Smoothing technique and its applications in Semidefinite Optimization

Yu. Nesterov *

October 11, 2004

Abstract

In this paper we extend the smoothing technique [7], [9] onto the problems of Semidefinite Optimization. For that, we develop a simple framework for estimating a Lipschitz constant for the gradient of some symmetric functions of eigenvalues of symmetric matrices. Using this technique, we can justify the Lipshitz constants for some natural approximations of maximal eigenvalue and the spectral radius of symmetric matrices. We analyze the complexity of the problem-oriented gradient-type schemes onto the problems of minimizing the maximal eigenvalue or the spectral radius of the matrix, which depends linearly on the design variables. We show that in the first case the number of iterations of the method is bounded by $O(\frac{1}{\epsilon})$, where ϵ is the required *absolute* accuracy of the problem. In the second case, the number of iterations is bounded by $\frac{4}{\delta}\sqrt{(1+\delta)r\ln r}$, where δ is the required *relative* accuracy and r is the maximal rank of corresponding linear matrix inequality. Thus, the latter method is a fully polynomial approximation scheme.

Keywords: convex optimization, non-smooth optimization, complexity theory, black-box model, optimal methods, structural optimization, smoothing technique.

CORE DP #2004/73

*Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium; e-mail: nesterov@core.ucl.ac.be.

This paper presents research results of the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office.
The scientific responsibility rests with its author(s).

1 Introduction

Motivation. Recently it was shown [7] that a proper use of structure of nonsmooth convex optimization problems leads to very efficient gradient schemes, whose performance significantly better than the lower complexity bounds derived from the black box assumptions [5]. However, this observation leads to implementable algorithms only if we are able to form a computable smooth approximation for the objective function of our problem. In this case, applying to this approximation an optimal method for minimizing smooth convex functions (see [6], Section 2.2.1), we can easily obtain a good solution to our initial problem. In papers [7] and [8], a special smoothing technique was developed mainly for piece-wise linear functions. Later, in [9] it was shown that this technique on some problem classes allows to compute approximate solutions with a required *relative accuracy*.

In this paper we extend the results of [7], [9] onto the problems of Semidefinite Optimization. For that, we justify the computable smooth approximations for two most important nonsmooth functions of symmetric matrices, these are the *maximal eigenvalue* and the *spectral radius*.

Contents. In Section 2 we study the smooth approximations of symmetric functions of eigenvalues of symmetric matrices. The main question we are interested in is the Lipschitz continuity of the gradient of such functions. We develop a simple technique which allows to estimate the Lipschitz constant of the gradient for a special class of symmetric functions. The value of such a function is obtained as a sum of the values of the same univariate function as applied to all eigenvalues of the argument (see Theorem 1). The main assumption on the univariate function is that all coefficients of its Taylor series at zero, starting from the second one, are nonnegative. We show that the functions from this class deliver the required good approximations for the maximal eigenvalue and the spectral radius of symmetric matrix. In Section 3 we describe an optimal method for smooth convex optimization, using the notation and style of [7]. This method is our main tool for treating the applications considered in the next two sections. In Section 4 we show how to apply the smoothing technique for minimizing the maximal eigenvalue of a symmetric matrix dependent linearly on the design variables. We discuss the similarity between our approach and the *spectral bundle method*, which is, in accordance to the present state of art, one of the most powerful scheme for treating the problems of that type. We derive an upper bound for the number of iterations of our gradient-type scheme, which appears to be proportional to $\frac{1}{\epsilon}$, where ϵ is a required *absolute* accuracy of the approximate solution. In Section 5 we apply the smoothing technique for minimizing the spectral radius of a symmetric matrix dependent linearly on the design variables. The number of iterations of the proposed gradient method is bounded by $\frac{4}{\delta}\sqrt{(1+\delta)r\ln r}$, where δ is a required *relative* accuracy, and r is the maximal rank of the matrix arising in corresponding linear matrix inequality. This method is a *fully polynomial approximation scheme* since its complexity does not depend on a particular problem instance. Another advantage of this method is that it does not require to compute an eigenvalue decomposition at each iteration. We provide the paper with Appendix, which is devoted to the necessary and sufficient conditions for Lipschitz continuity of the gradient of convex function with respect to an arbitrary norm introduced in the space of variables.

Notation. In what follows we denote by $\mathcal{M}_n$ the space of real $n \times n$-matrices, and by $\mathcal{S}_n \subset \mathcal{M}_n$ the space of symmetric matrices. A particular matrix is always denoted by a capital letter. In the spaces R^n and $\mathcal{M}_n$ we use the standard inner products

$$\langle x, y \rangle = \sum_{i=1}^{n} x^{(i)} y^{(i)}, \; x, y \in R^n,$$

$$\langle X, Y \rangle_M = \sum_{i,j=1}^{n} X^{(i,j)} Y^{(i,j)}, \; X, Y \in \mathcal{M}_n,$$

For $X \in \mathcal{S}_n$, we denote by $\lambda(X) \in R^n$ the vector of its eigenvalues. We assume that the eigenvalues are ordered in a decreasing order:

$$\lambda^{(1)}(X) \geq \lambda^{(2)}(X) \geq \ldots \geq \lambda^{(n)}(X), \quad X \in \mathcal{S}_n.$$

Thus, $\lambda_{\max}(X) = \lambda^{(1)}(X)$. Notation $D(\lambda) \in \mathcal{S}_n$ is used for a diagonal matrix with vector $\lambda \in R^n$ on the main diagonal. Note that any $X \in \mathcal{S}_n$ admits an eigenvalue decomposition

$$X = U(X) D(\lambda(X)) U(X)^T$$

with $U(X) : U(X)U(X)^T = I_n$, where $I_n \in \mathcal{S}_n$ is the identity matrix.

