A STOCHASTIC SMOOTHING ALGORITHM FOR SEMIDEFINITE PROGRAMMING.

ALEXANDRE D'ASPREMONT AND NOUREDDINE EL KAROUI

ABSTRACT. We use a rank one Gaussian perturbation to derive a smooth stochastic approximation of the maximum eigenvalue function. We then combine this smoothing result with an optimal smooth stochastic optimization algorithm to produce an efficient method for solving maximum eigenvalue minimization problems. We show that the complexity of this new method is lower than that of deterministic smoothing algorithms in certain precision/dimension regimes.

1. INTRODUCTION

We discuss applications of stochastic smoothing results to the design of efficient first-order methods for solving semidefinite programs. We focus here on the problem of minimizing the maximum eigenvalue of a matrix over a simple convex set Q (the meaning of simple will be made precise later), i.e. solve

$$\min_{X \in Q} \lambda_{\max}(X), \tag{1}$$

in the variable $X \in \mathbf{S}_n$. Note that all primal semidefinite programs with fixed trace have a dual which can be written in this form. While moderately sized problem instances are solved very efficiently by interior point methods [Ben-Tal and Nemirovski, 2001] with very high precision guarantees, these methods fail on most large-scale problems because the cost of running even one iteration becomes too high. When coarser precision targets are sufficient (e.g. spectral methods in statistical or geometric applications), much larger problems can be solved using first-order algorithms, which tradeoff a lower cost per iteration in exchange for a degraded dependence on the target precision.

So far, roughly two classes of first-order algorithms have been used to solve large-scale instances of the semidefinite program in (1). The first uses subgradient descent or a variant of the mirror-prox algorithm of [Nemirovskii and Yudin, 1979] that takes advantage of the geometry of Q to directly minimize $\lambda_{\max}(X)$. These methods do not exploit the particular structure of problem (1) and need $O(D_Q^2/\epsilon^2)$ iterations to reach a target precision ϵ, where D_Q is the diameter of the set Q. Each iteration requires computing a leading eigenvector of the matrix X at a cost of roughly $O(n^2 \log n)$ and projecting X on Q at a cost written p_Q. Spectral bundle methods [Helmberg and Rendl, 2000] use more information on the spectrum of X to speed up convergence, but their complexity is not well understood. More recently, [Nesterov, 2007a] showed that one could exploit the particular min-max structure of problem (1) by first regularizing the objective (using a "soft-max" exponential smoothing), then using optimal first-order methods for smooth convex minimization. These algorithms only require $O(\sqrt{\log n}/\epsilon)$ iterations, but each iteration forms a matrix exponential at a cost of $O(n^3)$. In other words, depending on problem size and precision targets, existing first-order algorithms offer a choice between two complexity bounds

$$O\left(\frac{D_Q^2(n^2 \log n + p_Q)}{\epsilon^2}\right) \quad \text{and} \quad O\left(\frac{D_Q\sqrt{\log n}(n^3 + p_Q)}{\epsilon}\right) \tag{2}$$

Note that the constants in front of all these estimates can be quite large and actual numerical complexity depends heavily on the particular path taken by the algorithm, especially for adaptive variants of the methods

Date: April 4, 2012.

2010 *Mathematics Subject Classification.* 90C22,90C15,47A75.

Key words and phrases. Semidefinite programming, Gaussian smoothing, eigenvalue problems.

detailed here (see [Nesterov, 2007b, §6] for an illustration on a simpler problem). In practice of course, these asymptotic worst case bounds are useful for providing general guidance in algorithmic choices, but remain relatively coarse predictors of performance for reasonable values of n and ϵ.

Many recent works have sought to move beyond these two basic complexity options. Overton and Womersley [1995] directly applied Newton's method to the maximum eigenvalue function, given a priori information on the multiplicity of this eigenvalue. Burer and Monteiro [2003] and Journée et al. [2008] focus on instances where the solution is known to have low rank (e.g. matrix completion, combinatorial relaxations) and solve the problem directly over the set of low rank matrices. These formulations are nonconvex and their complexity cannot be explicitly bounded, but empirical performance is often very good. Lu et al. [2007] focus on the case where the matrix has a natural structure (close to block diagonal). Juditsky et al. [2008] use a variational inequality formulation and randomized linear algebra to reduce the cost per iteration of first-order algorithms. Subsampling techniques were also used in [d'Aspremont, 2011] to reduce the cost per iteration of stochastic averaging algorithms. Finally, in results that are similar, Baes et al. [2011] use stochastic approximations of the matrix exponential to reduce the cost per iteration of smooth first-order methods. The complexity tradeoff and algorithms in this last result are different from ours (roughly speaking, a $1/\epsilon$ term is substituted to the $\sqrt{n}$ term in our bound), but both methods seek to reduce the cost of smooth first-order algorithms for semidefinite programming using stochastic gradient oracles instead of deterministic ones.

In this paper, we use stochastic smoothing results, combined with an optimal accelerated algorithm for stochastic optimization recently developed by Lan [2009], to derive a stochastic algorithm for solving (1) which requires only $O(\sqrt{n}/\epsilon)$ iterations, with each iteration computing a few sample leading eigenvectors of $(X + \epsilon zz^T/n)$ where $z \sim \mathcal{N}(0, \mathbf{I}_n)$. While in most applications of stochastic optimization the noise level is seen as exogenous, we use it in the algorithm detailed here to control the tradeoff between number of iterations and cost per iteration. The algorithm requires fewer iterations than nonsmooth methods and has lower cost per iteration than smoothing techniques. In some configurations of the parameters (n, ϵ, p_Q, D_Q), its total worst-case floating-point complexity is lower than that of both smooth and nonsmooth methods. Overall, the method has a cost per iteration comparable to that of nonsmooth methods while retaining some of the benefits of accelerated methods for smooth optimization.

The paper is organized as follows. In the next section, be briefly outline our stochastic smoothing algorithm for maximum eigenvalue minimization and compare its complexity with existing first-order algorithms. Section 3 details our main smoothing results on random rank one perturbations of the maximum eigenvalue function, highlighting in particular a phase transition in the spectral gap depending on the spectrum of the original matrix. Section 4 uses these smoothing results to produce a stochastic algorithm for maximum eigenvalue minimization, and describes an extension of the optimal stochastic optimization algorithm in [Lan, 2009] where the scale of the step size is allowed to vary adaptively (but monotonically). Section 5 informally discusses extensions of our results to other smoothing techniques, together with their impact on complexity. Section 6 presents some preliminary numerical experiments. An appendix contains auxiliary material, including a detailed discussion of the cost of computing leading eigenpairs of a symmetric matrix and a proof of the phase transition result for random rank-one perturbations.

