^{n} \widetilde{H}_{ii}^2 + \frac{1}{2\mu} \sum_{i,j=1 \quad i \neq j}^{n} \widetilde{H}_{ij}^2 \\
\leq&\frac{1}{2\mu} \langle \widetilde{H}, \widetilde{H} \rangle \\
\leq&\frac{1}{2\mu} \langle \Gamma(h), \Gamma(h) \rangle \\
\leq&\frac{1}{\mu} \langle \mathcal{A}(h), \mathcal{A}(h) \rangle \\
\leq&\frac{1}{\mu} \|G\| \|h\|^2.
\end{aligned} \tag{3.20}$$

Thus, the Lipschitz constant for the gradient of the smoothing function $\psi^\mu(\text{cot})$ is

$$L = \frac{1}{\mu} \|G\|. \tag{3.21}$$

In particular, if we take $\mu := \mu(\varepsilon) = \frac{\epsilon}{2R}$, where$R := m \ln m - \kappa \ln \kappa - (m-\kappa) \ln(m-\kappa)$ then we have

$$L = \frac{2R}{\epsilon} \|G\|. \tag{3.22}$$

Chapter 4

Numerical Experiments

Nesterov's method is to solve the following optimization problem:

$$\min_{x \in \mathcal{Q}} \{f(x)\}, \tag{4.1}$$

where f is a convex function with its gradient of $f(x)$ satisfied the Lipschitz condition:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \, \forall x, y \in \mathbb{R}^n. \tag{4.2}$$

where $L > 0$ is a constant.

Consider a prox-function $d(x)$ in $\mathcal{Q}$. We assume that $d(x)$ is continuous and strongly convex on $\mathcal{Q}$ with convexity parameter $\sigma > 0$. Denote x_0 by

$$x_0 = \min_{x \in \mathcal{Q}} d(x). \tag{4.3}$$

Without loss of generality we assume $d(x_0) = 0$. Thus, for any $x \in \mathcal{Q}$, we have

$$d(x) \geq \frac{1}{2}\sigma\|x - x_0\|^2. \tag{4.4}$$

Define

$$T_{\mathcal{Q}}(x) = \min_y \{\langle \nabla f(x), y - x\rangle + \frac{1}{2}L\|y - x\|^2 : y \in \mathcal{Q}\}. \tag{4.5}$$

Now we are ready to give Nesterov's smoothing algorithm [9]:

For $k \geq 0$ do

1. Compute $f(x_k)$ and $\nabla f(x_k)$.

2. Find $y_k = T_{\mathcal{Q}}(x_k)$.

3. Find $z_k = \arg\min_x\{\frac{L}{\sigma}d(x) + \sum_{i=0}^{k}\frac{i+1}{2}[f(x_i) + \langle\nabla f(x_i), x - x_i\rangle] : x \in \mathcal{Q}\}$.

4. Set $x_{k+1} = \frac{2}{k+3}z_k + \frac{k+1}{k+3}y_k$.

Nesterov prove the following Theorem [9, Theorem 3]:

Theorem 4.1. *Let the sequences $\{x_k\}_{k=0}^{\infty}$ and $\{y_k\}_{k=0}^{\infty}$ be generated by the above algorithm. Then for any $k \geq 0$, we have*

$$\frac{(k+1)(k+2)}{4}f(y_k) \leq \min_x\left\{\frac{L}{\sigma}d(x) + \sum_{i=0}^{k}\frac{i+1}{2}[f(x_i) + \langle\nabla f(x_i), x - x_i\rangle] : x \in \mathcal{Q}\right\}. \tag{4.6}$$

Therefore,

$$f(y_k) - f(x^*) \leq \frac{4Ld(x^*)}{\sigma(k+1)(k+2)}. \tag{4.7}$$

where x^* is an optimal solution to the problem. (4.1).

By applying Bregman's distance, Nesterov provided a modified algorithm, which gives a way to compute $T_{\mathcal{Q}}(x_k)$. In the new algorithm, we compute the $V_{\mathcal{Q}}$ instead of $T_{\mathcal{Q}}$. Bregman's distance was introduced in [3], as an extension to the usual metric discrepancy measure $(x, y) \rightarrow \|x - y\|^2$. If $f(\cdot)$ is a real convex function, then the Bregman distance between two parameters z and x is defined as

$$\xi(z, x) = f(x) - f(z) - \langle\nabla f(z), x - z\rangle, \quad x, z \in \mathcal{Q}. \tag{4.8}$$

The Bregman distance satisfies

$$\xi(z, x) \geq \frac{1}{2}\sigma\|x - z\|^2. \tag{4.9}$$

Define the Bregman projection of h as follows:

$$V_{\mathcal{Q}}(z, h) = \text{argmin}\{h^T(x - z) + \xi(z, x) : x \in \mathcal{Q}\}. \tag{4.10}$$

The following algorithm is Nesterov's algorithm via the Bregman distance [9]:

1. Choose $y_0 = z_0 = \arg\min_x\{\frac{L}{\sigma}d(x) + \frac{1}{2}[f(x_0) + \langle\nabla f(x_0), x - x_0\rangle] : x \in \mathcal{Q}\}$.

