Smooth Convex Approximation to the Maximum Eigenvalue Function [1]

Xin Chen [2] Houduo Qi [3], Liqun Qi [4] and Kok-lay Teo [5]

February 9, 2001

Abstract. In this paper, we consider smooth convex approximations to the maximum eigenvalue function. To make it applicable to a wide class of applications, the study is conducted on the composite function of the maximum eigenvalue function and a linear operator mapping $\mathbb{R}^m$ to $\mathcal{S}_n$, the space of n-by-n symmetric matrices. The composite function in turn is the natural objective function of minimizing the maximum eigenvalue function over an affine space in $\mathcal{S}_n$. This leads to a sequence of smooth convex minimization problems governed by a smoothing parameter. As the parameter goes to *zero*, the original problem is recovered. We then develop a computable Hessian formula of the smooth convex functions, matrix representation of the Hessian, and study the regularity conditions which guarantees the nonsingularity of the Hessian matrices. The study on the well-posedness of the smooth convex function leads to a regularization method. Finally, we illustrate the theoretical results with a combinatorial application. Our study strengthens the pioneer research of Lewis and Sendov [16] on twice differentiable spectral functions in a sense that, although most of the known spectral functions are not twice differentiable (not even differentiable), there are indeed many smooth spectral functions arising from approximating the nonsmooth spectral functions.

Key words. Symmetric function, spectral function, Tikhonov regularization, matrix representation.

AMS subject classifications. 49M45, 90C25, 90C33

---

[1]This work was supported by the Austrailan Research Council and the Research Grant Council of Hong Kong.

[2]Operations Research Center, Massachustts Institute of Technology, 77 Massachusetts Avenue, Building E40-194. Cambridge, MA 02139, U.S.A. E-mail: xchen@mit.edu.

[3]School of Mathematics, the University of New South Wales, Sydney 2052, Australia. E-mail: hdqi@maths.unsw.edu.au.

[4]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. E-mail: maqilq@polyu.edu.hk.

[5]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. E-mail: mateokl@polyu.edu.hk.

1 Introduction

Let $\mathcal{S}_n$ denote the space of n-by-n symmetric matrices endowed with the inner product $\langle X, Y\rangle := \mathrm{tr}(XY)$ for any $X, Y \in \mathcal{S}_n$. The maximum eigenvalue function is often defined as the first component of the eigenvalue function $\lambda : \mathcal{S}_n \to \mathbb{R}^n$, where for any $X \in \mathcal{S}_n$, $\lambda(X)$ is the vector of eigenvalues of X in nonincreasing order, i.e., $\lambda_1(X) \geq \lambda_2(X) \geq \ldots \geq \lambda_n(X)$. The minimization of the maximum eigenvalue function over various sets gives rise to probably the most important class of eigenvalue optimization problems [14]. In particular, the following problem which is to minimize the maximum eigenvalue function in an affine subspace of $\mathcal{S}_n$ is an equivalent reformulation of the semidefinite programming relaxation of some combinatorial problems, see Sec. 5. Let $A_0, A_1, \ldots, A_m \in \mathcal{S}_n$ be given, and define an operator $\mathcal{A} : \mathbb{R}^m \to \mathcal{S}_n$ by

$$\mathcal{A}y := \sum_{i=1}^{m} y_i A_i, \qquad \forall y \in \mathbb{R}^m.$$

Then the basic eigenvalue optimization problem we mentioned is given by

$$\inf_{y \in \mathbb{R}^m} \ \lambda_1(A(y)) \tag{1}$$

where

$$A(y) := A_0 + \mathcal{A}y \tag{2}$$

For more discussion on the problem, see [14, 11, 18, 26]. In combinatorial applications in [11], there is a linear term added to $\lambda_1(\cdot)$. We will see later on, this linear term can be viewed as a component of $\lambda_1(\cdot)$. Hence the model we considered is sufficient for covering this important class of applications.

It is well known that the eigenvalue function is usually not differentiable, which inevitably gives rise to extreme difficulties in extending classical optimization methods (often using the information of gradient and Hessian of objective functions) to eigenvalue optimization problems, see [20]. Pioneer works conducted recently by Lewis within a very general framework of spectral functions open ways in such extensions. A spectral function is usually defined as a composite function of a symmetric function $f : \mathbb{R}^n \to \mathbb{R}$ and the eigenvalue function $\lambda(\cdot)$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is symmetric if f is invariant under coordinate permutations, i.e., $f(P\mu) = f(\mu)$ for any $\mu \in \mathbb{R}^n$ and $P \in \mathcal{P}$, the set of all permutation matrices. Hence the spectral function defined by f and λ can be written as $(f \circ \lambda) : \mathcal{S}_n \to \mathbb{R}$ with $(f \circ \lambda)(X) = f(\lambda(X))$ for any $X \in \mathcal{S}_n$. It seems that the spectral function, thought as a composition of $\lambda(\cdot)$ and a symmetric function, would inherit the nonsmoothness of the eigenvalue function. However, according to the study of Lewis [13], the smoothness of a spectral function depends only on the smoothness of f. In fact, Lewis proved that $(f \circ \lambda)$ is differentiable if and only if f is differentiable (we talk about Fréchet differentiability throughout the paper.) Furthermore, Lewis and Sendov [16] proved that $(f \circ \lambda)$ is twice (continuously) differentiable if and only if f is twice (continuously) differentiable. And it is further conjectured for the general $C^{(k)}$ case. For modern nonsmooth analysis of spectral functions, we refer to a celebrated paper [15] by Lewis.

Although most of known spectral functions are not twice differentiable (not even differentiable), we will soon see that the smooth approximation denoted by f_ε ($\varepsilon > 0$ in general) to the nonsmooth function f gives rise to a smooth spectral function $(f_\varepsilon \circ \lambda)$, which is a smooth approximation to the spectral function $(f \circ \lambda)$. For example, let

$$f(x) := \max\{x_1, \ldots, x_n\} \tag{3}$$

Then

$$\lambda_1(X) = (f \circ \lambda)(X), \quad \forall\, X \in \mathcal{S}_n.$$

A well known smoothing function to the maximum function (3) is the exponential penalty function:

$$f_\varepsilon(x) := \varepsilon \ln \left(\sum_{i=1}^{n} e^{x_i/\varepsilon} \right), \qquad \varepsilon > 0. \tag{4}$$

It is a C^∞ convex function and has the following uniform approximation to f [4]:

$$0 \le f_\varepsilon(x) - f(x) \le \varepsilon \ln n.$$

The penalty function sometimes called the aggregation function is used in a number of occasions [2, 17, 30, 29, 9, 21, 24], In particular, Ben-Tal and Teboulle [2] studied this function in view of the recession function of the maximum function (3), and further studies along this line were conducted recently by Auslender [1] in a more general framework. As a matter of fact, this function is one of nonseparable penalty functions in [1].

It is easy to see that the exponential penalty function is symmetric in $\mathbb{R}^n$ and the well defined spectral function $(f_\varepsilon \circ \lambda)$ is a uniform approximation to $\lambda_1(\cdot)$, i.e.,

$$0 \le (f_\varepsilon \circ \lambda)(X) - \lambda_1(X) \le \varepsilon \ln n, \qquad \forall\, \varepsilon > 0,\ X \in \mathcal{S}_n. \tag{5}$$

Moreover, $(f_\varepsilon \circ \lambda)$ is twice continuously differentiable. It remains to consider its composition with the affine mapping $A(\cdot)$. For $\varepsilon > 0$, define $\theta_\varepsilon : \mathbb{R}^m \to \mathbb{R}$ by

$$\theta_\varepsilon(y) := (f_\varepsilon \circ \lambda \circ A)(y) = f_\varepsilon(\lambda(A(y))), \qquad \forall\, y \in \mathbb{R}^m.$$

The uniform approximation (5) implies

$$\lim_{\varepsilon \to 0} \theta_\varepsilon(y) = \lambda_1(A(y)), \qquad \forall\, y \in \mathbb{R}^m.$$

Therefore it is natural to consider the following convex minimization problem with twice continuously differentiable data:

$$\min_{y \in \mathbb{R}^m} \ \theta_\varepsilon(y). \tag{6}$$

Hence we arrive at a sequence of smooth functions which are uniform approximations to the maximum eigenvalue function with the accuracy controlled by a smoothing parameter. As the parameter goes to *zero*, the function $\lambda(\cdot)$ is recovered. The efficiency of numerical methods for solving (6) depends on availability of information of its gradient and Hessian, and efficiency of the way to calculate them. Another related question is the well-posedness

of the problem (6) in the sense of the book by Dontchev and Zolezzi [6]. This leads us to consider a regularized minimization method for solving (1).

The paper is organized as follows: In section 2, we recall the Hessian formula due to Lewis and Sendov [16] for twice differentiable spectral functions. Moreover, we prove that the gradient map of a spectral function $(f \circ \lambda)$ is continuous at $X \in \mathcal{S}_n$ if and only if the symmetric function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable at $\lambda(X)$. We stress that the proof is interesting for its own as it can be revised to give a different yet simple proof for Lewis-Sendov Hessian formula for twice differentiable spectral functions. In section 3, we derive formulae for the gradient as well as for the Hessian of function θ_ε, and study their properties. In particular, we address the regularity conditions ensuring the nonsingularity of the Hessian. We also develop a more explicit matrix representation of the Hessian. The study on the well-posedness of the problem (6) leads to a regularization method for solving (1) with convergence analysis in section 4. We illustrate in section 5 the theoretical results with a combinatorial application, which covers the semidefinite programming relaxation of the max-cut problem. Finally, conclusions as well as some possible future research topics are given in section 6.