Let us mention notations whose meanings are different for vectors and matrices. In R^n, we use a standard notation for l_p-norms:

$$\|x\|_{(p)} = \left[\sum_{i=1}^{n} |x^{(i)}|^p \right]^{1/p}, \quad x \in R^n,$$

where $p \geq 1$, and $\|x\|_\infty = \max\limits_{1 \leq i \leq n} |x^{(i)}|$. The corresponding norms in $\mathcal{S}_n$ are introduced by

$$\|X\|_{(p)} = \|\lambda(X)\|_{(p)}, \quad X \in \mathcal{S}_n,$$

Further, for vector $\lambda \in R^n$ we denote by $|\lambda| \in R^n$ a vector with entries $|\lambda^{(i)}|$, $i = 1, \ldots, n$. Notation $\lambda^k \in R^n$ is used for the vector with components $(\lambda^{(i)})^k$, $i = 1, \ldots, n$. However, for $X \in \mathcal{S}_n$ we define

$$|X| \stackrel{\text{def}}{=} U(X) D(|\lambda(X)|) U(X)^T,$$

and notation X^k is used for the standard matrix power.

2 Smooth symmetric functions of eigenvalues

For $k \geq 1$, consider the following function:

$$\pi_k(X) = \langle X^k, I_n \rangle_M = \sum_{i=1}^{n} (\lambda^{(i)}(X))^k, \quad X \in \mathcal{S}_n.$$

Let us derive an upper bound for its second derivative. Note that this bound is nontrivial only for $k \geq 2$.

The derivatives of this function along a direction $H \in \mathcal{S}_n$ are defined as follows:

$$\begin{aligned} \langle \nabla \pi_k(X), H \rangle_M &= k \langle X^{k-1}, H \rangle_M, \\ \langle \nabla^2 \pi_k(X) H, H \rangle_M &= k \sum_{p=0}^{k-2} \langle X^p H X^{k-2-p}, H \rangle_M. \end{aligned} \tag{2.1}$$

We need the following result.

Lemma 1 *For any $p, q \geq 0$, and X, H from $\mathcal{S}_n$ we have*

$$\langle X^p H X^q + X^q H X^p, H \rangle_M \leq 2 \langle |X|^{p+q}, H^2 \rangle_M \leq 2 \langle |\lambda(X)|^{p+q}, \lambda^2(H) \rangle. \tag{2.2}$$

Proof:
Indeed, denote $\lambda = \lambda(X)$, $D = D(\lambda)$, $U = U(X)$ and $\hat{H} = U^T H U$. Then

$$\begin{aligned} \langle X^p H X^q + X^q H X^p, H \rangle_M &= \langle U D^p U^T H U D^q U^T + U D^q U^T H U D^p U^T, H \rangle_M \\ &= \langle D^p \hat{H} D^q + D^q \hat{H} D^p, \hat{H} \rangle_M \\ &= \sum_{i,j=1}^n (\hat{H}^{(i,j)})^2 \left((\lambda^{(i)})^p (\lambda^{(j)})^q + (\lambda^{(i)})^q (\lambda^{(j)})^p \right) \\ &\leq \sum_{i,j=1}^n (\hat{H}^{(i,j)})^2 \left(|\lambda^{(i)}|^p |\lambda^{(j)}|^q + |\lambda^{(i)}|^q |\lambda^{(j)}|^p \right). \end{aligned}$$

Note that for arbitrary non-negative values a and b we always have

$$0 \leq (a^p - b^p)(a^q - b^q) = (a^{p+q} + b^{p+q}) - (a^p b^q + a^q b^p).$$

Thus, we can continue as follows:

$$\begin{aligned} \langle X^p H X^q + X^q H X^p, H \rangle_M &\leq \sum_{i,j=1}^n (\hat{H}^{(i,j)})^2 \left(|\lambda^{(i)}|^{p+q} + |\lambda^{(j)}|^{p+q} \right) \\ &= 2 \sum_{i,j=1}^n (\hat{H}^{(i,j)})^2 |\lambda^{(i)}|^{p+q} = 2 \langle D(|\lambda|)^{p+q} \hat{H}, \hat{H} \rangle_M \\ &= 2 \langle D^{p+q}(|\lambda|), \hat{H}^2 \rangle_M = 2 \langle |X|^{p+q}, H^2 \rangle_M. \end{aligned}$$

Thus, we get the first inequality in (2.2). The second inequality is standard. □

Corollary 1 *For any $k \geq 2$ we have*

$$\langle \nabla^2 \pi_k(X) H, H \rangle_M \leq k(k-1) \langle |\lambda(X)|^{k-2}, \lambda^2(H) \rangle. \tag{2.3}$$

Proof:
For $k = 2$ the bound is trivial. For $k \geq 3$, in representation (2.1) we can unify the terms in the expression $\sum\limits_{p=0}^{k-2} \langle X^p H X^{k-2-p}, H \rangle_M$ in symmetric pairs

$$\langle X^p H X^{k-2-p} + X^{k-2-p} H X^p, H \rangle_M.$$

Applying to each pair inequality (2.2), we get the estimate (2.3). □

Let $f(\tau)$ be a function of real variable τ, defined by a power series

$$f(\tau) = a_0 + \sum_{k=1}^{\infty} a_k \tau^k$$

with $a_k \geq 0$ for $k \geq 2$. We assume that its domain $\mathrm{dom}\, f = \{\tau : \ |\tau| < R\}$ is nonempty. For $X \in \mathcal{S}_n$ consider the following symmetric function of eigenvalues:

$$F(X) = \sum_{i=1}^{n} f(\lambda^{(i)}(X)).$$

Clearly, $\mathrm{dom}\, F = \{X \in \mathcal{S}_n : \ \lambda^{(1)}(X) < R, \ \lambda^{(n)}(X) > -R\}$.