Notation. Throughout the paper, we denote by $\lambda_i(X)$ the eigenvalues of the matrix $X \in \mathbf{S}_n$, in decreasing order. For clarity, we will also use $\lambda_{\max}(X)$ for the leading eigenvalue of X. When z denotes a vector in $\mathbb{R}^n$, its i-th coordinate is denoted by z_i. We denote equality in law (for random variables) by $\overset{\mathcal{L}}{=}$ and $\Longrightarrow$ stands for convergence in law.

2. Stochastic smoothing algorithm

We will solve a smooth approximation of problem (1), written

$$\begin{array}{ll}\text{minimize} & f(X) \triangleq \mathbf{E}\left[\max_{i=1,\ldots,k} \lambda_{\max}\left(X + \frac{\epsilon}{n} z_i z_i^T\right)\right] \\ \text{subject to} & X \in Q,\end{array} \tag{3}$$

in the variable $X \in \mathbf{S}_n$, where $Q \subset \mathbb{R}^n$ is a compact convex set, z_i are i.i.d. Gaussian vectors $z_i \sim \mathcal{N}(0, \mathbf{I}_n)$ and $k > 0$ is a small constant (typically 3). We call f^* the optimal value of this problem. The fact that $\lambda_{\max}(\cdot)$ is 1-Lipschitz with respect to the spectral norm with $\lambda_{\max}(z_i z_i^T) = \|z_i\|_2^2$, yields

$$\mathbf{E}\left[\max_{i=1,\ldots,k} \lambda_{\max}\left(X + \frac{\epsilon}{n} z_i z_i^T\right)\right] \leq \lambda_{\max}(X) + c_k \epsilon$$

where

$$c_k = \mathbf{E}\left[\max_{i=1,\ldots,k} \|z_i\|_2^2/n\right] \leq \mathbf{E}\left[\sum_{i=1}^k \|z_i\|_2^2/n\right] = k$$

depends only on k. Jensen's inequality and $\mathbf{E}[z_i z_i^T] = \mathbf{I}_n$ also yield

$$\lambda_{\max}(X) + \frac{\epsilon}{n} \leq \mathbf{E}\left[\max_{i=1,\ldots,k} \lambda_{\max}\left(X + \frac{\epsilon}{n} z_i z_i^T\right)\right].$$

This means that $f(X)$ will be a $c_k\epsilon$-uniform approximation of $\lambda_{\max}(X)$. We begin by briefly introducing the smoothing results on (3) detailed in Section 3, then describe our main algorithm.

2.1. **Smoothness of $f(X)$.** In Section 3, we will show that the function f has a Lipschitz continuous gradient w.r.t. the Frobenius norm, i.e.

$$\|\nabla f(X) - \nabla f(Y)\| \leq L\|X - Y\|_F$$

with constant L satisfying

$$L \leq C_k \frac{n}{\epsilon} \tag{4}$$

where $C_k > 0$ depends only on k and is bounded whenever $k \geq 3$. We will see in Section 3 that this bound is quite conservative and that much better regularity is achieved when the spectrum of X is well-behaved (see Theorem 3.8).

2.2. **Gradient variance.** Section 3 also shows that the function f is differentiable. Let ϕ be a leading eigenvector of the matrix $X + \frac{\epsilon}{n} z_{i_0} z_{i_0}^T$ where

$$i_0 = \underset{i=1,\ldots,k}{\operatorname{argmax}}\, \lambda_{\max}\left(X + \frac{\epsilon}{n} z_i z_i^T\right).$$

We will see that i_0 is unique with probability one. We have

$$\nabla f(X) = \mathbf{E}\left[\phi\phi^T\right] \quad \text{and} \quad \mathbf{E}\left[\left\|\phi\phi^T - \nabla f(X)\right\|_F^2\right] \leq 1\,. \tag{5}$$

Therefore the variance of the stochastic gradient oracle $\phi\phi^T$ is bounded by one. Once again, we will see in Section 3 that this bound is often quite conservative.

2.3. **Stochastic algorithm.** Given an unbiased estimator for ∇f with unit variance, the optimal algorithm for stochastic optimization derived in [Lan, 2009] will produce a matrix X_N such that

$$\mathbf{E}[f(X_N) - f^*] \leq \frac{4LD_Q^2}{N^2} + \frac{4D_Q}{\sqrt{Nq}} \tag{6}$$

after N iterations [Lan, 2009, Corollary 1], where $L \leq C_k n/\epsilon$ is the Lipschitz constant of ∇f discussed in the previous section and q is the number of sample matrices $\phi\phi^T$ averaged in approximating the gradient. Once again, we write D_Q the diameter of the set Q (see below for a precise definition) and p_Q, which appears in Table 1, the cost of projecting a matrix $X \in \mathbf{S}_n$ on the set Q.

Setting $N = 2D_Q\sqrt{n}/\epsilon$ and $q = \max\{1, D_Q/(\epsilon\sqrt{n})\}$ will then ensure $\mathbf{E}[f(X_N) - f^*] \leq 5\epsilon$. We compare in Table 1 the computational cost of the smooth stochastic algorithm in [Lan, 2009, Corollary 1] in this setting with that of the smoothing technique in [Nesterov, 2007a] and the nonsmooth stochastic

Algorithmic complexity	Num. of Iterations	Cost per Iteration
Nonsmooth	$O\left(\frac{D_Q^2}{\epsilon^2}\right)$	$O(p_Q + n^2 \log n)$
Stochastic smoothing	$O\left(\frac{D_Q\sqrt{n}}{\epsilon}\right)$	$O\left(p_Q + \max\left\{1, \frac{D_Q}{\epsilon\sqrt{n}}\right\} n^2 \log n\right)$
Deterministic Smoothing	$O\left(\frac{D_Q\sqrt{\log n}}{\epsilon}\right)$	$O(p_Q + n^3)$

TABLE 1. Worst-case computational cost of the smooth stochastic algorithm detailed here, the smoothing technique in [Nesterov, 2007a] and the nonsmooth subgradient descent method.

averaging method. Recall that the cost of computing one leading eigenvector of $X + vv^T$ is of order $O(n^2 \log n)$ while that of forming the matrix exponential $\exp(X)$ is $O(n^3)$ [Moler and Van Loan, 2003].