Notation: Vectors in $\mathbb{R}^n$ are viewed as columns and capital letters such as X, Y et.al. always denote matrices in $\mathcal{S}_n$. For $X \in \mathcal{S}_n$, we denote by X_{ij} the (i,j)th entry of X. We use $\circ$ to denote the Hadamard product between matrices, i.e.,

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

For $X \in \mathcal{S}_n$, $X \succ 0$ means that X is positive definite. Let the operator diag : $\mathcal{S}_n \to \mathbb{R}^n$ be defined by $\mathrm{diag}[X] := (X_{11}, \ldots, X_{nn})^T$, while for $\mu \in \mathbb{R}^n$, $\mathrm{Diag}[\mu_1, \ldots, \mu_n]$ will denote the diagonal matrix with its ith diagonal entry μ_i. Sometimes we write $\mathrm{Diag}[\mu]$ instead of $\mathrm{diag}[\mu_1, \ldots, \mu_n]$ for simplicity. Let $\mathcal{P}$ denote the set of all permutation matrices in $\mathbb{R}^{n\times n}$. For any given $\mu \in \mathbb{R}^n$, $\mathcal{P}_\mu$ denotes the stabilizer of μ defined by

$$\mathcal{P}_\mu := \{P \in \mathcal{P} |\ P\mu = \mu\}.$$

For any $\mu \in \mathbb{R}^n$ and $P \in \mathcal{P}$, we will frequently use the following fact.

$$\mathrm{Diag}[P\mu] = P\mathrm{Diag}[\mu]P^T.$$

Throughout, $\|\cdot\|$ denotes the Frobenius norm for matrices and the 2-norm for vectors. For any $x \in \mathbb{R}^n$, $X \in \mathcal{S}_n$ and any scalar $\gamma > 0$, we denote the γ-ball around X in $\mathcal{S}_n$ by

$$\mathcal{B}(X,\gamma) \ := \ \{Y \in \mathcal{S}_n |\ \|Y - X\| \le \gamma\}.$$

We let $\mathcal{A}^* : \mathcal{S}_n \to \mathbb{R}^m$ be the adjoint operator of the linear operator $\mathcal{A} : \mathbb{R}^m \to \mathcal{S}_n$ defined by (2) and satisfies for all $(d, D) \in \mathbb{R}^m \times \mathcal{S}_n$

$$d^T\mathcal{A}^*D := \langle D, \mathcal{A}d\rangle.$$

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

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

2 The formulae

Throughout this section, we let $f : \mathbb{R}^n \to \mathbb{R}$ be a symmetric function, usually twice differentiable. We will recall the formulae for the gradient and the Hessian of the twice continuously differentiable spectral functions $(f \circ \lambda) : \mathcal{S}_n \to \mathbb{R}$. Especially, we give an elementary proof concerning the continuity of the gradient map. So far we have not found a proof in the literature (it might be possible to prove the result following the approach by Lewis and Sendov, the resulting proof however would be much more complicated than ours.) It is also interesting to point out that the proof can be revised to give a different yet simple proof for Lewis-Sendov Hessian formula for twice differentiable spectral functions. First we review some useful perturbation results for the spectral decomposition of real symmetric matrices. These results will be used in our proof below.

Let $\mathcal{O}$ denote the group of $n \times n$ real orthogonal matrices. For each $X \in \mathcal{S}_n$, define the set of orthonormal eigenvectors of X by

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

Clearly $\mathcal{O}_X$ is nonempty for each $X \in \mathcal{S}_n$. The following lemma, proved in [5, Lemma 3] using results from [27, pp. 92 and 250], gives a key perturbation result for eigenvectors of symmetric matrices.

Lemma 2.1 *For any $X \in \mathcal{S}$, there exist scalars $\eta > 0$ and $\epsilon > 0$ such that*

$$\min_{P \in \mathcal{O}_X} \|P - Q\| \le \eta \|X - Y\| \quad \forall\ Y \in \mathcal{B}(X, \epsilon),\ \forall Q \in \mathcal{O}_Y. \tag{7}$$

We will use the following preliminary result obtained from the symmetry of f.

Lemma 2.2 *Let f be a symmetric function from $\mathbb{R}^n$ to $\mathbb{R}$. Then it holds that f is differentiable at $\mu \in \mathbb{R}^n$ if and only if f is differentiable at $P\mu$ for any $P \in \mathcal{P}$. In this case, $\nabla f(P\mu) = P\nabla f(\mu)$. If $P\mu = \mu$, then $\nabla f(\mu) = P^T \nabla f(\mu)$. Consequently, $(\nabla f(\mu))_i = (\nabla f(\mu))_j$ if $\mu_i = \mu_j$ for some $i, j \in \{1, \ldots, n\}$.*

Proof. The if part is trivial by just letting $P = I$. Suppose f is differentiable at $\mu \in \mathbb{R}^n$, i.e., $\nabla f(\mu)$ exists. Fixing a $P \in \mathcal{P}$, for any $h \in \mathbb{R}^n$, we have

$$f(P\mu + h) - f(P\mu) - (P\nabla f(\mu))^T h = f(\mu + P^T h) - f(\mu) - \nabla f(\mu)^T (P^T h) = o(\|h\|).$$

The last equality above uses the (Fréchet) differentiability of f at μ. This proves that $\nabla f(P\mu)$ exists, and in particular, $\nabla f(P\mu) = P\nabla f(\mu)$. This gives $\nabla f(\mu) = P^T \nabla f(\mu)$ for any $P \in \mathcal{P}$ satisfying $P\mu = \mu$. The last result is a direct consequence by choosing P as

$$P_{kl} = \begin{cases} 1 & \text{if } (k = i, l = j) \text{ or } (k = j, l = i) \\ 1 & \text{if } k = l \text{ but } k \ne i, j \\ 0 & \text{else} \end{cases} \qquad k, l = 1, \ldots, n.$$

□

Now we recall the formula for the gradient of a differential spectral function which is due to Lewis. But we were unable to find a proof in the literature for the continuity of the gradient map.

Proposition 2.3 *Let f be a symmetric function from $\mathbb{R}^n$ to $\mathbb{R}$ and $X \in \mathcal{S}_n$. Then the following hold:*

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

$$\nabla(f \circ \lambda)(X) = U\mathrm{Diag}[\nabla f(\lambda(X))]U^T, \quad \forall\, U \in \mathcal{O}_X. \tag{8}$$

More generally, the gradient of $(f \circ \lambda)$ has the following formula

$$\nabla(f \circ \lambda)(X) = V\mathrm{Diag}[\nabla f(\mu)]V^T, \tag{9}$$

for any orthogonal matrix $V \in \mathcal{O}$ and $\mu \in \mathbb{R}^n$ satisfying $X = V\mathrm{Diag}[\mu]V^T$.

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

Proof. (a) The formula (8) is discovered by Lewis [13, Thm. 1.1]. We only prove (9). Fix any $V \in \mathcal{O}$ and $\mu \in \mathbb{R}^n$ satisfying $X = V\mathrm{Diag}[\mu]V^T$, there must exist $P \in \mathcal{P}$ such that $\mu = P\lambda(X)$. By Lemma 2.2 (a), we have

$$\begin{array}{rl} & V\mathrm{Diag}[\nabla f(\mu)]V^T = V\mathrm{Diag}[\nabla f(P\lambda(X))]V^T \\ = & V\mathrm{Diag}[P\nabla f(\lambda(X))]V^T = VP\mathrm{Diag}[\nabla f(\lambda(X))]P^TV^T \end{array} \tag{10}$$

and

$$X = V\mathrm{Diag}[\mu]V^T = V\mathrm{Diag}[P\lambda(X)]V^T = VP\mathrm{Diag}[\lambda(X)]P^TV^T. \tag{11}$$

Apparently, $U := VP \in \mathcal{O}_X$. Then (10), (11) and (8) together imply (9).

(b) The part for the differentiability is contained in (a). We need only to prove the part for the continuity. We note that by Lemma 2.1 there exists $\eta > 0$ and $\epsilon > 0$ such that (7) holds at X.

Suppose f is continuously differentiable at $\lambda(X)$. Consider any $Y \in \mathcal{B}(X, \epsilon)$ and any $Q \in \mathcal{O}_Y$. Then by (7), there exists $P \in \mathcal{O}_X$ such that

$$\|P - Q\| \le \eta\|Y - X\|.$$

This relation means that $\|P - Q\| \to 0$ as $Y \to X$. Then (8), $\lambda(Y) \to \lambda(X)$ as $Y \to X$, and the continuity of ∇f at $\lambda(X)$ together yield

$$\begin{array}{rcl} \nabla(f \circ \lambda)(Y) - \nabla(f \circ \lambda)(X) & = & Q\mathrm{Diag}[\nabla f(\lambda(Y))]Q^T - P\mathrm{Diag}[\nabla f(\lambda(X))]P^T \\ & = & Q\mathrm{Diag}[\nabla f(\lambda(Y)) - \nabla f(\lambda(X))]Q^T \\ & + & (Q - P)\mathrm{Diag}[\nabla f(\lambda(X))]Q^T + P\mathrm{Diag}[\nabla f(\lambda(X))](Q - P)^T \\ & \to & 0 \qquad \text{as } \ Y \to X. \end{array}$$

Thus $\nabla(f \circ \lambda)$ is continuous at X.