Theorem 1 *For any $X \in \mathrm{dom}\, F$ and $H \in \mathcal{S}_n$ we have*

$$\langle \nabla^2 F(X) H, H \rangle \leq \sum_{i=1}^{n} f''(|\lambda^{(i)}(X)|)(\lambda^{(i)}(H))^2.$$

Proof:
Indeed,

$$\begin{aligned} F(X) &= n \cdot a_0 + \sum_{i=1}^{n} \sum_{k=1}^{\infty} a_k (\lambda^{(i)}(X))^k \\ &= n \cdot a_0 + \sum_{k=1}^{\infty} a_k \sum_{i=1}^{n} (\lambda^{(i)}(X))^k \ = \ n \cdot a_0 + \sum_{k=1}^{\infty} a_k \pi_k(X). \end{aligned}$$

Thus, in view of inequality (2.3),

$$\begin{aligned} \langle \nabla^2 F(X) H, H \rangle_M &= \sum_{k=2}^{\infty} a_k \langle \nabla^2 \pi_k(X) H, H \rangle_M \\ &\leq \sum_{k=2}^{\infty} k(k-1) a_k \langle |\lambda(X)|^{k-2}, \lambda^2(H) \rangle \\ &= \sum_{i=1}^{n} \sum_{k=2}^{\infty} k(k-1) a_k |\lambda^{(i)}(X)|^{k-2} (\lambda^{(i)}(H))^2 \\ &= \sum_{i=1}^{n} f''(|\lambda^{(i)}(X)|)(\lambda^{(i)}(H))^2. \qquad \square \end{aligned}$$

Let us consider now two important examples of symmetric functions of eigenvalues.

1. Squared l_p matrix norm. For integer $p \geq 1$ consider the following function:

$$F_p(X) = \tfrac{1}{2}\|\lambda(X)\|^2_{(2p)} = \tfrac{1}{2}\langle X^{2p}, I_n\rangle_M^{1/p}, \quad X \in \mathcal{S}_n. \tag{2.4}$$

Thus, $F_p(X) = \frac{1}{2}(\pi_{2p}(X))^{1/p}$. Therefore, in view of (2.3), for any $X, H \in \mathcal{S}_n$ we have

$$\begin{aligned}
\langle \nabla F_p(X), H\rangle_M &= \tfrac{1}{2p}(\pi_{2p}(X))^{\frac{1}{p}-1}\langle \nabla \pi_{2p}(X), H\rangle_M, \\
\langle \nabla^2 F_p(X)H, H\rangle_M &= \tfrac{1}{2p}\cdot\left(\tfrac{1}{p}-1\right)\cdot(\pi_{2p}(X))^{\frac{1}{p}-2}\langle \nabla \pi_{2p}(X), H\rangle_M^2 \\
&\quad + \tfrac{1}{2p}(\pi_{2p}(X))^{\frac{1}{p}-1}\langle \nabla^2 \pi_{2p}(X)H, H\rangle_M \qquad (2.5)\\
&\leq (2p-1)(\pi_{2p}(X))^{\frac{1}{p}-1}\langle |\lambda(X)|^{2p-2}, \lambda^2(H)\rangle.
\end{aligned}$$

Let us apply Hölder inequality $\langle x, y\rangle \leq \|x\|_{(\beta)}\|y\|_{(\gamma)}$ with $\beta = \frac{p}{p-1}$, $\gamma = \frac{\beta}{\beta-1} = p$, and

$$x^{(i)} = |\lambda^{(i)}(X)|^{2p-2}, \quad y^{(i)} = (\lambda^{(i)}(H))^2, \quad i = 1, \ldots, n.$$

Then, we can continue:

$$\langle \nabla^2 F_p(X)H, H\rangle_M \leq (2p-1)\|\lambda(H)\|^2_{(2p)} = (2p-1)\|H\|^2_{(2p)}. \tag{2.6}$$

2. Entropy smoothing of maximal eigenvalue. Consider the function

$$E(X) = \ln \sum_{i=1}^n e^{\lambda^{(i)}(X)} \stackrel{\text{def}}{=} \ln F(X), \quad X \in \mathcal{S}_n. \tag{2.7}$$

Note that

$$\begin{aligned}
\langle \nabla E(X), H\rangle_M &= \tfrac{1}{F(X)}\langle \nabla F(X), H\rangle_M, \\
\langle \nabla^2 E(X)H, H\rangle_M &= -\tfrac{1}{F^2(X)}\langle \nabla F(X), H\rangle_M^2 + \tfrac{1}{F(X)}\langle \nabla^2 F(X)H, H\rangle_M \\
&\leq \tfrac{1}{F(X)}\langle \nabla^2 F(X)H, H\rangle_M.
\end{aligned}$$

Let us assume first that $X \succeq 0$. Function $F(X)$ is formed by auxiliary function $f(\tau) = e^\tau$, which satisfies assumptions of Theorem 1. Therefore

$$\langle \nabla^2 E(X)H, H\rangle_M \leq \left[\sum_{i=1}^n e^{\lambda^{(i)}(X)}\right]^{-1} \sum_{i=1}^n e^{\lambda^{(i)}(X)}(\lambda^{(i)}(H))^2 \leq \|H\|^2_\infty. \tag{2.8}$$

It remains to note that $E(X + \tau I_n) = E(X) + \tau$. Hence, the Hessian $\nabla^2 E(X + \tau I_n)$ does not depend on τ, and we conclude that the estimate (2.8) is valid for arbitrary $X \in \mathcal{S}_n$.