Table 1 shows a clear tradeoff in this group of algorithms between the number of iterations and the cost of each iteration. In certain regimes for (n, ϵ), the total worst-case complexity of the smooth stochastic algorithm is lower than that of both smooth and nonsmooth methods. This is the case for instance when

$$c_1 \max\left\{1, \frac{D_Q}{\epsilon\sqrt{n}}\right\} n^2 \log n \leq p_Q \leq c_2 n^{5/2} \sqrt{\log n}$$

for some absolute constants $c_1, c_2 > 0$. In practice of course, the constants in front of all these estimates can be quite large and the key contribution of the algorithm detailed here is to preserve some of the benefits of smooth accelerated methods (e.g. fewer iterations), while requiring a much lower computational (and memory) cost per iteration by exploiting the very specific structure of the $\lambda_{\max}(X)$ function.

3. EFFICIENT STOCHASTIC SMOOTHING

In this section, we show how to regularize the function $\lambda_{\max}(X)$ using stochastic smoothing arguments. We begin by recalling a classical argument about Gaussian regularization; we then improve smoothing performance by exploiting some explicit structural results on the spectrum of rank one updates of symmetric matrices.

3.1. **Gaussian smoothing.**

We first recall a standard result on Gaussian smoothing which does not exploit any structural information on the function $\lambda_{\max}(X)$ except its Lipschitz continuity.

Lemma 3.1. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz continuous with constant μ with respect to the Euclidean norm. The function g such that*

$$g(x) = \mathbf{E}[f(x + \epsilon z)]$$

where $z \sim \mathcal{N}(0, \mathbf{I}_n)$ and $\epsilon > 0$, has a Lipschitz continuous gradient with

$$\|\nabla g(x) - \nabla g(y)\| \leq \frac{2\sqrt{n}\mu}{\epsilon} \|x - y\|.$$

Proof. See Nesterov [2011] for a short proof and applications in gradient-free optimization. ■

Let us consider the function

$$\mathbf{E}[\lambda_{\max}(X + (\epsilon/\sqrt{n})U)] \ ,$$

where $U \in \mathbf{S}_n$ is a symmetric matrix with standard normal upper triangle coefficients. Using again Jensen's inequality, the fact that $\lambda_{\max}(X)$ is 1-Lipschitz with respect to the spectral norm and bounds on the largest eigenvalue of U (which follow easily from either Trotter [1984] or Davidson and Szarek [2001]), we see that this function is an ϵ-approximation of $\lambda_{\max}(X)$.

The lemma above shows that it has a Lipschitz continuous gradient with constant bounded by $O\left(n^{3/2}/\epsilon\right)$. This approach was used e.g. in [d'Aspremont, 2008] to reduce the cost per iteration of a smooth optimization algorithm with approximate gradient, and by [Nesterov, 2011] to derive explicit complexity bounds on gradient free optimization methods.

3.2. **Gradient smoothness.** We recall the following classical result (which can be derived from results in [Kato, 1995] and is proved in the appendix for the sake of completeness) showing that the gradient of $\lambda_{\max}(X)$ is smooth when the largest eigenvalue of X has multiplicity one, with (local) Lipschitz constant controlled by the spectral gap.

Theorem 3.2. *Suppose $X \in \mathbf{S}_n$ and call $\{\lambda_i(X)\}_{i=1}^n$ the decreasingly ordered eigenvalues of X. Suppose $\lambda_{\max}(X)$, the largest eigenvalue of X, has multiplicity one. Let Y be a symmetric matrix with $\|Y\|_F = 1$ and call*

$$g(X,Y) = \lim_{t \to 0} \frac{\partial^2 \lambda_{\max}(X+tY)}{\partial t^2}.$$

Then the local Lipschitz constant of the gradient is given by

$$\|\nabla \lambda_{\max}(X)\|_L = \sup_{Y \in \mathbf{S}_n, \|Y\|_F = 1} g(X,Y) = \frac{1}{2} \frac{1}{\lambda_{\max}(X) - \lambda_2(X)}. \tag{7}$$

This result shows that if we want to produce smooth approximations of the function $\lambda_{\max}(X)$ using random perturbations, we need these perturbations to increase the spectral gap by a sufficient margin. We will see below that, up to a small trick, random rank one Gaussian perturbations of the matrix X will suffice to achieve this goal.

3.3. **Rank one updates.** For $X \in \mathbf{S}_n$, we call $\lambda \in \mathbb{R}^n$ the spectrum of the matrix X, in decreasing algebraic order. Whenever $v \neq 0$ is not an eigenvector of X, the leading eigenvalue l_1 of the matrix $X + (\epsilon/n)vv^T$, with $\epsilon > 0$, is always strictly larger than λ_1 [see Golub and Van Loan, 1990, §8.5.3] and we write $l_1 = \lambda_1 + t$.

We assume without loss of generality that X is diagonal. If X were not diagonal, we could simply rotate v. The variable t is the unique positive solution of the *secular equation*

$$s(t) \triangleq \frac{n}{\epsilon} - \frac{\mathsf{v}_1^2}{t} - \sum_{i=2}^{n} \frac{\mathsf{v}_i^2}{(\lambda_1 - \lambda_i) + t} = 0. \tag{8}$$

where v_i are the coefficients of the vector v. We plot the function s for a sample matrix X in Figure 1.