Suppose instead $(f \circ \lambda)$ is continuously differentiable at X. Without loss of generality, we assume that $(f \circ \lambda)$ is differentiable in $\mathcal{B}(X, \epsilon)$ (if necessary, shrinking ϵ). Fix $Q \in \mathcal{O}_X$. Let $h \in \mathbb{R}^n$ and

$$H := Q\mathrm{Diag}[h_1, \ldots, h_n]Q^T$$

so that $\lambda(X)+h$ consists of eigenvalues of $X+H$. We note that $(f \circ \lambda)$ is differentiable at $X+H$ whenever $H \in \mathcal{B}(X, \epsilon)$. Equivalently, by (a) of this proposition, f is differentiable at $\lambda(X+H)$ whenever $\|h\| \leq \epsilon$. Lemma 2.2 implies that f is differentiable at $\lambda(X)+h$ since there exists $P \in \mathcal{P}$ satisfying $\lambda(X)+h = P\lambda(X+H)$. Moreover, any limit of P as $h \to 0$ belongs to $\mathcal{P}_{\lambda(X)}$. It follows from Lemma 2.1 that for any $U \in \mathcal{O}_{X+H}$ there exists $V \in \mathcal{O}_X$ such that $\|U-V\| \leq \eta$ and $U^TV \to I$ as $h \to 0$. Then the continuity of $\nabla(f \circ \lambda)$ at X and the formula (8) yield

$$\begin{aligned}
\nabla(f \circ \lambda)(X) &= \lim_{H \to 0} \nabla(f \circ \lambda)(X+H) \\
&= \lim_{H \to 0} U\mathrm{Diag}[\nabla f(\lambda(X+H))]U^T \\
&= \lim_{h \to 0} U\mathrm{Diag}[\nabla f(P^T(\lambda(X)+h))]U^T \\
&= \lim_{h \to 0} U\mathrm{Diag}[P^T\nabla f(\lambda(X)+h)]U^T \\
&= \lim_{h \to 0} UP^T\mathrm{Diag}[\nabla f(\lambda(X)+h)]PU^T.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathrm{Diag}[\nabla f(\lambda(X))] &= V^T\nabla(f \circ \lambda)(X)V \\
&= \lim_{h \to 0} V^TUP^T\mathrm{Diag}[\nabla f(\lambda(X)+h)]PU^TV.
\end{aligned}$$

Taking into account of the facts that $U^TV \to I$ as $h \to 0$ and any limit of P as $h \to 0$ must belong to $\mathcal{P}_{\lambda(X)}$, we have from the above relations that, any limit of $\mathrm{Diag}[\nabla f(\lambda(X)+h)]$ must belong to the set

$$\left\{P\mathrm{Diag}[\nabla f(\lambda(X))]P^T |\ P \in \mathcal{P}_{\lambda(X)}\right\}.$$

For any $P \in \mathcal{P}_{\lambda(X)}$, we have from Lemma 2.2 that

$$P\mathrm{Diag}[\nabla f(\lambda(X))]P^T = \mathrm{Diag}[P\nabla f(\lambda(X))] = \mathrm{Diag}[\nabla f(P\lambda(X))] = \mathrm{Diag}[\nabla f(\lambda(X))].$$

Thus this set contains only one element $\mathrm{Diag}[\nabla f(\lambda(X))]$ so that

$$\lim_{h \to 0} \mathrm{Diag}[\nabla f(\lambda(X)+h)] = \mathrm{Diag}[\nabla f(\lambda(X))],$$

or equivalently $\lim_{h \to 0} \nabla f(\lambda(X)+h) = \nabla f(\lambda(X))$ establishing the continuity of ∇f at $\lambda(X)$. □

In a recent paper [16], Lewis and Sendov found a formula for calculating the Hessian of the spectral function $(f \circ \lambda)$, when it exists, via calculating the Hessian of f. This facilitates the possibility of using second-order methods for solving the convex minimization problem

which we will study later. Suppose that f is twice differentiable at $\mu \in \mathbb{R}^n$. Define the matrix $\mathcal{C}(\mu) \in \mathbb{R}^{n\times n}$:

$$(\mathcal{C}(\mu))_{ij} := \begin{cases} 0 & \text{if } \ i = j \\ (\nabla^2 f(\mu))_{ii} - (\nabla^2 f(\mu))_{ij} & \text{if } \ i \neq j \text{ and } \mu_i = \mu_j \\ \dfrac{(\nabla f(\mu))_i - (\nabla f(\mu))_j}{\mu_i - \mu_j} & \text{else.} \end{cases} \tag{12}$$

It is easy to see that $\mathcal{C}(\mu)$ is symmetric due to the symmetry of f. The following result is proved by Lewis and Sendov [16, Thm. 3.3, 4.2].

Proposition 2.4 *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*

$$\nabla^2(f \circ \lambda)(X)[H] = U\left(\mathrm{Diag}[\nabla^2 f(\lambda(X))\mathrm{diag}[\tilde{H}]] + \mathcal{C}(\lambda(X)) \circ \tilde{H}\right) U^T, \quad \forall H \in \mathcal{S}_n \tag{13}$$

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

Remark. We stress that the formulae (8) and (13) do not depend on the particular choice of $U \in \mathcal{O}_X$.

3 Smooth convex minimization problems

Through the rest of the paper, we let f be the maximum function defined by (3), f_ε be the penalty function defined by (4), and θ_ε be the merit function in the convex minimization problem (6). The primal task in this section is to derive explicit formulae for the gradient and the Hessian of θ_ε so that the numerical methods for (6) can take advantage of those information. We then discuss the regularity conditions ensuring the nonsingularity of the Hessian of θ_ε. Finally, we give a matrix representation for the Hessian.

3.1 Basic calculation

The following result is specialization of the results of [21, Lemma 7] and [24, Prop. 3.1] on a general form of f_ε to the maximum function (3), and can also be calculated directly.

Lemma 3.1 *The following results hold:*

(a) For any $\varepsilon_1 > \varepsilon_2 > 0$, $f_{\varepsilon_1}(x) \geq f_{\varepsilon_2}(x)$ for all $x \in \mathbb{R}^n$.

(b) The gradient of f_ε is given by

$$\nabla f_\varepsilon(x) = \mu(\varepsilon, x) := (\mu_1(\varepsilon, x), \ldots, \mu_n(\varepsilon, x))^T \tag{14}$$

where for $i = 1, \ldots, n$

$$\mu_i(\varepsilon, x) = \frac{e^{x_i/\varepsilon}}{\sum_{j=1}^n e^{x_j/\varepsilon}} \in (0, 1] \quad \text{and} \quad \sum_{i=1}^n \mu_i(\varepsilon, x) = 1. \tag{15}$$

(c) The Hessian of f_ε is given by

$$\nabla^2 f_\varepsilon(x) = \frac{1}{\varepsilon}\left(\mathrm{Diag}[\mu(\varepsilon,x)] - \mu(\varepsilon,x)(\mu(\varepsilon,x))^T\right). \tag{16}$$

It follows Proposition 2.3 that the spectral function $(f_\varepsilon \circ \lambda)$ is continuously differentiable in $\mathcal{S}_n$. We now relate the gradient of $(f_\varepsilon \circ \lambda)$ to the subdifferential of $\lambda_1(\cdot)$ (it is well known that $\lambda_1(\cdot)$ is convex and hence its subdifferential is well defined at any point $X \in \mathcal{S}_n$.) In fact, it is shown in [19, Thm. 1] that

$$\partial\lambda_1(X) = \left\{Q_1 Y Q_1^T :\ Y \in \mathcal{C}_r\right\} \tag{17}$$

where Q_1 is an $n \times r$ matrix whose columns form an orthogonal basis of the eigenspace associated with $\lambda_1(X)$ (it has dimension r), and $\mathcal{C}_r$ is the so-called spectraplex of $\mathcal{S}_r$:

$$\mathcal{C}_r := \{V \in \mathcal{S}_r :\ V \text{ is positive semidefinite, } \mathrm{tr}(V) = 1\}. \tag{18}$$

The following result means that the gradient $\nabla(f_\varepsilon \circ \lambda)(X)$ is an approximate element with respect to $\partial\lambda_1(X)$.

Proposition 3.2 *For any $X \in \mathcal{S}_n$, we have*

$$\lim_{\varepsilon\to 0} \nabla(f_\varepsilon \circ \lambda)(X) \in \partial\lambda_1(X).$$

Proof. Let $X \in \mathcal{S}_n$ and $X = Q\mathrm{Diag}[\lambda(X)]Q^T$, $Q \in \mathcal{O}_X$. Denote $\lambda := \lambda(X)$ for simplicity. Assume that X has multiplicity r of the largest eigenvalue λ_1. Then the first r columns of Q must form an orthogonal basis of the eigenspace associated with λ_1. These columns form a matrix, say $Q_1 \in \mathbb{R}^{n\times r}$.

It follows from (8) and (14) that

$$\nabla(f_\varepsilon \circ \lambda)(X) = Q\mathrm{Diag}[\nabla f_\varepsilon(\lambda)]Q^T = Q\mathrm{Diag}[\mu(\epsilon,\lambda)]Q^T,$$

with

$$\mu_i(\varepsilon,\lambda) = \frac{e^{\lambda_i/\varepsilon}}{\sum_{j=1}^n e^{\lambda_j/\varepsilon}} = \frac{e^{(\lambda_i-\lambda_1)/\varepsilon}}{\sum_{j=1}^n e^{(\lambda_j-\lambda_1)/\varepsilon}}.$$

Noticing that λ has multiplicity r of λ_1 and λ_1 is the largest element in λ, we obtain

$$\lim_{\varepsilon\to 0}\mu_i(\varepsilon,\lambda) = \begin{cases} \frac{1}{r} & i = 1,\ldots,r \\ 0 & i = r+1,\ldots,n. \end{cases}$$

Therefore

$$\lim_{\varepsilon\to 0}\nabla(f_\varepsilon \circ \lambda)(X) = Q\mathrm{Diag}[(\frac{1}{r},\ldots,\frac{1}{r},0,\ldots,0)]Q^T = \frac{1}{r}Q_1 I_r Q_1^T,$$

where I_r is the identity matrix in $\mathcal{S}_r$. Our result then follows from the characterization (17) and (18) of subdifferential of $\lambda_1(\cdot)$ at X. $\square$

The next step is to develop the formulae for the gradient and the Hessian of θ_ε. Viewed as the composition of the twice differentiable spectral function $(f_\varepsilon \circ \lambda)$ and the affine mapping $A(\cdot) : \mathbb{R}^m \to \mathcal{S}_n$, it seems that we can use the chain rule to obtain the formulae from (8) and (13). We stress however that it is not straightforward to use the chain rule particularly due to the special formulation of the Hessian formula (13) of twice differentiable spectral functions. So we give a complete characterization of the formulae in the next result.