2. For $k \geq 0$ iterate:

a. Find $z_k = \arg\min_x\{\frac{L}{\sigma}d(x) + \sum_{i=0}^{k}\frac{i+1}{2}[f(x_i) + \langle\nabla f(x_i), x - x_i\rangle] : x \in \mathcal{Q}\}$.

b. Set $\tau_k = \frac{2}{k+3}$ and $x_{k+1} = \tau_k z_k + (1-\tau_k)y_k$.

c. Find $\hat{x}_{k+1} = V_{\mathcal{Q}}(z_k, \frac{\sigma}{L}\tau_k\nabla f(x_{k+1}))$.

d. Set $y_{k+1} = \tau_k\hat{x}_{k+1} + (1-\tau_k)y_k$.

Nesterov pointed out for the above method, Theorem 4.1 holds. The computational results shown in the next section are achieved by applying the above algorithm.

4.1 Computational Results

First, we solve the smoothing problem:

$$\min_{x\in\mathcal{Q}}\phi^{\mu}(x),$$

where the closed convex set $\mathcal{Q}$ is given by

$$\mathcal{Q} = \{x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1,\ 0 \leq x_i \leq 1,\ \ i = 1, 2, \ldots, n\}.$$

Let P_i be a m-by-m random matrix with its entries in [-1,1], define $C,\ A_1, A_2, \cdots, A_n$ as $C = \frac{1}{2}(P_0 + P_0^T)$ and $A_i = i \times \frac{1}{2}(P_i + P_i^T)$, $i = 1, \cdots, n$.

Let ϵ be the desired accuracy, i.e., $\phi(\bar{x}) - \phi^* \leq \epsilon$. R is defined by 2.5 and $d(x) = lnn + \sum_{i=1}^{n} x_i$. According to Theorem 4.1, we have the iteration bound $N := [\frac{2\sqrt{2R\ln n}}{\epsilon}\sqrt{\|G\|}]$. We composed the matlab codes for this problem. Tables 4.1, 4.2 and 4.3 are the numerical results for this problem.

Table 4.1: for $m = 10$, $n = 4$ and different κ

κ	ϵ	N	time	starting value	optimal value
3	0.01	12186	21”	18.26	12.74
3	0.001	123668	197”	21.328	16.184
4	0.01	17835	34”	23.14	14.23
4	0.001	175299	376”	21.965	12.172
5	0.01	18830	39”	19.83	12.71
5	0.001	188304	377”	21.159	11.974

Table 4.2: for $m = 30$, $n = 4$ and different κ

κ	ϵ	N	time	starting value	optimal value
3	0.01	62906	374”	59.08	35.43
3	0.001	653111	4230”	60.952	36.184
4	0.01	71845	393”	62.51	37.57
4	0.001	742312	4505”	61.253	35.942
5	0.01	73390	408”	84.81	39.52
5	0.001	753401	4669”	86.695	40.994

Table 4.3: for $n = 4$, $\kappa = 4, \epsilon = 0.01$ and different m

m	N	time	starting value	optimal value
10	17835	34”	23.13	14.23
30	71845	393”	62.51	37.57
50	136471	1568”	60.952	36.184

Next, we solve the smoothing problem :

$$\min_{x \in \mathcal{Q}} \psi^{\mu}(x).$$

The parameters $\mathcal{Q}$, C, $A_i, i = 1, \cdots, n$ are defined as for $\phi^{\mu}(x)$. The maximal number of iterations is $N = [\frac{4\sqrt{R \ln n}}{\epsilon}\sqrt{\|G\|}]$. We composed the matlab codes for this problem. Tables 4.4 contains the computational results.

Table 4.4: for $n = 4$, $\kappa = 3$ and different m

m	ϵ	N	time	starting value	optimal value
10	0.01	29529	162"	21.45	13.37
10	0.001	268361	15562"	21.623	13.405
20	0.01	64050	306"	39.71	25.80
20	0.001	684150	2958"	41.125	25.867
30	0.01	107190	744"	59.73	36.67
30	0.001	1050400	7230"	62.452	37.336

4.2 Conclusions

Note that the two problems we have discussed can both be converted into semidefinite programming problems. One may then consider second order approaches like Newton's method to solve these problems. However, for high dimensional problems, the efficiency of such approaches are not satisfactory.

In this chapter, we have done some experiments, but our final goal is to solve high dimension problems. In our experiments, the number of iterations is very high. In order to achieve the required accuracy with efficiency, we make the following observation on the improvement of the smoothing algorithm.