Equation (8) implies that we have almost explicit expressions for the eigenvalue decomposition of the matrix

$$X + \frac{\epsilon}{n} vv^T$$

where $v \in \mathbb{R}^n$ and $\epsilon > 0$. Having assumed that X is diagonal. Golub and Van Loan [1990, Th. 8.5.3] also shows that if $\mathsf{v}_i \neq 0$ for $i = 1, \ldots, n$ and $\epsilon > 0$, then $t > 0$ and the eigenvalues of X and $X + (\epsilon/n)vv^T$ are interlaced, i.e.

$$\lambda_n(X) \leq \lambda_n(X + \frac{\epsilon}{n}vv^T) \leq \ldots \leq \lambda_{\max}(X) < \lambda_{\max}(X + \frac{\epsilon}{n}vv^T).$$

By construction, the function

$$s^+(t) \triangleq \frac{n}{\epsilon} - \frac{\mathsf{v}_1^2}{t}$$

is an upper bound on $s(t)$ on the interval $(0, \infty)$. This means that the positive root of $s^+(t)$ is a lower bound on the positive root t^* of $s(t)$ and we get

$$t^* \geq \frac{\epsilon \mathsf{v}_1^2}{n}.$$

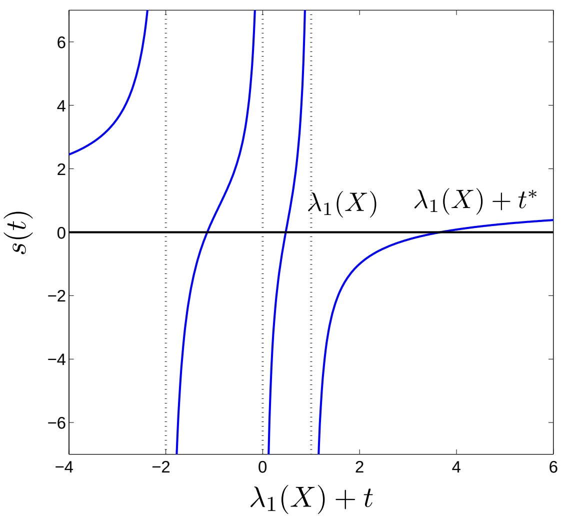

FIGURE 1. Plot of $s(t)$ versus $\lambda_1(X)+t$. The matrix has dimension four and its spectrum is here $\{-2,-2,0,1\}$. The three leading eigenvalues of $X+\epsilon vv^T$ are the roots of $s(t)$, the fourth eigenvalue is at -2.

Using interlacing, we have

$$\lambda_2(X+\frac{\epsilon}{n}vv^T) \leq \lambda_1(X) \leq \lambda_1(X)+t^* = \lambda_1(X+\frac{\epsilon}{n}vv^T).$$

This gives a lower bound on the spectral gap of the perturbed matrix

$$\frac{\epsilon \mathsf{v}_1^2}{n} \leq t^* \leq \lambda_1(X+\frac{\epsilon}{n}vv^T) - \lambda_2(X+\frac{\epsilon}{n}vv^T)\ , \tag{9}$$

which will allow us to control the smoothness of $\nabla f(X)$.

3.4. **Low rank Gaussian smoothing.** We now come back to the objective function of problem (3), written

$$f(X) \triangleq \mathbf{E}\left[\max_{i=1,\ldots,k} \lambda_{\max}\left(X+\frac{\epsilon}{n}z_iz_i^T\right)\right]$$

where z_i are i.i.d $\mathcal{N}(0,\mathbf{I}_n)$ and $k>0$ is a small constant. We first show that we can differentiate under the expectation in the definition of $f(X)$.

Lemma 3.3. *Let λ_1+T be the largest eigenvalue of the matrix $X+(\epsilon/n)zz^T$, where X is a given deterministic matrix and $z\sim\mathcal{N}(0,\mathbf{I}_n)$. Then the random variable T has a density.*

Proof. As usual, we call $\{\lambda_i\}_{i=1}^n$ the decreasingly ordered eigenvalues of X and assume here that $\lambda_{\max}(X)$ has multiplicity $l<n$ (if $l=n$ there is nothing to show). By rotational invariance of the standard Gaussian distribution, we can and do assume that X is diagonal. As we have seen before, T is therefore the positive root of the equation

$$0 = s(T) = \frac{n}{\epsilon} - \frac{\sum_{i=1}^l \mathsf{z}_i^2}{T} - \sum_{i=l+1}^n \frac{\mathsf{z}_i^2}{(\lambda_1-\lambda_i)+T},$$

and note that $s(t)$ is increasing in t. Hence,

$$\begin{aligned} P(T \geq t) &= P(s(T) \geq s(t)) = P(0 \geq s(t)) \\ &= P\left(\frac{\sum_{i=1}^{l} \mathsf{z}_i^2}{t} + \sum_{i=l+1}^{n} \frac{\mathsf{z}_i^2}{(\lambda_1 - \lambda_i) + t} \geq \frac{n}{\epsilon}\right), \\ &= \int_{\frac{1}{\epsilon}}^{\infty} p_t(u) du \triangleq I(t) \ . \end{aligned}$$

where p_t is the density of the random variable

$$Y_t = \frac{1}{n}\left(\frac{\sum_{i=1}^{l} \mathsf{z}_i^2}{t} + \sum_{i=l+1}^{n} \frac{\mathsf{z}_i^2}{(\lambda_1 - \lambda_i) + t}\right) .$$

If the integral $I(t)$ can be differentiated under the integral sign, then we can differentiate $P(T \geq t)$ and we will have established the existence of a density for T and hence for $\lambda_1 + T$. Now $p_t(x)$ is a very smooth function of both t and x. Indeed, it is a convolution of $n - l$ densities that are smooth in t and x. Recall that if X has density f and $t > 0$, X/t has density $tf(t\cdot)$. Recall also that a random variable with χ_l^2 distribution has density

$$f_l(x) = \frac{2^{-l/2}}{\Gamma(l/2)} x^{l/2-1} \exp(-x/2) \ .$$

So it is clear that for any k, any $t > 0$, and any $\alpha \geq 0$, $(t + \alpha)f_l((t + \alpha)x)$ is C^∞ in t. Applying this result in connection to [Durrett, 2010, Th. A.5.1], we see that Y_t has a density which is a smooth function of $t > 0$. Indeed, it is C^∞ on $(0, \infty)$. Moreover, it is easy to see that the conditions of [Durrett, 2010, Th. A.5.1] are satisfied for p_t, which guarantees that we can differentiate under the integral sign. This shows that for any $t > 0$, the function g such that $g(t) = P(T \geq t)$ is differentiable in t. It is also clear that $P(T = 0)$ is 0, so this distribution has no atoms at 0. We conclude that T has a density on $(0, \infty)$. ■

We then directly obtain the following corollaries. The first shows that two perturbed eigenvalues obtained from independent rank one perturbations are different with probability one.