Proposition 3.3 *Let $\varepsilon > 0$ be given. Then the followings hold:*

(a) The function $\theta_\varepsilon(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is continuously differentiable, and the gradient of $\theta_\varepsilon(\cdot)$ at $y \in \mathbb{R}^m$ is given by

$$\nabla\theta_\varepsilon(y) = \mathcal{A}^* \left(U(\mathrm{Diag}[\nabla f_\varepsilon(\lambda(A(y)))])U^T \right), \qquad \forall\, U \in \mathcal{O}_{A(y)}.$$

(b) The function $\theta_\varepsilon(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is twice continuously differentiable, and the Hessian of $\theta_\varepsilon(\cdot)$ at $y \in \mathbb{R}^m$ is given by

$$\nabla^2\theta_\varepsilon(y)[h] = \mathcal{A}^* \left(\nabla^2(f_\varepsilon \circ \lambda)(A(y))[\mathcal{A}h] \right), \qquad \forall\, h \in \mathbb{R}^m$$

where

$$\nabla^2(f_\varepsilon \circ \lambda)(A(y))[\mathcal{A}h] = U \left(\mathrm{Diag}[\nabla^2 f_\varepsilon(\lambda(A(y)))\mathrm{diag}[\tilde{H}]] + \mathcal{C} \circ \tilde{H} \right) U^T \qquad (19)$$

and U is any orthogonal matrix in $\mathcal{O}_{A(y)}$, $H = \mathcal{A}h$, $\tilde{H} = U^T H U$, and $\mathcal{C} = \mathcal{C}(\lambda(A(y)))$ is defined as in (12) with f replaced by f_ε.

Proof. (a) Let $y \in \mathbb{R}^m$ be given. It follows from Proposition 2.3 (a) that $(f_\varepsilon \circ \lambda)$ is differentiable at $A(y)$. Then for any $h \in \mathbb{R}^m$ we have

$$\begin{aligned}
\theta_\varepsilon(y+h) - \theta_\varepsilon(y) &= (f_\varepsilon \circ \lambda \circ A)(y+h) - (f_\varepsilon \circ \lambda \circ A)(y) \\
&= (f_\varepsilon \circ \lambda)(A(y+h)) - (f_\varepsilon \circ \lambda)(A(y)) \\
&= (f_\varepsilon \circ \lambda)(A(y) + \mathcal{A}h) - (f_\varepsilon \circ \lambda)(A(y)) \\
&= \langle \nabla(f_\varepsilon \circ \lambda)(A(y)), \mathcal{A}h \rangle + o(\|h\|) \\
&= \sum_{i=1}^m \langle \nabla(f_\varepsilon \circ \lambda)(A(y)), A_i h_i \rangle + o(\|h\|) \\
&= \sum_{i=1}^m h_i \langle \nabla(f_\varepsilon \circ \lambda)(A(y)), A_i \rangle + o(\|h\|) \\
&= \langle \mathcal{A}^* \left(\nabla(f_\varepsilon \circ \lambda)(A(y)) \right), h \rangle + o(\|h\|).
\end{aligned}$$

Hence, (8) gives

$$\nabla\theta_\varepsilon(y) = \mathcal{A}^* \left(\nabla(f_\varepsilon \circ \lambda)(A(y)) \right) = \mathcal{A}^* \left(U(\mathrm{Diag}[\nabla f_\varepsilon(\lambda(A(y)))])U^T \right)$$

where U is any orthogonal matrix in $\mathcal{O}_{A(y)}$. From above, we see that the continuity of $\nabla\theta_\varepsilon$ depends on the continuity of $\nabla(f_\varepsilon \circ \lambda)(A(\cdot))$. Hence, the continuity of $\nabla(f_\varepsilon \circ \lambda)(\cdot)$ (Proposition 2.3 (b)) and the continuity of $A(\cdot)$ imply the continuity of $\nabla\theta_\varepsilon(\cdot)$.

(b) It is proved in (a) that θ_ε is continuously differentiable. Also we note from Proposition 2.4 that $(f_\varepsilon \circ \lambda)$ is twice continuously differentiable since f_ε is twice continuously differentiable. Hence, for any $y, h \in \mathbb{R}^m$, we have

$$\begin{aligned} \nabla\theta_\varepsilon(y+h) - \nabla\theta_\varepsilon(y) &= \mathcal{A}^* \left(\nabla(f_\varepsilon \circ \lambda)(A(y) + \mathcal{A}h) - \nabla(f_\varepsilon \circ \lambda)(A(y))\right) \\ &= \mathcal{A}^* \left(\nabla^2(f_\varepsilon \circ \lambda)(A(y))[\mathcal{A}h]\right) + o(\|h\|) \end{aligned}$$

and $\nabla^2(f_\varepsilon \circ \lambda)(A(y))[\mathcal{A}h]$ is given in (19). It is easy to see that the operator $\nabla^2\theta_\varepsilon : \mathbb{R}^m \to \mathcal{S}_m$ given by

$$\nabla^2\theta_\varepsilon(y)[h] := \mathcal{A}^* \left(\nabla^2(f_\varepsilon \circ \lambda)(A(y))[\mathcal{A}h]\right), \qquad \forall\, h \in \mathbb{R}^m$$

is linear (it is also easy to prove the operator is symmetric). By the definition of Fréchet differentiability, we know that θ_ε is twice differentiable and it has the Hessian formula as given in (b). Moreover, from the formula, we see that the continuity of $\nabla^2\theta_\varepsilon$ depends on the continuity of the Hessian of the spectral function $(f_\varepsilon \circ \lambda)$, which has been stated in Proposition 2.4. This completes our proof. □

We now develop a more explicit formula for calculating $\nabla\theta_\varepsilon(y)$. Let $U \in \mathcal{O}_{A(y)}$ and define

$$\tilde{A}_i := U^T A_i U, \qquad i = 1, \ldots, m.$$

Let $z := \lambda(A(y))$ and $\mu := \mu(\varepsilon, z)$ (cf. (15)), then

$$\langle A_i,\ U(\mathrm{Diag}[\nabla f_\varepsilon(\lambda(A(y)))])U^T\rangle = \langle U^T A_i U,\ \mathrm{Diag}[\mu]\rangle = \mu^T \mathrm{diag}[\tilde{A}_i].$$

We further let $W \in \mathbb{R}^{n\times m}$ with its ith column given by the vector $\mathrm{diag}[\tilde{A}_i]$, i.e.,

$$W(:, i) := \mathrm{diag}[\tilde{A}_i], \qquad i = 1, \ldots, m. \tag{20}$$

Then taking into account of the formula of $\nabla\theta_\varepsilon(y)$ in Proposition 3.3 (a), we have

$$\nabla\theta_\varepsilon(y) = W^T\mu. \tag{21}$$

In the next two subsections, we will pay attention to the Hessian of θ_ε, study its nonsingularity and develop a matrix representation for it.

3.2 Nonsingularity

The nonsingularity of the Hessian matrix $\nabla^2\theta_\varepsilon(y)$ plays a very important role in Newton-type methods for solving the (smooth) convex problem (6). It is clear to see that the Hessian matrix is split into two parts, namely

$$\nabla^2\theta_\varepsilon(y) = \mathcal{T}_1(y) + \mathcal{T}_2(y)$$

where $\mathcal{T}_1, \mathcal{T}_2 : \mathbb{R}^m \to \mathcal{S}_n$ are defined respectively by

$$\mathcal{T}_1(y)[h] := \mathcal{A}^* \left(U(\mathrm{Diag}(\nabla^2 f_\varepsilon(\lambda(A(y)))\mathrm{diag}[\tilde{H}]))U^T \right), \qquad \forall\, h \in \mathbb{R}^m$$
$$\mathcal{T}_2(y)[h] := \mathcal{A}^* \left(U(\mathcal{C} \circ \tilde{H})U^T \right), \qquad \forall\, h \in \mathbb{R}^m$$

where $U \in \mathcal{O}_{A(y)}$, $H = \mathcal{A}h$, $\tilde{H} = U^T H U$ and $\mathcal{C}$ is defined by $\mathcal{C}(\lambda(A(y)))$ as in (12) with f being replaced by f_ε. An implicit fact used in the above definitions is the independence of the choice of $U \in \mathcal{O}_{A(y)}$. Since θ_ε is convex, $\nabla^2\theta_\varepsilon$ is positive semidefinite. We will see that it can be split into two positive semidefinite operators, i.e., $\mathcal{T}_1$ and $\mathcal{T}_2$. If one of them is positive definite, so is $\nabla^2\theta_\varepsilon$. This provides a way to study the nonsingularity of $\nabla^2\theta_\varepsilon$.