3 Optimal method for smooth convex optimization

In the next sections we discuss two possible applications of the results presented in Section 2. In both of them we use an optimal method for minimizing a convex function with Lipschitz continuous gradient (see, for example, [6, 7]). For the sake of completeness, we provide the reader with a short description of this scheme.

Let function $f(x)$ be differentiable and convex on a closed convex set $Q \subseteq E$, where E is a finite dimensional real vector space. Let us fix some norm $\|\cdot\|$ on E. In this section we consider an efficient optimization scheme for solving the following problem:

$$\min_x \{f(x) : \ x \in Q\}, \tag{3.1}$$

where f satisfies on Q the Lipschitz condition for its gradient:

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|, \quad \forall x, y \in Q.$$

Recall that the standard gradient projection method at this problem converges as $O(\frac{1}{k})$, where k is the iteration counter (see, e.g. [6], Section 2.1.5).

Let us assume that the constant $L > 0$ is known. Then we can use in our methods the *gradient mapping* $T_Q(x) \in Q$, which is defined as an optimal solution to the following minimization problem:

$$\min_y \left\{ \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} L\|y - x\|^2 : \ y \in Q \right\}. \tag{3.2}$$

If the norm $\|\cdot\|$ is not strictly convex, the problem (3.2) can have multiple solutions. In this case we stick the notation $T_Q(x)$ to any of them.

Denote by $d(x)$ a *prox-function* of the set Q. This means that $d(x)$ is continuous and strongly convex on Q with respect to the norm $\|\cdot\|$ with convexity parameter $\sigma > 0$, and that $d(x_0) = 0$, where x_0 is the *prox-center* of the set Q:

$$x_0 = \arg\min_x \{d(x) : \ x \in Q\}.$$

In our scheme we update recursively three sequences of points $\{x_k\}_{k=0}^{\infty}$, $\{y_k\}_{k=0}^{\infty}$, and $\{z_k\}_{k=0}^{\infty}$ from Q.

For $k \geq 0$ **do**

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

2. Find $y_k = T_Q(x_k)$.

3. Find $z_k = \arg\min_x \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^{k} \frac{i+1}{2} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] : \ x \in Q \right\}$. (3.3)

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

Note that at each iteration of this scheme we need to solve two auxiliary problems: the computation of the gradient mapping (Step 2), and the problem

$$\min_x \{\tfrac{L}{\sigma} d(x) + \langle u_k, x \rangle\},$$

with $u_k = \frac{2}{(k+1)(k+2)} \sum_{i=0}^{k} (i+1) \nabla f(x_i)$ (Step 3). We assume that the exact solutions for both problems are available.

Theorem 2 (see [7]). *Let sequences $\{x_k\}_{k=0}^{\infty}$ and $\{y_k\}_{k=0}^{\infty}$ be generated by method (3.3). Then for any $k \geq 0$ we have*

$$\frac{(k+1)(k+2)}{4} f(y_k) \leq \min_x \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^{k} \frac{i+1}{2} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] : \ x \in Q \right\}. \quad (3.4)$$

Therefore,

$$f(y_k) - f(x^*) \leq \frac{4Ld(x^*)}{\sigma(k+1)(k+2)}, \quad (3.5)$$

where x^* is an optimal solution to the problem (3.1).

In the next sections we will use the above scheme for minimizing smooth approximations of nonsmooth convex functions. The smoothing will be done by the use of spectral functions considered in Section 2.

4 Minimizing the maximal eigenvalue of symmetric matrix

Consider the following problem:

$$\text{Find } \phi^* = \min_y \{\phi(y) \stackrel{\text{def}}{=} \lambda_{\max}(C + A(y)) : \ y \in Q\}, \quad (4.1)$$

where Q is a closed convex set in R^m and $A(\cdot)$ is a linear operator from R^m to $\mathcal{S}_n$:

$$A(y) = \sum_{i=1}^{m} y^{(i)} A_i \in \mathcal{S}_n, \quad y \in R^m.$$

Note that the objective function in (4.1) is nonsmooth. Therefore this problem can be solved either by interior-point methods, or by general methods of nonsmooth convex optimization (see, e.g., [6], Chapter 3). However, due to a very special structure of the objective function, for problem (4.1) there were proposed so-called *spectral bundle methods* (see [1, 2, 3, 4, 10]).