Firstly, we can reduce the time of eigenvalue decomposition in each iteration. In our algorithm, we only need the first κ eigenvalues. In many cases, $\kappa \ll m$, we can try to do the partial eigenvalue decomposition. In each decomposition steps, we only decompose the first l largest eigenvalues, where l is a heuristic parameter, and $\kappa < l \leq m$. For instance, let l be $\kappa + 3$, $\kappa + 5$, or 2κ, 3κ. Then we compare the lth largest eigenvalue and the κth eigenvalue. If the lth eigenvalue is far less than the κth eigenvalue, the eigenvalues less than the lth have little contribution to the optimal value $v^T\lambda$ in each iteration, which means that the correspondence v_is are very small for $i > l$. When we apply (2.29) to compute the gradient of $\phi^\mu(x)$ or $\psi^\mu(x)$, v_is also have little contribution to the gradient matrix for $i > l$. From the above discussion, the partial eigenvalues decomposition will not cause great loss in the process. How to find a proper l in each decomposition step, and how to estimate the loss need to be taken into consideration in further research.

Secondly, Nesterov has provided an excessive gap algorithm [10], which is based on the upper bound and lower bound of the optimal value. In further research, we may try to apply this algorithm to our problems and reduce the number of total iterations.

Finally, in our analysis, we prove that the gradients of the two classes of problems are Lipschitz continuous, and derive the Lipschitz constant. The maximal iteration number depends on the Lipschitz constant, which depends on the norm of the matrix G. If $\|G\|$ becomes larger, the Lipschitz constant becomes larger, thereby the iteration number becomes larger. In Nesterov's analysis, $\|\mathcal{A}\|$ is the infinity norm, while in our result, $\|G\|$ is the normal matrix norm, its property still needs investigation.

Bibliography

[1] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, "Bregman monotone optimization algorithms", *SIAM Journal on Control and Optimization*, vol. 42, pp. 596-636, 2003.

[2] A. Ben-Tal and M. Teboulle, "A smoothing technique for nondifferentiable optimization problems", in Optimization. Fifth French German Conference, Lecture Notes in Math. 1405, Springer-Verlag, New York, 1989, 1–11.

[3] L. M. Bregman "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming", U.S.S.R *Computational Mathematics and Mathematical Physics*, vol. 7, pp. 200-217, 1967.

[4] X. Chen, H.D. Qi, L. Qi and K. Teo, "Smooth convex approximation to the maximum eigenvalue function" to appear in *J. of Global Optimization*.

[5] G.H. Golub and C.F. Van Loan, "Matrix Computations", the Johns Hopkins University Press, Baltimore, USA, Third Edition,(1996) .

[6] A.S. Lewis, “Derivatives of spectral functions”, *Mathematics of Operations Research, 21* (1996), 576–588.

[7] A.S. Lewis and H.S. Sendov, “Twice differentiable spectral functions”, *SIAM J. Matrix Anal. Appl., 22* (2001), 368–386.

[8] Yu. Nesterov, “Unconstrained convex minimization in relative scale”, CORE DP 2003/96.

[9] Yu. Nesterov, “Smooth minimization of non-smooth functions”, CORE DP 2003/12, February 2003. Accepted by *Mathematcal Programming.*

[10] Yu. Nesterov, “Excessive gap technique in non-smooth convex minimizarion”, CORE DP 2003/35, May 2003. Accepted by *SIAM J. Optimization.*

[11] Yu. Nesterov, “Smoothing technique and its applications in semidefinite Optimizarion”, CORE DP 2004/73, October 2004. Accepted by *SIAM J. Optimization.*

[12] S. Shi. “Smooth convex approximation and its applications“, Master’s thesis, National University of Singapore, 2004.

[13] D. Sun and J. Sun, “Strong semismoothness of eigenvalues of symmetric matrices and its application to inverse eigenvalue problems”, *SIAM J. Numer. Anal., 40* (2003), 2352–2367.

Name: Yu Qi
Degree: Master of Science
Department: Mathematics
Thesis Title: Smoothing Approximations for Two Classes of Convex Eigenvalue Optimization Problems

Abstract

In this thesis, we consider two problems: minimizing the sum of the κ largest eigenvalues and the sum of the κth largest absolute values of eigenvalues of a parametric linear operator. In order to apply Nesterov's smoothing algorithm to solve these two problems, we construct two computable smoothing functions whose gradients are Lipschitz continuous. This construction is based on Shi's thesis [12] and new techniques introduced in this thesis. Numerical results on the performance of Nesterov's smooth algorithm are also reported.

Keywords:
Smoothing functions, Lipschitz constants, Smoothing algorithm, Eigenvalue problems.

SMOOTHING APPROXIMATIONS FOR TWO CLASSES OF CONVEX EIGENVALUE OPTIMIZATION PROBLEMS

YU QI

NATIONAL UNIVERSITY OF SINGAPORE

2005