Proposition 3.4 *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be defined as above, then both of them are positive semidefinite operators. And $\nabla^2\theta_\epsilon(y)$ is positive definite iff the matrices $A_1, \ldots, A_m$ are linearly independent and $I \notin Range(\mathcal{A})$, where $Range(\mathcal{A}) = \{\mathcal{A}h | h \in \mathbb{R}^m\}$ and I is the identity matrix.*

Proof. We first prove the positive semidefiniteness of $\mathcal{T}_1(\cdot)$. It suffices to show that for any given $y \in \mathbb{R}^m$

$$\langle h, \mathcal{T}_1(y)[h]\rangle \geq 0, \qquad \forall\, h \in \mathbb{R}^m.$$

It is straightforward to see

$$\begin{aligned}
\langle h, \mathcal{T}_1(y)[h]\rangle &= \langle h, \mathcal{A}^* \left(U(\mathrm{Diag}(\nabla^2 f_\varepsilon(\lambda(A(y)))\mathrm{diag}[\tilde{H}]))U^T \right)\rangle \\
&= \langle \mathcal{A}h, U(\mathrm{Diag}(\nabla^2 f_\varepsilon(\lambda(A(y)))\mathrm{diag}[\tilde{H}]))U^T\rangle \\
&= \langle U^T(\mathcal{A}h)U, \mathrm{Diag}(\nabla^2 f_\varepsilon(\lambda(A(y)))\mathrm{diag}[\tilde{H}])\rangle \\
&= (\mathrm{diag}[\tilde{H}])^T \nabla^2 f_\varepsilon(\lambda(A(y)))\mathrm{diag}[\tilde{H}] \\
&\geq 0.
\end{aligned}$$

The last inequality used the semidefiniteness of $\nabla^2 f_\varepsilon(\cdot)$ as f_ε is a convex function.

Now we prove $\mathcal{T}_2$ is positive semidefinite. For simplicity, we let $z := \lambda(A(y))$. Then taking into account of Lemma 3.1 and (12), we obtain

$$\mathcal{C} = \mathcal{C}(z) = \begin{cases} 0 & \text{if } i = j \\ \frac{1}{\varepsilon}\mu_i(\varepsilon, z) & \text{if } i \neq j,\ z_i = z_j \\ \dfrac{\mu_i(\varepsilon, z) - \mu_j(\varepsilon, z)}{z_i - z_j} & \text{if } z_i \neq z_j. \end{cases} \tag{22}$$

Here we used the relation

$$(\nabla^2 f_\varepsilon(z))_{ii} = \mu_i(\varepsilon, z) - (\mu_i(\varepsilon, z))^2, \quad (\nabla^2 f_\varepsilon(z))_{ij} = -\mu_i(\varepsilon, z)\mu_j(\varepsilon, z)$$

and $\mu_i(\varepsilon, z) = \mu_j(\varepsilon, z)$ if $z_i = z_j$ and $\mu_i(\varepsilon, z) > \mu_j(\varepsilon, z)$ if $z_i > z_j$. Thus, $\mathcal{C}_{ij} \geq 0$ for all $i, j = 1, \ldots, n$. We also note that for any $B, C, D \in \mathcal{S}_n$ it is easy to verify that

$$\langle B, C \circ D\rangle = \langle C, B \circ D\rangle = \langle D, B \circ C\rangle.$$

Using this and the nonnegativity of all elements of $\mathcal{C}$, we have for any $h \in \mathbb{R}^m$

$$\begin{aligned}
\langle h, \mathcal{T}_2(y)[h]\rangle &= \langle h, \mathcal{A}^*\left(U(\mathcal{C}\circ\tilde{H})U^T\right)\rangle \\
&= \langle \mathcal{A}h, U(\mathcal{C}\circ\tilde{H})U^T\rangle \\
&= \langle U^T(\mathcal{A}h)U, \mathcal{C}\circ\tilde{H}\rangle \\
&= \langle \tilde{H}, \mathcal{C}\circ\tilde{H}\rangle \\
&= \langle \mathcal{C}, \tilde{H}\circ\tilde{H}\rangle \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}(\mathcal{C}_{ij})(\tilde{H}_{ij})^2 \geq 0.
\end{aligned}$$

This proves the semidefiniteness of $\mathcal{T}_2$.

From the above argument, we see that $\nabla^2\theta_\varepsilon(y)$ is positive semidefinite and it is positive definite iff there exist no nonzero h such that $\langle h, \mathcal{T}_1[h]\rangle = \langle h, \mathcal{T}_2[h]\rangle = 0$. Now note that $\langle h, \mathcal{T}_1[h]\rangle = 0$ iff $\tilde{H} = \rho I$ for some $\rho \in \mathbb{R}$ due to the special structure of $\nabla^2 f_\epsilon$, and $\langle h, \mathcal{T}_2[h]\rangle = 0$ iff $\tilde{H}_{ij} = 0$ for $i \neq j$ since $\mathcal{C}_{ij} > 0$ for $i \neq j$. Hence $\nabla^2\theta_\varepsilon[h] = 0$ for nonzero $h \in \mathbb{R}^m$ iff $\tilde{H} = \rho I$ for some $\rho \neq 0$. The possibility of $\rho = 0$ is removed by the linear independence of the matrices $A_1, \ldots, A_m$. But $\tilde{H} = \rho I$ is equivalent to $I \in Range(\mathcal{A})$. Thus we complete the proof. □

The condition $I \notin Range(\mathcal{A})$ is not restrictive as it looks like. In fact, it can be dropped from the statement by using a transformation procedure below, as long as the problem (1) has a solution. To see this, we first present a lemma which is also useful in a regularization method proposed in the next section. Let

$$\nu := \inf_{y\in\mathbb{R}^m} \lambda_1(A(y)).$$

Lemma 3.5 *The optimal value ν is finite iff $\lambda_1(\mathcal{A}y) \geq 0$ for any $y \in \mathbb{R}^m$. Moreover, the solution set of the problem (1), denoted by Ω_0, is nonempty and bounded iff $\lambda_1(\mathcal{A}y) > 0$ for any $0 \neq y \in \mathbb{R}^m$.*

Proof. If there exists one $y \in \mathbb{R}^m$ such that $\lambda_1(\mathcal{A}y) < 0$, then $\lambda_1(A(ty)) \to -\infty$ as $t \to \infty$, therefore $\nu = -\infty$. On the other hand, if $\nu = -\infty$, there must exist an $y \in \mathbb{R}^m$ such that $\lambda_1(\mathcal{A}y) < 0$.

Now we prove the second part. The "if" direction is easy as the level set of the problem (1)

$$\mathcal{L}(\alpha) := \{y \in \mathbb{R}^m |\ \lambda_1(A(y)) \leq \alpha\}$$

with $\alpha = \lambda_1(A_0)$ is nonempty and compact. So we only prove the "only if" direction. Suppose that there is $y^* \in \Omega_0$ and Ω_0 is bounded. It follows from the first part of the lemma, $\lambda_1(\mathcal{A}(y)) \geq 0$ for any $y \in \mathbb{R}^m$. If $\lambda_1(\mathcal{A}y) = 0$ for some $0 \neq y \in \mathbb{R}^m$, then for any $t \geq 0$,

$$\lambda_1(A(y^* + ty)) = \lambda_1(A(y^*) + t\mathcal{A}y) \leq \lambda_1(A(y^*)) + t\lambda_1(\mathcal{A}y) = \lambda_1(A(y^*)).$$

If the equality holds for all $t \geq 0$, then $y^* + ty \in \Omega_0$ for all $t \geq 0$, contradicting the boundedness of Ω_0. Therefore, we must have

$$\lambda_1(A(y^* + t^* y) < \lambda_1(A(y^*))$$

for some large $t^* > 0$. That is, $y^* + t^* y$ is a point which yields a value less than that given by y^*, and hence y^* could not be a solution of (1), a contradiction. This finishes the proof for the "only if" part □

If $A_0 \succ 0$, i.e., A_0 is positive definite, then $A(y) \succ \mathcal{A}y$, therefore $\lambda_1(A(y)) > \lambda_1(\mathcal{A}y)$. If in addition ν is finite, then we have $\nu > 0$ which follows from Lemma 3.5. So from now on, we can assume $A_0 \succ 0$ without loss of generality; otherwise, we can change A_0 to $A_0 + \rho I$ such that the latter matrix is positive definite, while the optimal solution set is the same. With the assumption that $A_0 \succ 0$, we know $\nu > 0$ or $\nu = -\infty$.

Now we can augment $A_i \in \mathcal{S}_n$ to $\bar{A}_i \in \mathcal{S}_{n+1}$ for all i as follows:

$$\bar{A}_i = \begin{pmatrix} A_i & 0 \\ 0^T & 0 \end{pmatrix}.$$

Consider a new eigenvalue optimization problem:

$$\bar{\nu} := \inf_{y \in \mathbb{R}^m} \ \lambda_1(\bar{A}(y))$$

where

$$\bar{A}(y) := \bar{A}_0 + \bar{\mathcal{A}}y = \bar{A}_0 + \sum_{i=1}^{m} \bar{A}_i y_i.$$

We can see that $\bar{\nu} = \nu$ if $\nu > 0$ and $\bar{\nu} = 0$ if $\nu = -\infty$ with the assumption $A_0 \succ 0$. Now using the exponential penalty function to the new problem and suppress the constant eigenvalue 0 of $\bar{A}(y)$, we have

$$\bar{f}_\epsilon(y) = \epsilon \ln(1 + \sum_{i=1}^{n} e^{y_i/\epsilon}).$$

Note that $\bar{f}_\epsilon$ is an approximation of

$$\bar{f}(y) = \max\{0, y_1, \cdots, y_n\}.$$

Define

$$\bar{\theta}_\epsilon(y) := \bar{f}_\epsilon(\lambda_1(A(y))).$$

We know that the Hessian of $\bar{\theta}_\epsilon$ is positive definite everywhere following the proof of Proposition 3.4 if $\{A_1, \ldots, A_m\}$ are linearly independent. In the above statement, we have used an obvious fact: $I_{n+1} \notin Range(\bar{\mathcal{A}})$.

It is clear that the augmentation step is only a process for understanding that what matters the problem (6) is the linear independence condition as long as the original problem has a solution. In some cases, it is easy to remove some matrices from $A_1, \ldots, A_m$ so that they become linearly independent. We will see this happens in our combinatorial applications. However, in some other cases, it is sometimes hard to detect the linear independence. For the latter case, in the next section, we will use a regularization technique to enforce the problem (6) to be well-posed.