The idea of spectral bundle methods is very simple. Indeed, the subdifferential of the function $\lambda_{\max}(X)$, $X \in \mathcal{S}_n$, has the following structure:

$$\partial \lambda_{\max}(X) = \left\{ G = \sum_{i=1}^{d(X)} \tau_i u_i(X) u_i(X)^T, \ \tau_i \geq 0, \ i = 1, \ldots, d(X), \ \sum_{i=1}^{d(X)} \tau_i = 1 \right\},$$

where $u_i(X)$ are the columns of the matrix $U(X)$ and $d(X)$ is defined as multiplicity of the maximal eigenvalue of matrix X. Note that the general optimization methods require computation of a single subgradient at each test point. Consequently, at each point these methods can employ only a single linear inequality of the type

$$\lambda_{\max}(Y) \geq \lambda_{\max}(X) + \langle G, Y - X \rangle_M = \langle G, Y \rangle_M, \quad Y \in \mathcal{S}_n,$$

with certain $G \in \partial \lambda_{\max}(X)$. However, in our situation we can construct a better lower bound for the objective function. Indeed, let us fix an arbitrary r, $1 \leq r \leq n$. Then

$$\lambda_{\max}(Y) = \max_{\|x\|_{(2)}=1} \langle Yx, x \rangle \geq \max_{1 \leq i \leq r} \langle Y u_i(X), u_i(X) \rangle, \quad Y \in \mathcal{S}_n.$$

Of course, the best model corresponds to $r = n$. But in this case the amount of accumulated information and the complexity of generating new test points grow too quickly. Therefore, different versions of spectral bundle methods apply different strategies for defining a reasonably small parameter r. Unfortunately, in accordance to our knowledge, up to now there is no complexity analysis done for these schemes. So, we can compare them only by computational results.

We are going to solve the problem (4.1) by a smoothing technique in a manner similar to [7]. This means that we replace the function $\lambda_{\max}(X)$ by its smooth approximation $f_\mu(X) = \mu E(\frac{1}{\mu}X)$, defined by (2.7) and a tolerance parameter $\mu > 0$. Note that

$$\begin{aligned} f_\mu(X) &= \mu \ln \left[\sum_{i=1}^n e^{\lambda^{(i)}(X)/\mu} \right] \geq \lambda_{\max}(X), \\ f_\mu(X) &\leq \lambda_{\max}(X) + \mu \ln n. \end{aligned} \tag{4.2}$$

At the same time,

$$\nabla f_\mu(X) = \left[\sum_{i=1}^n e^{\lambda^{(i)}(X)/\mu} \right]^{-1} \cdot \sum_{i=1}^n e^{\lambda^{(i)}(X)/\mu} u_i(X) u_i(X)^T. \tag{4.3}$$

Similarly to the spectral bundle methods, at each test point X the gradient $\nabla f_\mu(X)$ takes into account different eigenvectors of the matrix X. Since the factors $e^{\lambda^{(i)}(X)/\mu}$ decrease very rapidly, this gradient actually depends only on few largest eigenvalues. However, their selection is made automatically by expression (4.3). In some sense, the ranking of importance of the eigenvalues is done in a logarithmic scale controlled by the tolerance parameter μ.

Let us analyze now the efficiency of smoothing technique as applied to problem (4.1). Our goal is to find an ϵ-solution $\bar{x} \in Q$ to the problem (4.1):

$$\phi(\bar{y}) - \phi^* \leq \epsilon. \tag{4.4}$$

For that we will try to find an $\frac{1}{2}\epsilon$-solution to the smooth problem

$$\text{Find } \phi_\mu^* = \min_y \{ \phi_\mu(y) \stackrel{\text{def}}{=} f_\mu(C + A(y)) : \ y \in Q \}, \tag{4.5}$$

with $\mu = \mu(\epsilon)$, defined as

$$\mu(\epsilon) = \tfrac{\epsilon}{2 \ln n}. \tag{4.6}$$

Clearly, if $\phi_\mu(\bar{y}) - \phi_\mu^* \leq \frac{1}{2}\epsilon$, then in view of (4.2) we have

$$\phi(\bar{y}) - \phi^* \leq \phi_\mu(\bar{y}) - \phi_\mu^* + \mu \ln n \leq \epsilon.$$

Let us analyze now the complexity of finding $\frac{1}{2}\epsilon$-solution to problem (4.5) by the optimal method (3.3).

Let us fix some norm $\|h\|$ for $h \in R^m$. Consider a prox-function $d(x)$ of the set Q with the prox-center $x_0 \in Q$. We assume this function to be strongly convex on Q with convexity parameter $\sigma > 0$. Define

$$\|A\| = \max_{h \in R^m} \{ \|A(h)\|_\infty : \ \|h\| = 1 \}.$$

Let us estimate the second derivative of function $\phi_\mu(y)$. For any y and h from R^m, in view of inequality (2.8) we have

$$\begin{aligned}\langle \nabla \phi_\mu(y), h\rangle &= \langle \nabla f_\mu(C + A(y)), h\rangle = \langle \nabla E(\tfrac{1}{\mu}(C + A(y))), A(h)\rangle_M,\\ \langle \nabla^2 \phi_\mu(y)h, h\rangle &= \tfrac{1}{\mu}\langle \nabla^2 E(C + A(y))A(h), A(h)\rangle_M\\ &\le \tfrac{1}{\mu}\|A(h)\|_\infty^2 \le \tfrac{1}{\mu}\|A\|^2 \cdot \|h\|^2.\end{aligned}$$

Thus, by Theorem 4 function $\phi_\mu(y)$ has a Lipschitz continuous gradient with the constant

$$L = \tfrac{1}{\mu}\|A\|^2 = \tfrac{2\ln n}{\epsilon}\|A\|^2.$$

Taking now into account the estimate (3.5), we conclude that the method (3.3), as applied to the problem (4.5), has the following rate of convergence:

$$\phi_\mu(y_k) - \phi_\mu^* \le \tfrac{8\ln n\|A\|^2 d(y_\mu^*)}{\epsilon\cdot\sigma(k+1)(k+2)},$$

where $y_\mu^* \in Q$ is the solution to (4.5). Hence, it is able to generate an $\frac{1}{2}\epsilon$-solution to this problem (which is an ϵ-solution to problem (4.1)) at most after

$$\tfrac{4\|A\|}{\epsilon}\sqrt{\tfrac{\ln n}{\sigma}\, d(y_\mu^*)} \tag{4.7}$$

iterations.