Corollary 3.4. *Suppose $l_{1,1} = \lambda_{\max}(X + (\epsilon/n)z_1 z_1^T)$ and $l_{1,2} = \lambda_{\max}(X + (\epsilon/n)z_2 z_2^T)$, where z_1 and z_2 are independent with distribution $\mathcal{N}(0, \mathbf{I}_n)$. Then $l_{1,1} \neq l_{1,2}$ with probability one.*

Proof. Follows from Lemma 3.3 since $l_{1,1}$ and $l_{1,2}$ are two independent draws from a distribution with a density on $(0, \infty)$ and $P(l_{1,1} = 0) = P(l_{1,2} = 0) = 0$. ■

The second corollary shows that the maximum of (independent) perturbed eigenvalues is differentiable with probability one.

Corollary 3.5. *Suppose $l_{1,i} = \lambda_{\max}(X + (\epsilon/n)z_i z_i^T)$, where z_i are i.i.d. with distribution $\mathcal{N}(0, \mathbf{I}_n)$ for $i = 1, \ldots, k$. Then the mapping $X \to \max_{i=1,\ldots,k} l_{1,i}$ is differentiable with probability one.*

Proof. This corollary follows from the previous one and the fact that if g and h are two differentiable functions, then $\max(g, h)$ is differentiable at all points x such that $g(x) \neq h(x)$. Indeed, in that case $[\max(g, h)]'(x) = g'(x)1_{g(x)>h(x)} + h'(x)1_{g(x)<h(x)}$. ■

We now use these preliminary results to produce a bound on the Lipschitz constant of the gradient of $f(X)$ defined above.

Proposition 3.6. *Let $\{z_i\}_{i=1}^k$ be i.i.d. $\mathcal{N}(0, \mathbf{I}_n)$. The function $f(X)$ such that*

$$f(X) = \mathbf{E}\left[\max_{i=1,\ldots,k} \lambda_{\max}(X + (\epsilon/n)z_i z_i^T)\right]$$

is smooth and the Lipschitz constant of its gradient w.r.t. the Frobenius norm is bounded by

$$L \leq C_k \frac{n}{\epsilon} \quad \textit{where} \quad C_k = \frac{1}{\sqrt{2}} \frac{k}{k-2}$$

when $k \geq 3$.

Proof. Writing z_i the first coordinate of the vector z_i and combining Theorem 3.2, Corollary 3.5 and the spectral gap bound in (9) shows

$$\|\nabla \lambda_{\max}(X)\|_L \leq \mathbf{E}\left[\frac{n}{2\epsilon} \min_{i=1,\ldots,k} \frac{1}{\mathsf{z}_i^2}\right] \leq \mathbf{E}\left[\frac{n}{2\epsilon} \frac{k}{\chi_k^2}\right]$$

where χ_k^2 is χ^2 distributed with k d.f. The fact that

$$\mathbf{E}\left[\frac{1}{\chi_k^2}\right] = \frac{1}{2^{k/2}\Gamma(k/2)} \int_0^\infty t^{\frac{k-2}{2}-1} e^{-t/2} dt = \frac{\Gamma\left(\frac{k}{2}-1\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{2}},$$

whenever $k \geq 3$ yields the desired result, since $\Gamma(x+1) = x\Gamma(x)$. ■

Note that the bound above is a bit coarse; numerical simulations show that for independent $\mathcal{N}(0,1)$ random variables $\{\mathsf{z}_i\}_{i=1}^3$,

$$\frac{1}{2}\mathbf{E}\left[1/\max\{\mathsf{z}_1^2, \mathsf{z}_2^2, \mathsf{z}_3^2\}\right] = .75...$$

while $C_3 = 2.12...$, for example. We could of course use the density of the minimum above to get a more accurate bound but C_k would not have a simple closed form then.

3.5. **Gradient variance.** In this section, we will bound the variance of the stochastic gradient oracle approximating ∇f. Let us call

$$g(X,z) = \lambda_{\max}(X + \frac{\epsilon}{n} z z^T) \ .$$

Because of the rotational invariance of both $\lambda_{\max}(\cdot)$ and of the Gaussian distribution, we can assume without loss of generality that X is diagonal and that its largest eigenvalue has multiplicity l.

Lemma 3.7. *Suppose w.l.o.g. that the matrix* $X \in \mathbf{S}_n$ *is diagonal, and* $z_i \sim \mathcal{N}(0, \mathbf{I}_n)$*, the gradient of*

$$f(X) = \mathbf{E}\left[\max_{i=1,\ldots,k} \lambda_{\max}(X + (\epsilon/n) z_i z_i^T)\right]$$

is also diagonal. It is given by

$$\nabla f(X) = \mathbf{E}[\phi_{i_0}\phi_{i_0}^T]$$

where ϕ_{i_0} *is the leading eigenvector of the matrix* $X + \frac{\epsilon}{n} z_{i_0} z_{i_0}^T$*, and*

$$i_0 = \operatorname*{argmax}_{i=1,\ldots,k} \lambda_{\max}\left(X + \frac{\epsilon}{n} z_i z_i^T\right) \ .$$

We have

$$\mathbf{E}\left[\|\phi_{i_0}\phi_{i_0}^T - \mathbf{E}[\phi_{i_0}\phi_{i_0}^T]\|_F^2\right] = 1 - \mathbf{Tr}\left(\nabla f(X)^2\right) \leq 1, \tag{10}$$

where $\mathbf{Tr}\left(\nabla f(X)\right) = 1$ *by construction.*

Proof. As above, $z \sim \mathcal{N}(0, \mathbf{I}_n)$ means that z is not an eigenvector of X with probability one. Call $\lambda_i(X)$ the eigenvalues of X in decreasing order. This means that $g(X,z)$ has multiplicity one and we call $\phi(X,z)$ the corresponding eigenvector. We know that $\nabla g(X,z) = \phi(X,z)\phi(X,z)^T$ with

$$\phi(X,z)_i = c \frac{\mathsf{z}_i}{g(X,z) - \lambda_i},$$

where $c > 0$ is a normalizing factor. Recall that $g(X, z)$ is the largest root of $\chi(\lambda) = 0$, where

$$\chi(\lambda) = 1 + \frac{n}{\epsilon} \frac{\sum_{i=1}^{l} \mathrm{z}_i^2}{\lambda_1(X) - \lambda} + \frac{n}{\epsilon} \sum_{i=l+1}^{n} \frac{\mathrm{z}_i^2}{\lambda_i(X) - \lambda} .$$