3.3 Matrix representation

It is ideal to have a matrix representation $V \in \mathcal{S}_m$ for $\nabla^2\theta_\varepsilon(y)$ so that for any $h \in \mathbb{R}^m$,

$$\nabla^2\theta_\varepsilon(y)[h] = Vh. \tag{23}$$

Whence the Newton equation

$$Vh = -\nabla\theta_\varepsilon(y)$$

can be solved in a number of ways [10]. To this end, let $U \in \mathcal{O}_{A(y)}$ be used in the definitions of $\mathcal{T}_1$ and $\mathcal{T}_2$, and recall that

$$\tilde{A}_i := U^T A_i U \qquad i = 1, \ldots, m.$$

Note that

$$\tilde{H} = U^T(\mathcal{A}h)U = U^T(\sum_{j=1}^m A_j h_j)U = \sum_{j=1}^m \tilde{A}_j h_j.$$

Let matrix $W \in \mathbb{R}^{n\times m}$ be defined by (20). Then

$$\mathrm{diag}[\tilde{H}] = \mathrm{diag}[\sum_{j=1}^m \tilde{A}_j h_j] = Wh.$$

For simplicity, further let $B := \nabla^2 f_\varepsilon(\lambda(A(y)))$. It follows that for $i = 1, \ldots, m$

$$\begin{aligned}
(\mathcal{T}_1(y)[h])_i &= \langle A_i,\ U(\mathrm{Diag}(B\mathrm{diag}[\tilde{H}]))U^T\rangle \\
&= \langle U^T A_i U,\ \mathrm{Diag}[BWh]\rangle \\
&= \langle \tilde{A}_i,\ \mathrm{Diag}[BWh]\rangle \\
&= \sum_{j=1}^n (\tilde{A}_i)_{jj}(BWh)_j.
\end{aligned}$$

Putting in matrix form, we have

$$\mathcal{T}_1(y)[h] = \left(W^T B W\right) h. \tag{24}$$

Now we consider the linear operator $\mathcal{T}_2$. For $i = 1, \ldots, m$, we have

$$\begin{aligned}
(\mathcal{T}_2(y)[h])_i &= \langle A_i,\ U(\mathcal{C} \circ \tilde{H})U^T\rangle \\
&= \langle U^T A_i U,\ \mathcal{C} \circ \tilde{H}\rangle \\
&= \langle \tilde{A}_i,\ \mathcal{C} \circ \tilde{H}\rangle \\
&= \langle \tilde{H},\ \tilde{A}_i \circ \mathcal{C}\rangle \\
&= \sum_{j=1}^m \langle \tilde{A}_j,\ \tilde{A}_i \circ \mathcal{C}\rangle h_j \\
&= \sum_{j=1}^m \langle \mathcal{C},\ \tilde{A}_i \circ \tilde{A}_j\rangle h_j.
\end{aligned}$$

Putting in matrix form, we then have

$$\mathcal{T}_2(y)[h] = Mh, \tag{25}$$

where $M \in \mathcal{S}_m$ is defined by

$$M_{ij} := \langle \mathcal{C},\ \tilde{A}_i \circ \tilde{A}_j \rangle \qquad i, j = 1, \ldots, m. \tag{26}$$

Putting (24) and (25) together, we have the matrix representation (23) with $V \in \mathcal{S}_m$ given by

$$V := W^T B W + M. \tag{27}$$

Proposition 3.4 implies that both the matrix $W^T B W$ and M are positive semidefinite. In fact the positive semidefiniteness of $W^T B W$ follows directly from that of B since B is always positive semidefinite (not positive definite), see the structure of B in (15). We restate Proposition 3.4 in terms of V.

Proposition 3.6 *Let V be defined as above depending on given $y \in \mathbb{R}^m$. Then*

$$\nabla^2 \theta_\varepsilon(y) = V.$$

Moreover, if V is positive definite iff the matrices $A_1, \ldots, A_m$ are linearly independent and $I \notin Range(\mathcal{A})$.

To see clearly the implication of Proposition 3.6, let us consider the simplest case that $A(y)$ has multiplicity one of the largest eigenvalue, i.e., $\lambda_1(A(y)) > \lambda_2(A(y))$. In this case, the function $\lambda_1(A(\cdot))$ is twice continuously differentiable. After calculation by using the formula (13) and repeating similar arguments as for the operator $\mathcal{T}_2$, we see that

$$\nabla^2 \lambda_1(A(y)) = M$$

where M is defined as in (26) with $\mathcal{C}$ replaced by

$$\mathcal{C} := \begin{pmatrix} 0 & \frac{1}{\lambda_1 - \lambda_2} & \frac{1}{\lambda_1 - \lambda_3} & \cdots & \frac{1}{\lambda_1 - \lambda_n} \\ \frac{1}{\lambda_1 - \lambda_2} & 0 & 0 & \cdots & 0 \\ \frac{1}{\lambda_1 - \lambda_3} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\lambda_1 - \lambda_n} & 0 & 0 & \cdots & 0 \end{pmatrix}$$

and $\lambda_i := \lambda_i(A(y)), i = 1, \ldots, n$. We stress that M is independent of choice of $U \in \mathcal{O}_{A(y)}$ due to the structure of $\mathcal{C}$. Then conditions ensuring the nonsingularity of M becomes the same conditions which guarantees the nonsingularity of the Hessian of $\lambda_1(A(\cdot))$ at y, which is the case for optimization problems with twice continuously differentiable data.

However, for the degenerate case, i.e., $A(y)$ has multiplicity r ($r > 1$) of the largest eigenvalue, the situation becomes a bit complicated. For simplicity, let again $z := \lambda(A(y))$.

Hence, $z_i = \lambda_1(A(y))$ for $i = 1, \ldots, r$. We are interested in the case where ε is small. It follows from (15) that

$$\lim_{\varepsilon \to 0} \mu_i(\varepsilon, z) = \begin{cases} \frac{1}{r} & i = 1, \ldots, r \\ 0 & i = r+1, \ldots, n. \end{cases}$$

We observe from (22) that some elements of $\mathcal{C}$ grow to infinity as ε approaches to zero. The same problem occurs with the matrix $\nabla^2 f_\varepsilon(y)$. Fortunately, those two matrices are well scaled, i.e., they are magnitude of $1/\varepsilon$. Let $\mathcal{C}_0, \mathcal{D}_0 \in \mathcal{S}_n$ be defined by

$$\mathcal{C}_0 := \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathcal{D}_0 := \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

where $C := \frac{1}{r}(E_r - I_r)$, $D := \frac{1}{r^2}(rI_r - E_r)$, and E_r is the matrix of all ones in $\mathcal{S}_r$ and I_r is the identity matrix in $\mathcal{S}_r$. It is easy to see from (22) and (16) that

$$\lim_{\varepsilon \to 0} \varepsilon \mathcal{C} = \mathcal{C}_0 \quad \text{and} \quad \lim_{\varepsilon \to 0} \varepsilon \nabla^2 f_\varepsilon(y) = \mathcal{D}_0. \tag{28}$$

The limits in (28) mean that eventually the eigenvectors corresponding to the largest eigenvalue of $A(y)$ dominate the Hessian of $\theta_\varepsilon(y)$ as ε goes to zero.

4 A regularization method

In previous sections, we have seen that the function $\theta_\varepsilon(\cdot)$ is well conditioned in the sense that it is twice continuously differentiable and strictly convex given the linear independence condition. Moreover, we have

$$\lim_{\varepsilon \to 0} \theta_\varepsilon(y) = \lambda_1(A(y)), \qquad \forall y \in \mathbb{R}^m.$$

This suggests that, instead of minimizing the nondifferentiable function $\lambda_1(A(\cdot))$, we can minimize the well conditioned function $\theta_\varepsilon(\cdot)$, and let ε go to zero. A crucial question to this approach is that if any selection of the minimizers of $\theta_\varepsilon(\cdot)$ converges to the set of minimizers of $\lambda_1(A(\cdot))$ as ε goes to zero. To answer this question, we apply Auslender's general penalty approach [1] to our problem (1) of minimizing the maximum eigenvalue function over an affine space.

Recall that the recession function of the max-function (3) is (see Example (ii) of [1, Introduction])

$$h(x) := \ln \left(\sum_{i=1}^{n} e^{x_i} \right), \qquad \forall\, x \in \mathbb{R}^n.$$

Then the smoothing function f_ε defined by (4) can be obtained by

$$f_\varepsilon(x) = \varepsilon h(x/\varepsilon).$$

For simplicity, let $g : \mathbb{R}^m \to \mathbb{R}^n$

$$g(y) := (\lambda \circ A)(y) = \lambda(A(y)),$$

and the components of g be given respectively by

$$g_i(y) := \lambda_i(A(y)) = (\lambda_i \circ A)(y), \quad i = 1, \ldots, n.$$

We then have

$$\theta_\varepsilon(y) = \varepsilon h(g(y)/\varepsilon).$$

For each $i = 1, \ldots, n$ and $y \in \mathbb{R}^m$, the recession function of g_i, denoted by $(g_i)_\infty$, is defined by (see, [1, (2.1)]

$$(g_i)_\infty(y) := \inf \left\{ \liminf_{k\to\infty} \frac{g_i(t_k y_k)}{t_k} : \ t_k \to +\infty, \ y_k \to y \right\}.$$

Let

$$g_\infty(y) := ((g_1)_\infty(y), \ldots, (g_n)_\infty(y)), \qquad \forall\, y \in \mathbb{R}^m.$$

A general convergence result of Auslender [1, Thm. 2.3], which in our case concerns the convergence of any selection of the minimizers of $\theta_\varepsilon(\cdot)$ to the solution set of the minimizers of $\lambda_1(A(\cdot))$, involves the condition

$$h(g_\infty(y)) > 0, \qquad \forall\, 0 \neq y \in \mathbb{R}^m. \tag{29}$$

The other assumptions of [1, Thm. 2.3] are automatically satisfied in our case.