5 Minimizing the spectral radius of symmetric matrix

For matrix $X \in \mathcal{S}_n$, define its spectral radius:

$$\rho(X) = \max_{1\le i\le n} |\lambda^{(i)}(X)| = \max\{\lambda^{(1)}(X), -\lambda^{(n)}(X)\}.$$

Clearly, $\rho(X)$ is a convex function on $\mathcal{S}_n$. In this section we consider the following optimization problem:

$$\text{Find } \phi_* = \min_{y\in R^m}\{\phi(y) \stackrel{\text{def}}{=} \rho(A(y)) :\ y \in Q\}, \tag{5.1}$$

where $Q \subset R^m$ is a closed convex set separated from the origin, and $A(\cdot)$ is a linear operator from R^m to $\mathcal{S}_n$:

$$A(y) = \sum_{i=1}^m y^{(i)} A_i \in \mathcal{S}_n, \quad y \in R^m.$$

We assume that the matrices $\{A_i\}_{i=1}^m$ are linearly independent. Hence, the matrix $G \in \mathcal{S}_m$ with elements

$$G^{(i,j)} = \langle A_i, A_j\rangle_M, \quad i, j = 1, \ldots, m,$$

is positive definite. Denote by r the maximal rank of $A(y)$:

$$r = \max_{y \in R^m} \operatorname{rank} A(y) \leq \min \left\{ n, \sum_{i=1}^{m} \operatorname{rank} A_i \right\}.$$

We are going to solve (5.1) using a variant of smoothing technique, suggested in [9] for solving structural convex optimization problems in *relative scale*. Note that in view of our assumptions ϕ^* is strictly positive.

First of all, we approximate a non-smooth objective function in (5.1) by a smooth one. For that, we use $F_p(X)$ defined by (2.4). Note that

$$\begin{array}{rcl} F_p(X) & = & \frac{1}{2} \langle X^{2p}, I_n \rangle_M^{1/p} \geq \frac{1}{2} \rho^2(X), \\ F_p(X) & \leq & \frac{1}{2} \rho^2(X) \cdot (\operatorname{rank} X)^{1/p}. \end{array} \tag{5.2}$$

Consider the problem

$$\text{Find } f_p^* = \min_{y \in R^m} \{ f_p(y) \stackrel{\text{def}}{=} F_p(A(y)) : \ y \in Q \}, \tag{5.3}$$

From (5.2) we can see that

$$\frac{1}{2} \phi_*^2 \leq f_p^* \leq \frac{1}{2} \phi_*^2 \cdot r^{1/p}. \tag{5.4}$$

Our goal is to find a point $\bar{y} \in Q$, which solves (5.1) with relative accuracy $\delta > 0$:

$$\phi(\bar{y}) \leq (1 + \delta) \phi_*.$$

Let us choose an integer p, which satisfies the following inequality

$$p(\delta) \stackrel{\text{def}}{=} \frac{1+\delta}{\delta} \ln r \leq p \leq 2p(\delta). \tag{5.5}$$

Assume that $\bar{y} \in Q$ solves (5.3) with relative accuracy δ. Then, in view of (5.2) and (5.4), we have

$$\phi(\bar{y}) / \phi_* \ \leq \ r^{\frac{1}{2p}} \cdot \sqrt{f_p(\bar{y}) / f_p^*} \ \leq \ r^{\frac{1}{2p}} \cdot \sqrt{1 + \delta} \ \leq \ e^{\frac{\delta}{2(1+\delta)}} \cdot \sqrt{1 + \delta} \ \leq \ 1 + \delta.$$

Thus, we need to estimate the efficiency of the method (3.3) as applied to the problem (5.3). Let us introduce the norm

$$\|h\|_G = \langle Gh, h \rangle^{1/2}, \quad h \in R^m.$$

Assuming that $p(\delta) \geq 1$ and using the estimate (2.6), for any y and h from R^m we obtain

$$\begin{array}{rcl} \langle \nabla^2 f_p(y) h, h \rangle & = & \langle \nabla^2 F_p(A(y)) A(h), A(h) \rangle_M \\ & \leq & (2p-1) \|A(h)\|_{2p}^2 \ \leq \ (2p-1) \|A(h)\|_{(2)}^2 \\ & = & (2p-1) \langle A(h), A(h) \rangle_M \ = \ (2p-1) \langle Gh, h \rangle \\ & = & (2p-1) \|h\|_G^2. \end{array}$$

Thus, in view of Theorem 4, function $f_p(y)$ has Lipschitz continuous gradient on R^m with respect to the nowm $\|\cdot\|_G$ with Lipschitz constant

$$L = 2p - 1 \leq 4p(\delta). \tag{5.6}$$

On the other hand, for any $X \in \mathcal{S}_n$ with $\operatorname{rank} X \leq r$, and $p \geq 1$ we have

$$\tfrac{1}{r}\|X\|_{(2)}^2 \leq \|X\|_\infty^2 \leq \|X\|_{(2p)}^2.$$

Hence, $\frac{1}{2r}\|y\|_G^2 \leq f_p(y)$ for any $y \in R^m$. In particular,

$$\tfrac{1}{2r}\|y_p^*\|_G^2 \leq f_p^*, \tag{5.7}$$

where y_p^* is a solution to (5.3).