Call $\mathfrak{s}$ a vector of ± 1 and $z[\mathfrak{s}] = \mathfrak{s} \circ z$, i.e $z[\mathfrak{s}]_i = \mathfrak{s}_i \mathrm{z}_i$ for $i = 1, \ldots, n$. The secular equation above depends only on z_i^2, hence $l_1(X, z) = g(X, z)$ also depends only on z_i^2 and we have, for any z and $\mathfrak{s}$,

$$\lambda_1(X + \frac{\epsilon}{n} z z^T) = \lambda_1(X + \frac{\epsilon}{n} z[\mathfrak{s}] z[\mathfrak{s}]^T),$$

which means

$$g(X, z) = g(X, \mathfrak{s} \circ z) = g(X, z[\mathfrak{s}]).$$

Let us call

$$M(X, z) = \phi(X, z) \phi(X, z)^T .$$

We use an invariance argument to show that the symmetric matrix $\nabla f(X) = \mathbf{E}[M(X, z)]$ is in fact diagonal. We write $A = \nabla f(X)$ in what follows to simplify notations. Take a vector z, change its i-th coordinate to $-\mathrm{z}_i$ and call $\mathfrak{s}^{(i)}$ the corresponding $\mathfrak{s}$ sign vector. The i^{th} coordinate of ϕ is changed to its opposite, but all the other coordinates remain the same, while $g(X, z) = g(X, \mathfrak{s}^{(i)} \circ z)$. This means that

$$M_{i,j}(X, z) = -M_{i,j}(X, \mathfrak{s}^{(i)} \circ z) \text{ , when } i \neq j.$$

The coordinatewise product of $\mathfrak{s}^{(i)}$ and z has the same law as z, i.e. $z \overset{\mathcal{L}}{=} \mathfrak{s}^{(i)} \circ z$, so

$$M_{i,j}(X, z) \overset{\mathcal{L}}{=} M_{i,j}(X, \mathfrak{s}^{(i)} \circ z),$$

and we conclude from the first equation that $\mathbf{E}[M_{i,j}(X, z)] = -\mathbf{E}[M_{i,j}(X, \mathfrak{s}^{(i)} \circ z)]$ and from the second equation that $\mathbf{E}[M_{i,j}(X, z)] = \mathbf{E}[M_{i,j}(X, \mathfrak{s}^{(i)} \circ z)]$, hence

$$A_{i,j} = \mathbf{E}[M_{i,j}(X, z)] = -\mathbf{E}[M_{i,j}(X, z)] = 0 \text{ when } i \neq j,$$

and $\mathbf{E}[M]$ is diagonal, as announced. We now focus on the variance $\mathbf{E}[\|\phi\phi^T - \mathbf{E}[\phi\phi^T]\|_F^2]$. We can rewrite this expression as

$$\|\phi\phi^T - \mathbf{E}[\phi\phi^T]\|_F^2 = \mathbf{Tr}\,(\phi\phi^T - \mathbf{E}[\phi\phi^T])^2 .$$

Using the fact that $\phi^T\phi = 1$, and $\mathbf{E}[\phi\phi^T] = A$, we see that

$$\mathbf{E}[\mathbf{Tr}\,(\phi\phi^T - \mathbf{E}[\phi\phi^T])^2] = \mathbf{Tr}\,A - \mathbf{Tr}\,A^2 = 1 - \mathbf{Tr}\,A^2 \leq 1$$

Furthermore, recall that A diagonal means $\mathbf{Tr}\,A^2 = \sum_{i=1}^n A_{i,i}^2$, and $\sum_{i=1}^n A_{i,i} = 1$ with $A_{i,i} \geq 0$. ■

Simply using the fact that ϕ_{i_0} is an eigenvector, we have of course

$$\|\phi_{i_0}\phi_{i_0}^T - \mathbf{E}[\phi_{i_0}\phi_{i_0}^T]\|_F^2 \leq 4 \tag{11}$$

which means that the gradient will naturally satisfy the "light-tail" condition A2 in [Lan, 2009] for $\sigma^2 = 4$. The bound in (10) together with the proof above show that when the spectral gaps $\lambda_1(X) - \lambda_i(X)$ are large, the diagonal of $\nabla f(X)$ is approximately sparse. In that scenario, $Tr(\nabla f(X)^2)$ is close to $Tr(\nabla f(X))$, hence close to one, and the variance of the gradient oracle is small.

3.6. **A phase transition.** We are here investigating the properties of a random rank one perturbation of a deterministic matrix X, specifically $X(\epsilon) = X + (\epsilon/\sqrt{n})zz^T$, where $z \sim \mathcal{N}(0, \mathbf{I}_n)$. As we will see, the bounds we obtained above are quite conservative and the Lipschitz constant of the gradient is in fact much lower than n/ϵ when the spectrum of X is well-behaved (in a sense that will be made clear later).

In particular, we will observe that there is a *phase transition phenomenon* in ϵ. Let us call $T = \lambda_{\max}(X(\epsilon)) - \lambda_{\max}(X)$. If the perturbation scale ϵ is small, T is of order $1/n$ (the worst-case bound we obtained above). If ϵ is large, T is of order one. And if ϵ has a critical value, then T is $O_P(1/\sqrt{n})$.

The main idea behind the results we present below is the following. We are looking for the zeros of a certain random function, which can be seen as a perturbation of a deterministic function. Hence, it is natural to use ideas used in asymptotic root finding problems [see Miller, 2006, pp. 36-43], to expand the solution in powers of the size of the perturbation. We note that a similar idea was used in [Nadler, 2008], which focused on a different random matrix problem. We have the following theorem.