Lemma 4.1 *Suppose that the solution set of (1), denoted by Ω_0, is nonempty and bounded. Then (29) is satisfied for all $0 \neq y \in \mathbb{R}^m$.*

Proof. Let $0 \neq y \in \mathbb{R}^m$ be given, and let $t_k \to +\infty$ and $y_k \to y$. We have

$$(\lambda_i \circ A)(t_k y_k)/t_k = \lambda_i(A(t_k y_k))/t_k = \lambda_i(A_0 + t_k \mathcal{A} y_k)/t_k = \lambda_i(A_0/t_k + \mathcal{A} y_k).$$

Hence

$$(g_i)_\infty(y) = \lim_{k\to\infty} \lambda_i(A_0/t_k + \mathcal{A} y_k) = \lambda_i(\mathcal{A} y).$$

Therefore, we have

$$h(g_\infty(y)) = \lambda_1(\mathcal{A} y).$$

Then Lemma 3.5 implies (29) for all $0 \neq y \in \mathbb{R}^m$ since Ω_0 is nonempty and bounded. □

Now we are ready to specialize to our problem a general convergence result for penalty and barrier methods due to Auslender [1, Thm. 2.3] (all assumptions except (2.7) of this theorem are automatically satisfied with our problem, and (2.7) is verified in Lemma 4.1).

Theorem 4.2 *Suppose that the solution set Ω_0 of problem (1) is nonempty and compact. Then the solution set Ω_ε of problem (6) is also nonempty and compact for any $\varepsilon > 0$. Moreover, every selection $y_\varepsilon \in \Omega_\varepsilon$ stays bounded with all its limit points in Ω_0. In particular, if the matrices $A_1, \ldots, A_m$ are linearly independent, then for every $\varepsilon > 0$ the minimization problem (6) is strictly convex, admits a unique solution y_ε, and*

$$\lim_{\varepsilon\to 0} \operatorname{dist}(y_\varepsilon, \Omega_0) = 0$$

where dist *denotes the distance of y_ε to Ω_0.*

The boundedness condition on the solution set cannot be dropped in the theorem. Here is an example: Let $n = 2,\ m = 1,\ A_0 = \mathrm{Diag}[0, 0]$ and $A_1 = \mathrm{Diag}[0, 1]$. Then the solution set of the eigenvalue optimization problem is $\{x \in \mathbb{R} |\ x \leq 0\}$, which is obviously unbounded. However, for any $\varepsilon > 0$,

$$\theta_\varepsilon(x) = \varepsilon \ln(1 + e^{x/\varepsilon}),$$

and hence the convex minimization problem (6) has no solution. In this example, the linear independence condition holds. On the one hand, the linear independence increases the solvability of the smooth problem (6) as it is strictly convex. However, sometimes it is very difficult to check the linear independence. On the other hand, the linear independence sometimes is not sufficient to guarantee the solvability of the smooth problem. To overcome these difficulties, we consider the so-called Tikhonov regularization of the problem (6):

$$\min_{y \in \mathbb{R}^m} \ \hat{\theta}_\varepsilon(y), \tag{30}$$

where

$$\hat{\theta}_\varepsilon(y) := \theta_\varepsilon(y) + \frac{1}{2}\|y\|^2, \qquad \forall\ y \in \mathbb{R}^m.$$

Since $\theta_\varepsilon(\cdot)$ is convex for any $\varepsilon > 0$, $\hat{\theta}_\varepsilon(\cdot)$ is strongly convex, and hence the problem (30) always has a (unique) solution and is well-posed in the sense of Dontchev and Zolezzi [6]. The Tikhonov regularization has been extensively studied in the book [6] and been successfully used to study complementarity problems and variational inequalities by a number of authors recently [7, 8, 22, 23, 25, 28]. The following result concerning the convergence of the solution sequence of the regularized problem (30) can be proved similarly as that for [1, Thm. 2.3] by using Lemma 4.1.

Theorem 4.3 *The smooth problem (30) has always a unique solution y_ε for any $\varepsilon > 0$, and the solution sequence $\{y_\varepsilon\}_{\varepsilon>0}$ remains bounded and the distance between y_ε and Ω_0 approaches zero as $\varepsilon \to 0$ providing Ω_0 is nonempty and bounded.*

Our regularization method based on Theorem 4.3 can be stated as follows.

Algorithm 4.4 *(A regularization method)*

(S.1) Let $\{\varepsilon_1, \varepsilon_2, \ldots\}$ be a given sequence decreasing to zero. Let $y^0 \in \mathbb{R}^m$ be given. Set $k := 1$.

(S.2) Using an unconstrained convex minimization method with initial point y^{k-1} to find the unique solution of the problem (30) with $\varepsilon = \varepsilon_k$.

(S.3) Repeat (S.2) until a termination criterion is satisfied.

Remark. Although the problem (30) is well-posed for each $\varepsilon_k > 0$, we stress that sometimes it is not an easy task to find the exact solution of (30). Fortunately, it is possible to design an iterative algorithm based on Algorithm 4.4 by exploiting the regularization term of (30) as done in [23, 28] for box variational inequalities and complementarity problems. It is beyond the scope of the current paper to design such an algorithm with convergence analysis. We will leave it with future research.

5 A combinatorial application

In this section, we will illustrate the theory developed so far with a combinatorial application. After formulating the combinatorial problem as the standard eigenvalue optimization problem (1), we emphasize two aspects of the application: Firstly, we verify that the two regularity conditions in Proposition 3.6 are satisfied so that the smooth convex problem (6) is strictly convex. Secondly, we obtain a more explicit formula for calculating the Hessian matrix. This strengthens the theoretical study of Lewis and Sendov [16] in a sense that, although the general Hessian formula for twice differentiable spectral functions is a bit complicated, it has much simpler counterparts in some real applications.

Our application starts with the eigenvalue optimization reformulation of the quadratic programming in $\{-1, 1\}$ variables

$$\max\ x^T C x \quad \text{s.t.} x \in \{-1, 1\}^n. \tag{31}$$

In the case that C is the Laplace matrix of a (possibly weighted) graph, the problem is known to be equivalent to the max-cut problem. Note that

$$\mathrm{tr}(x^T C x) = \mathrm{tr}(C x x^T) := \langle C, X \rangle,$$

and for all $\{-1, 1\}^n$ vectors, xx^T is positive semidefinite matrix with all diagonal elements equal to 1, which also implies $\mathrm{tr}(xx^T) = n$, so the semidefinite relaxation of (31) is as follows (with the redundant constraint $\mathrm{tr}(X) = n$)

$$\begin{array}{ll} \max & \langle C, X \rangle \\ \text{s.t.} & \mathrm{diag}(X) = e \\ & \mathrm{tr}(X) = n \\ & X \succeq 0. \end{array} \tag{32}$$

Here e is the vector of all ones in $\mathbb{R}^n$ (noticing that e in previous sections denotes the exponential function always with a power, so there should be no ambiguity). By considering its dual problem, (32) is equivalent to the following eigenvalue optimization problem

$$\min_{y \in \mathbb{R}^n}\ n\lambda_1(C - \mathrm{Diag}[y]) + e^T y.$$

See [11, Sects. 2, 7] for detailed arguments. This eigenvalue optimization problem can be transformed into the standard form

$$\min_{y \in \mathbb{R}^n}\ \lambda_1(A_0 + \sum_{i=1}^n A_i y_i) \tag{33}$$

with $A_0 = nC$, $A_i = -nE_i + I$, $i = 1, \ldots, n$ and $E_i \in \mathcal{S}_n$ is the zero matrix with the exception of the ith diagonal element equal to 1. It is easy to see

$$\sum_{i=1}^n A_i = 0,$$

which means linear dependence of the matrices $\{A_1, \ldots, A_n\}$. However, the matrices $\{A_1, \ldots, A_{n-1}\}$ are linearly independent. To see this, let $h \in \mathbb{R}^{n-1}$ satisfy

$$\sum_{i=1}^{n-1} A_i h_i = 0.$$

Simple calculation yields that $h_1 = h_2 = \ldots = h_{n-1} = 0$. So problem (33) is equivalent to the following problem in $\mathbb{R}^{n-1}$:

$$\min_{y \in \mathbb{R}^{n-1}} \ \lambda_1(A_0 + \sum_{i=1}^{n-1} A_i y_i) \tag{34}$$

Now we verify that I does not belong to the range of $\mathcal{A}$, i.e.,

$$I \notin Range(\mathcal{A}) := \{\sum_{i=1}^{n-1} A_i h_i |\ h \in \mathbb{R}^{n-1}\}.$$

Suppose that there is $h \in \mathbb{R}^{n-1}$ satisfying

$$\mathcal{A}h = I,$$

which means

$$\begin{cases} -nh_j + \sum_{i=1}^{n-1} h_i &= 1, \quad j = 1, \ldots, n-1 \\ \sum_{i=1}^{n-1} h_i &= 1. \end{cases}$$

Hence, $h_1 = h_2 = \ldots = h_{n-1} = 0$, contradicting the last equality of the above system. Then it follows from Proposition 3.4 that the function θ_ε when applied to the problem (34) is strictly convex and its Hessian is positive definite everywhere. Moreover, the computational formulae of the gradient and the Hessian become much simple, partially due to the sparse structure of A_i's.