Denote $x_0 = \arg\min_y\{\|y\|_G : \ y \in Q\}$. Since the norm $\|\cdot\|_G$ is Euclidean, we have

$$\|y_p^* - x_0\|_G^2 \leq \|y_p^*\|_G^2 - \|x_0\|_G^2 < \|y_p^*\|_G^2.$$

Combining this inequality with estimate (5.7), we get

$$\tfrac{1}{2}\|y_p^* - x_0\|_G^2 \leq \tfrac{1}{2}\|y_p^*\|_G^2 \leq r f_p^*. \tag{5.8}$$

In order to apply to the problem (5.3) method (3.3), let us choose the following prox-function:

$$d(x) = \tfrac{1}{2}\|x - x_0\|_G^2. \tag{5.9}$$

Since, the convexity parameter σ of this function is equal to one, in view of the bounds (5.6) and (5.8), the method (3.3) launched from the starting point x_0 converges as follows:

$$f_p(y_k) - f_p^* \quad \leq \quad \tfrac{16(1+\delta)r\ln r}{\delta\cdot(k+1)(k+2)} \cdot f_p^*. \tag{5.10}$$

Hence, in order to solve problem (5.3) with relative accuracy δ (and, therefore, solve (5.1) with the same relative accuracy), method (3.3) needs at most

$$\tfrac{4}{\delta}\sqrt{(1+\delta)r\ln r} \tag{5.11}$$

iterations. Note that this bound does not depend on a particular problem instance.

At each iteration of method (3.3) as applied to the problem (5.3) with $d(x)$ defined by (5.9) it is necessary to compute twice a projection of a point onto the set Q with respect to Euclidean metric $\|\cdot\|_G$. This operation is easy in the following cases.

- The set Q is an affine subspace in R^m. Then the projection can be computed by inverting the matrix G. An important example of such a problem is as follows:

$$\min_{y \in R^m}\left\{\rho\left(\sum_{i=1}^m y^{(i)} A_i\right) : \ y^{(1)} = 1\right\}.$$

- The matrix G and the set Q are both simple. For example, if $\langle A_i, A_j\rangle = 0$ for $i \neq j$, then G is a diagonal matrix. In this case, a projection onto a box, for example, is easy to compute. Such a situation occurs when the matrix $A(y)$ is parameterized directly by its entries.

Finally, note that the computation of the value and the gradient of function $f_p(y)$ can be done without eigenvalue decomposition of matrix $A(y)$. Indeed, let $p = 2^k$ satisfies condition (5.5). Consider the following of sequence of matrices:

$$\begin{aligned} X_0 &= A(y), \quad Y_0 = I_n, \\ X_i &= X_{i-1}^2, \quad Y_i = Y_{i-1} X_{i-1}, \quad i = 1, \ldots, k. \end{aligned} \tag{5.12}$$

By induction, it is easy to see that $X_k = A^p(y)$ and $Y_k = A^{p-1}(y)$. Hence, in accordance to (2.1), (2.4), and definition of function $f_p(y)$ in (5.3), we have:

$$\begin{aligned} f_p(y) &= \tfrac{1}{2} \langle X_k, I_n \rangle_M^{2/p}, \\ \nabla f_p(y)^{(i)} &= \tfrac{2 f_p(y)}{\langle X_k, I_n \rangle_M} \cdot \langle Y_k, A_i \rangle_M, \quad i = 1, \ldots, m. \end{aligned}$$

Note that the complexity of computing the matrix $A(y)$ is of the order $O(n^2 m)$ arithmetic operations. The auxiliary computation (5.12) takes

$$O(n^3 \ln p) = O\left(n^3 \ln \tfrac{\ln r}{\delta}\right)$$

operations. After that, the vector $\nabla f_p(y)$ can be computed in $O(n^2 m)$ arithmetic operations. Clearly, the complexity of the first and the last computation is much lower if the matrices A_i are sparse.

Note also, that the computation (5.12) can be performed more efficiently if the matrix $A(y)$ is represented in the form

$$A(y) = U T U^T, \quad U U^T = I_n,$$

where T is a three-diagonal matrix. This representation takes $O(n^3)$ arithmetic operations.

References

[1] C. Helmberg and F. Rendl. A spectral bundle method for semidenite programming. Technical Report SC 97-37, Konrad-Zuse-Zentrum fur Informationstechnik Berlin, August 1997.

[2] C. Helmberg and F. Oustry. Bundle methods to minimize the maximum eigenvalue function. In Lieven Vandenberghe R. Saigal and H. Wolkovicz, editors, *Hanbook on Semidefinite Programming. Theory, Algorithms and Applications*. Kluwer Academic Publisher, 1999.

[3] C. Lemarechal and F. Oustry. Nonsmooth algorithms to solve semidefinite programs. In L. El Ghaoui and S-I. Niculescu, editors, *Recent Advances on LMI methods in Control*, Advances in Design and Control series. SIAM, 1999.