Theorem 3.8 (**Phase transition for the largest eigenvalue: rank one perturbation**). *Let X be a symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Suppose λ_1 has multiplicity l and the gap between λ_1 and λ_{l+1}, γ, stays bounded away from 0. (Implicitly everything can change with n; l is held bounded.). Call $\lambda_1 - \lambda_i = \gamma + \delta_i$, for $i > l$. Consider the matrix*

$$X(\epsilon) = X + \frac{\epsilon}{n} zz^T \ ,$$

where z is a vector with i.i.d $\mathcal{N}(0,1)$ entries. Call $l_1(\epsilon)$ the largest eigenvalue of the perturbed matrix $X(\epsilon)$. We assume that $\epsilon \asymp 1$. Define ϵ_0 by

$$\frac{1}{\epsilon_0} = \frac{1}{n} \sum_{j=l+1}^{n} \frac{1}{\gamma + \delta_j} \ .$$

We note that ϵ_0 is actually a function of n, but we do not write it explicitly to simplify notations. We also assume that there exists $C > 0$, independent of n, such that

$$\frac{1}{\gamma^2} > \frac{1}{n} \sum_{l+1}^{n} \frac{1}{(\gamma + \delta_j)^2} > C \ ,$$

and that n is going to infinity. Call, for i.i.d $\mathcal{N}(0,1)$ random variables $\{\mathsf{z}_j\}_{j=1}^n$,

$$\xi_1 = \frac{1}{\sqrt{n}} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2 - 1}{\gamma + \delta_j} = O_P(1) \ ,$$

$$\zeta_1 = \frac{1}{n} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2}{(\gamma + \delta_j)^2} = O_P(1) \ ,$$

and $\sum_{i=1}^{l} \mathsf{z}_i^2 = \chi_l^2$ a χ_l^2 random variable with l degrees of freedom, independent of ξ_1 (note that the estimates of the size of ξ_1 and ζ_1 are key in all the results that follow). We have the following three situations:

(1) *If $0 < \epsilon < \epsilon_0$, i.e $\epsilon_0 - \epsilon$ stays bounded away from 0 when n grows,*

$$l_1(\epsilon) = \lambda_1 + \frac{W_1}{n} + \frac{W_2}{n^{3/2}} + O_P\left(\frac{1}{n^2}\right) ,$$

where

$$W_1 = \frac{\chi_l^2}{1/\epsilon - 1/\epsilon_0} \quad \text{and} \quad W_2 = W_1^2 \xi_1.$$

(2) *If $\epsilon = \epsilon_0$,*

$$l_1(\epsilon) = \lambda_1 + \frac{W_1}{\sqrt{n}} + O_P\left(\frac{1}{n}\right) ,$$

where

$$W_1 = \frac{\xi_1 + \sqrt{\xi_1^2 + 4\chi_l^2 \zeta_1}}{2\zeta_1}.$$

(3) *If $\epsilon > \epsilon_0$, i.e $\epsilon - \epsilon_0$ stays bounded away from 0 when n grows, call $t_0 > 0$, the solution of*

$$\frac{1}{\epsilon} = \frac{1}{n} \sum_{j=l+1}^{n} \frac{1}{t_0 + \gamma + \delta_j}.$$

Note that $t_0 \leq (1 - l/n)\epsilon$. Then

$$l_1(\epsilon) = \lambda_1 + t_0 + \frac{W_1}{\sqrt{n}} + O_P\left(\frac{1}{n}\right).$$

Here,

$$W_1 = \frac{\xi(t_0)}{\zeta(t_0)} \ ,$$

where

$$\xi(t_0) = \frac{1}{\sqrt{n}} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2 - 1}{t_0 + \gamma + \delta_i} = O_P(1) \ ,$$

$$\zeta(t_0) = \frac{1}{n} \sum_{j=l+1}^{n} \frac{1}{(t_0 + \gamma + \delta_i)^2} = O(1) \ .$$

The proof of this theorem is in the Appendix. The phase transition can be further explored in the situation where $\epsilon - \epsilon_0$ is infinitesimal in n but not exactly 0.

We are especially concerned in this paper with random variables of the type

$$\max_{i=1,\ldots,k} \lambda_{\max}(X + (\epsilon/n) z_i z_i^T) - \lambda_{\max}(X)$$

for i.i.d z_i's. The previous theorem gives us an idea of the scale of this difference, which clearly depends on ϵ and the whole spectrum of X. It is also clear that taking a max over finitely many k's does not change anything to the previous result as far as scale is concerned. The previous theorem shows that our uniform bound on the inverse of the gap cannot be improved. However, in many situations, the gap is much greater than $1/n$ and the worst case bound on the Lipschitz constant of $f(X)$ is very conservative.

4. Stochastic composite optimization

In this section, we will develop a variant of the algorithm in [Lan, 2009] which allows for adaptive (monotonic) scaling of the step size parameter. For the sake of completeness, we first recall the principal definitions in [Lan, 2009], adopting the same notation, with only a few minor modifications to allow the full problem to be stochastic. We focus on the following optimization problem

$$\min_{x \in Q} \Psi(x) := f(x) + h(x), \tag{12}$$

where $Q \subset \mathbb{R}^n$ is a compact convex set. We let $\|\cdot\|$ be a norm and write $\|\cdot\|_*$ the dual norm. We assume that the functions $f(x)$ and $h(x)$ are defined by

$$f(x) = \mathbf{E}[f(x,\xi)] \quad \text{and} \quad h(x) = \mathbf{E}[h(x,\xi)],$$

for some random variable $\xi \in \mathbb{R}^d$, we write $\Psi(x,\xi) := f(x,\xi) + h(x,\xi)$.

We also assume that $\Psi(\cdot,\xi)$ is convex for any $\xi \in \mathbb{R}^d$, that $\Psi(x,\xi_t) - \Psi(x) \geq 0$ a.s., and that the function $f(x)$ is convex with a Lipschitz continuous gradient

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|, \quad \text{for all } x, y \in Q,$$

and that $h(x)$ is a convex Lipschitz continuous function with

$$|h(x) - h(y)| \leq \mathcal{M}\|x - y\|, \quad \text{for all } x, y \in Q.$$

Furthermore, we assume that we observe a subgradient of Ψ through a stochastic oracle $G(x, \xi)$, satisfying

$$\mathbf{E}[G(x, \xi)] = g(x) \in \partial\Psi(x), \tag{13}$$

$$\mathbf{E}[\|G(x, \xi) - g(x)\|_*^2] \leq \sigma^2. \tag{14}$$

We let $\omega(x)$ be a distance generating function, i.e. a function such that

$$Q^o = \left\{ x \in Q : \exists y \in \mathbb{R}^p, \; x \in \operatorname*{argmin}_{u \in Q} [y^T u + \omega(u)] \right\}$$

is a convex set. We assume that $\omega(x)$ is strongly convex on Q^o with modulus α with respect to the norm $\|\cdot\|$, which means

$$(y - x)^T(\nabla\omega(y) - \nabla\omega(x)) \geq \alpha\|y - x\|^2, \quad x, y \in Q^o.$$