Recall from the problem (34) that

$$m = n - 1, \ A_0 = nC, \ A_i = -nE_i + I$$

where C is the matrix associated with the problem (31). Hence $A(y) = A_0 + \sum_{i=1}^{m} A_i y_i$. Given $y \in \mathbb{R}^{n-1}$, we first have a spectral decomposition

$$U^T(A(y))U = \lambda(A(y)). \tag{35}$$

As a matter of fact, U can be chosen satisfying

$$U^T C_y U = \lambda(C_y), \quad C_y := C - \mathrm{Diag}[(y^T, 0)]$$

and hence

$$\lambda(A(y)) = n\lambda(C_y) + \left(\sum_{i=1}^{n-1} y_i\right) e.$$

For each $i = 1, \ldots, n-1$, we have

$$\tilde{A}_i = U^T A_i U = -nU^T E_i U + I = -n(U(i,:))^T U(i,:) + I$$

where $U(i,:)$ denotes the ith row vector of U. Hence

$$\mathrm{diag}(\tilde{A}_i) = -n \begin{pmatrix} (U(i,1))^2 \\ \vdots \\ (U(i,n))^2 \end{pmatrix} + e.$$

The matrix W defined by (20) is given by

$$W = -n \left(U(1:m,:) \circ U(1:m,:)\right)^T + E_n^m,$$

where $U(1:m,:)$ is a matrix consisting of the first m columns of U, and E_n^m is the $n \times m$ matrix of all ones. Again let $z := \lambda(A(y))$ and $\mu := \mu(\varepsilon, z)$ (cf. (15)). Then it follows from (21) that

$$\begin{aligned} \nabla\theta_\varepsilon(y) &= -n \left(U(1:m,:) \circ U(1:m,:)\right) \mu + e \\ &= -nW_0^T(\mu - e/n) \end{aligned} \tag{36}$$

where we used

$$W_0 := \left(U(1:m,:) \circ U(1:m,:)\right)^T, \quad \sum_{i=1}^n \mu_i = 1 \ \text{ and } \ \sum_{j=1}^n (U(i,j))^2 = 1, \ \forall i = 1, \ldots, m.$$

We now use (27) to simplify the calculation of the Hessian of θ_ε. With the notion above, we further let

$$B := \nabla f_\varepsilon(\lambda(A(y))) = \frac{1}{\varepsilon}\left(\mathrm{Diag}[\mu] - \mu\mu^T\right).$$

It is easy to verify that

$$BE_n^m = \frac{1}{\varepsilon}\left(\mathrm{Diag}[\mu] - \mu\mu^T\right) E_n^m = 0,$$

which implies

$$\begin{aligned} W^T B W &= (-nW_0 + E_n^m)^T B(-nW_0 + E_n^m) \\ &= n^2 W_0^T B W_0 - 2nW_0^T B E_n^m + E_m^n B E_n^m \\ &= n^2 W_0^T B W_0. \end{aligned} \tag{37}$$

We now come forward to calculate M defined in (26). With $\mathcal{C} = \mathcal{C}(z)$ defined by (22) and $\tilde{E}_i = U^T E_i U$, we have

$$\tilde{A}_i = -n\tilde{E}_i + I \quad \text{and} \quad \mathcal{C} \circ \tilde{A}_i = -n(\mathcal{C} \circ \tilde{E}_i)$$

since the diagonal elements of $\mathcal{C}$ are zeros. Then, for $i, j = 1, \ldots, m$, we have

$$\begin{aligned} M_{ij} &= \langle \mathcal{C},\ \tilde{A}_i \circ \tilde{A}_j \rangle \\ &= \langle \mathcal{C} \circ \tilde{A}_i,\ \tilde{A}_j \rangle \\ &= -n\langle \mathcal{C} \circ \tilde{E}_i,\ -n\tilde{E}_j + I \rangle \\ &= n^2 \langle \mathcal{C} \circ \tilde{E}_i,\ \tilde{E}_j \rangle \end{aligned} \tag{38}$$

where the last equality used the fact that the diagonal elements of $(\mathcal{C} \circ \tilde{E}_i)$ are zeros. Hence it is easy to calculate the gradient and the Hessian of θ_ε according to (36)-(38).

6 Conclusions and future research

In this paper, we studied smooth convex approximations to the maximum eigenvalue function. To make it applicable to a wide class of applications, the study are conducted on the composite function of the maximum eigenvalue function and a linear operator mapping $\mathbb{R}^m$ to $\mathcal{S}_n$, which in turn is the natural objective function of minimizing the maximum eigenvalue function over an affine space in $\mathcal{S}_n$. This leads us to a sequence of smooth convex minimization problems governed by a smoothing parameter. As the parameter goes to *zero*, the original problem is recovered. Many efforts are then given on deriving a computable Hessian formula of the smooth convex functions, and on the regularity conditions which guarantees the nonsingularity of the Hessian matrices. We also proposed a regularization technique ensuring the well-posedness of the smooth convex problems. In our combinatorial applications, we obtained much more explicit formulae for calculating the gradient and the Hessian of the corresponding convex functions, and verified the regularity conditions.

Our study strengthens the pioneer research of Lewis and Sendov [16] on twice differentiable spectral functions in a sense that, although most of the known spectral functions are not twice differentiable (not even differentiable), there are indeed many smooth spectral functions arising from approximating nonsmooth spectral functions, as verified in this paper. More interestingly, the gradient and the Hessian are easy to obtain for these smooth approximations. This could lead to a class of numerical methods for minimizing the maximum eigenvalue function over an affine space. The regularization method proposed in this paper is only the beginning toward this purpose. It is possible to design an iterative method based on Algorithm 4.4 and conduct numerical tests. We leave this in future research. Finally we point out that the proof for the continuity of the gradient map in Proposition 2.3 (b) can be revised to give a different yet simple proof for the Lewis-Sendov Hessian formula [16, Thm. 3.3].

References

[1] A. Auslender, *Penalty and barrier methods: a unified framework*, SIAM J. Optim., 10 (1999), pp. 211–230.

[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, pp. 1–11.

[3] D.P. Bertsekas, *Constrained Optimization and Lagrangian Multiplier Methods*, Academic Press, New York 1982.

[4] P.-L. Chang, *A minimax approach to nonlinear programming*, Doctoral Dissertation, University of Washington, Department of Mathematics, 1980.

[5] Xin Chen and P. Tseng, *Non-interior continuation methods for solving semidefinite complementarity problems*, Report, Department of Mathematics, University of Washington, Seattle, October 1999.

[6] A.L Dontchev and T. Zolezzi, *Well-posed optimization problems*, Lecture Notes in Mathematics, 1543. Springer-Verlag, Berlin, 1993.

[7] F. Facchinei *Structural and stability properties of P_0 nonlinear complementarity problems*, Math. Oper. Res., 23 (1998), pp. 735–745.

[8] F. Facchinei and C. Kanzow, *Beyond monotonicity in regularization methods for nonlinear complementarity problems*, SIAM J. Control Optim., 37 (1999), pp. 1150–1161.

[9] A.A. Goldstein, *Chebyshev approximation and linear inequalities via exponentials*, Report, Department of Mathematics, University of Washington, Seattle, 1997.

[10] G.H. Golub and C.F. Val Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore 1996.

[11] C. Helmberg and F. Rendl, *A spectral bundle method for semidefinite programming*, SIAM J. Optim., 10 (2000), pp. 673–696.

[12] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge Press, Cambridge, United Kingdom, 1985.

[13] A.S. Lewis *Derivatives of spectral functions*, Math. Oper. Res., 21 (1996), pp. 576–588.

[14] A.S. Lewis and M.L. Overton, *Eigenvalue optimization*, Acta Numer., 5 (1996), pp. 149–190.

[15] A.S. Lewis, *Nonsmooth analysis of eigenvalues*, Math. Programming, 84 (1999), pp. 1-24.

[16] A.S Lewis and H.S. Sendov, *Twice differentiable spectral functions*, to appear in SIAM J. Matrix Anal. Appl..

[17] X.-S. Li, *An aggregation function method for nonlinear programming*, Science in China (Series A), 34 (1991), pp. 1467–1473.

[18] F. Oustry, *The $\mathcal{U}$-Lagrangian of the maximum eigenvalue function*, SIAM J. Optim., 9 (1999) pp. 526–549.

[19] F. Oustry, *A second-order bundle method to minimize the maximum eigenvalue function*, to appear in Math. Programming.

[20] M.L. Overton and R.S. Womersley, *Second derivatives for optimizing eigenvalues of symmetric matrices*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 667–718.

[21] J.-M. Peng and Z.-H. Lin, *A non-interior continuation method for generalized linear complementarity problems*, Math. Programming, 86 (1999), pp. 533–563.

[22] H.-D. Qi, *Tikhonov regularization methods for variational inequality problems*, J. Optim. Theory Appl., 102 (1999), pp. 193–201.

[23] H.-D. Qi, *A regularized smoothing Newton method for box constrained variational inequality problems with P_0-functions*, SIAM J. Optim. 10 (2000), pp. 315–330.

[24] H.-D. Qi and L.-Z. Liao, *A smoothing Newton method for extended vertical linear complementarity problems*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 45-66.

[25] G. Ravindran and M.S. Gowda, *Regularization of P_0-functions in box variational inequality problems*, to appear in SIAM J. Optim..

[26] A. Shapiro and M.K.H. Fan, *On eigenvalue optimization*, SIAM J. Optim., 5 (1995), pp. 552–568.

[27] G.W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[28] D. Sun, *A regularization Newton method for solving nonlinear complementarity problems*, Appl. Math. Optim., 40 (1999), pp. 315–339.

[29] H. Tang and L. Zhang, *A maximum entropy method for convex programming*, Chinese Sci. Bull., 39 (1994), pp. 682–684.

[30] P. Tseng and D.P. Bertsekas, *On the convergence of the exponential multiplier method for convex programming*, Math. Programming, 60 (1993), pp. 1–19.

[31] L. Zhang and H. Tang, *A further study on a penalty function of Bertsekas*. Advances in nonlinear programming (Beijing, 1996), 345–351, Appl. Optim., 14, Kluwer Acad. Publ., Dordrecht, 1998.