[4] M. V. Nayakkankuppam and Y. Tymofejev. A parallel implementation of the spectral bundle method for large-scale semidefinite programs. *Proceedings of the 8th SIAM Conference on Applied Linear Algebra*, Williamsburg (VA), 2003.

[5] A.S. Nemirovskii and D.B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley, Chichester (1983).

[6] Yu. Nesterov, *Introductory lectures on Convex Optimization. A Basic course.* Kluwer, Boston/Dordrech/London, 2004.

[7] Yu. Nesterov. Smooth minimization of nonsmooth functions. CORE DP 2003/12. Accepted by *Mathematical Programming*.

[8] Yu. Nesterov. Excessive gap technique in non-smooth convex minimization. CORE DP 2003/35. Accepted by *SIOPT*.

[9] Yu. Nesterov. Unconstrained convex minimization in relative scale. CORE DP 2003/96.

[10] F. Oustry, A second order bundle method to minimize the maximum eigenvalue function, Mathematical Programming 89 (2000), pp. 1-33.

6 Appendix: Some properties of smooth convex functions

In this section we present some useful inequalities for smooth convex function, whose smoothness is measured with respect to an arbitrary norm. These facts seems to be well known. However, in the literature they are usually proved for Euclidean norms (see, for example, [6]).

Let E be a finite dimensional real vector space. Denote by E^* the dual space, which is formed by linear functions on E. Let $\langle s, x\rangle$ be a scalar product of elements $s \in E^*$ and $x \in E$. Then, any norm $\|\cdot\|$ on E defines a dual norm on E^*:

$$\|s\|_* = \max_{x \in E}\{\langle s, x\rangle : \|x\| \le 1\}, \quad s \in E^*.$$

Thus, any $x \in E$ and $s \in E^*$ satisfy Cauchy-Schwartz inequality

$$\langle s, x\rangle \le \|s\|_* \cdot \|x\|.$$

For a differentiable function $f(x)$, $x \in E$, we write $f \in \mathcal{F}_L^{1,1}(E)$ if f is convex and

$$\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|, \quad \forall x, y \in E. \tag{6.1}$$

In other words, such a function has Lipschitz continuous gradient.

The proof of the following theorem almost coincide with that of Theorem 2.1.2 [6].

Theorem 3 *All conditions below, holding for all x, $y \in E$, are equivalent to inclusion $f \in \mathcal{F}_L^{1,1}(E)$:*

$$0 \le f(y) - f(x) - \langle \nabla f(x), y - x\rangle \le \tfrac{L}{2} \| x - y \|^2, \tag{6.2}$$

$$f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{2L} \| \nabla f(x) - \nabla f(y) \|_*^2 \le f(y), \tag{6.3}$$

$$\tfrac{1}{L} \| \nabla f(x) - \nabla f(y) \|_*^2 \le \langle \nabla f(x) - \nabla f(y), x - y\rangle, \tag{6.4}$$

$$0 \le \langle \nabla f(x) - \nabla f(y), x - y\rangle \le L \| x - y \|^2, \tag{6.5}$$

Proof:
Assume that f is convex and (6.1) holds. Then, for any x and y from Q we have

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y - x\rangle + \int_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x\rangle d\tau \\ &\le f(x) + \langle \nabla f(x), y - x\rangle + \int_0^1 \tau \cdot L\|y - x\|^2 d\tau, \end{aligned} \tag{6.6}$$

and (6.2) follows from the definition of convex functions. Further, let us fix $x_0 \in E$. Consider the function

$$\phi(y) = f(y) - \langle \nabla f(x_0), y\rangle.$$

Note that $\phi \in \mathcal{F}_L^{1,1}(E)$ and its point of global minimum is $y^* = x_0$. Therefore, in view of (6.2), we have

$$\begin{aligned} \phi(y^*) &\leq \min_{x \in E}[\phi(y) + \langle \nabla\phi(y), x - y\rangle + \tfrac{L}{2}\|y - x\|^2] \\ &= \min_{x \in E}[\phi(y) - \|\nabla\phi(y)\|_* \cdot \|x - y\| + \tfrac{L}{2}\|y - x\|^2] \\ &= \phi(y) - \tfrac{1}{2L} \parallel \phi'(y) \parallel_*^2 . \end{aligned}$$

And we get (6.3) since $\phi'(y) = f'(y) - f'(x_0)$.

We obtain (6.4) from inequality (6.3) by adding two copies of it with x and y interchanged. Applying now Cauchy–Schwartz inequality, we get (6.1). Thus, we proved the following chain

$$(6.1) \Rightarrow (6.2) \Rightarrow (6.3) \Rightarrow (6.4) \Rightarrow (6.1).$$

Finally, we obtain (6.5) from inequality (6.2) by adding two copies of it with x and y interchanged. At the same time, (6.2) leads to (6.5) by integration (6.6). □

In this paper we often use the following condition

Theorem 4 *Two times continuously differentiable function f belongs to $\mathcal{F}_L^{1,1}(E)$ if for any x and h from R^n we have*

$$0 \leq \langle \nabla^2 f(x)h, h\rangle \leq L\|h\|^2. \tag{6.7}$$

Proof:
Indeed, for any x and y from E we have

$$0 \leq \langle \nabla f(y) - \nabla f(x), y - x\rangle = \int_0^1 \langle \nabla^2 f(x + \tau(y - x))(y - x), y - x\rangle d\tau \leq L\|y - x\|^2.$$

Hence, condition (6.5) holds and $f \in \mathcal{F}_L^{1,1}(E)$ by Theorem 3. □