We then define a prox-function $V(x, y)$ on $Q^o \times Q$ as follows

$$V(x, y) \equiv \omega(y) - [\omega(x) + \nabla\omega(x)^T(y - x)], \tag{15}$$

which is nonnegative and strongly convex with modulus α with respect to the norm $\|\cdot\|$. The prox-mapping associated to V is then defined as

$$P_x^{Q,\omega}(y) \equiv \operatorname*{argmin}_{z \in Q} \{y^T(z - x) + V(x, z)\}. \tag{16}$$

This prox-mapping can be rewritten

$$P_x^{Q,\omega}(y) = \operatorname*{argmin}_{z \in Q} \{z^T(y - \nabla\omega(x)) + \omega(z)\},$$

and the strong convexity of $\omega(\cdot)$ means that $P_x^{Q,\omega}(\cdot)$ is Lipschitz continuous with respect to the norm $\|\cdot\|$ with modulus $1/\alpha$ (see Nemirovski [2004] or [Hiriart-Urruty and Lemaréchal, 1993, Vol. II, Th. 4.2.1]). Finally, we define the ω diameter of the set Q as

$$D_{\omega,Q} \equiv (\max_{z \in Q} \omega(z) - \min_{z \in Q} \omega(z))^{1/2}, \tag{17}$$

finally, we let

$$x^w = \operatorname*{argmin}_{x \in Q} w(x),$$

which satisfies

$$\frac{\alpha}{2}\|x - x^w\|^2 \leq V(x^w, x) \leq w(x) - x(x^w) \leq D_{w,Q}^2, \quad \text{for all } x \in Q.$$

4.1. **Stochastic composite optimization for semidefinite optimization.** We can use the results of Section 3 to derive explicit performance bounds on the algorithm in [Lan, 2009, §3] for problem (3). If we define the stochastic gradient oracle

$$G(X, z) = \frac{1}{q} \sum_{l=1}^{q} \phi_l \phi_l^T \tag{18}$$

where ϕ be a leading eigenvector of the matrix $X + \frac{\epsilon}{n} z_{i_0} z_{i_0}^T$ where

$$i_0 = \operatorname*{argmax}_{i=1,\ldots,k} \lambda_{\max}\left(X + \frac{\epsilon}{n} z_i z_i^T\right),$$

where z_i are i.i.d. Gaussian vectors $z_i \sim \mathcal{N}(0, \mathbf{I}_n)$ and $k > 0$ is a small constant (typically 3). [Lan, 2009, Corollary 1] implies the following result on the complexity of solving (3) using the AC-SA algorithm in [Lan, 2009, §3].

Proposition 4.1. *Let $N > 0$, and write f^* the optimal value of problem* (3)*. Suppose that the sequences X_t, X_t^{md}, X_t^{ag} are computed as in [Lan, 2009, Corollary 1] using the stochastic gradient oracle in* (18)*. After N iterations of the AC-SA algorithm in [Lan, 2009, §3], we have*

$$\mathbf{E}[f(X_{N+1}^{ag}) - f^*] \leq \frac{4n\sqrt{2}C_k D_{\omega,C}^2}{\epsilon N(N+2)} + \frac{4\sqrt{2}D_{\omega,C}}{\sqrt{Nq}} \tag{19}$$

Proof. Using the bound on the variance of the stochastic oracle $G(X,z)$, we know that G satisfies (13) with $\sigma^2 = 1/q$. Section 3 also shows that the Lipschitz constant of the gradient is bounded by $C_k n/\epsilon$. If we pick $\|\cdot\|_F^2/2$ as the prox function, [Lan, 2009, Corollary 1] yields the desired result. ■

Setting $N = 2D_Q\sqrt{n}/\epsilon$ and $q = \max\{1, D_Q/(\epsilon\sqrt{n})\}$ in the convergence bound above will then ensure $\mathbf{E}[f(X_N) - f^*] = O(\epsilon)$. In the section that follows, we detail a version of the AC-SA algorithm with adaptive (but monotonically decreasing) step-size scaling parameter.

4.2. **Stochastic composite optimization with line search.** The algorithm in [Lan, 2009, §3] uses worst case values of the Lipschitz constant L and of the gradient's quadratic variation σ^2 to determine step sizes at each iteration. In practice, this is a conservative strategy and slows down iterations in regions where the function is smoother. In the deterministic case, adaptive versions of the optimal first-order algorithm in [Nesterov, 1983] have been developed by Nesterov [2007b] among others. These algorithms run a few line search steps at each iterations to determine the optimal step size while guaranteeing convergence. The algorithm in [Lan, 2009] is a generalization of the first-order methods in [Nesterov, 1983, 2003] and, in what follows, we adapt the line search steps in Nesterov [2007b] to the stochastic algorithm of [Lan, 2009, §3]. Here, we will study the convergence properties of an adaptive variant of the algorithm for stochastic composite optimization in [Lan, 2009, §3].

Algorithm 1 Adaptive algorithm for stochastic composite optimization.

```
Input: An initial point $x^{ag} = x_1 = x^w \in \mathbb{R}^n$, an iteration counter $t = 1$, the number of iterations $N$, line
    search parameters $\gamma^{min}$, $\gamma^{max}$, $\gamma^d$, $\gamma > 0$, with $\gamma^d < 1$.
 1: Set $\gamma = \gamma^{max}$.
 2: for $t = 1$ to $N$ do
 3:   Define $x_t^{md} = \frac{2}{t+1}x_t + \frac{t-1}{t+1}x_t^{ag}$
 4:   Call the stochastic gradient oracle to get $G(x_t^{md}, \xi_t)$.
 5:   repeat
 6:     Set $\gamma_t = \frac{(t+1)\gamma}{2}$.
 7:     Compute the prox mapping $x_{t+1} = P_{x_t}(\gamma_t G(x_t^{md}, \xi_t))$.
 8:     Set $x_{t+1}^{ag} = \frac{2}{t+1}x_{t+1} + \frac{t-1}{t+1}x_t^{ag}$.
 9:   until $\Psi(x_{t+1}^{ag}, \xi_{t+1}) \leq \Psi(x_t^{md}, \xi_t) + \langle G(x_t^{md}, \xi_t), x_{t+1}^{ag} - x_t^{md} \rangle + \frac{\alpha\gamma^d}{4\gamma}\|x_{t+1}^{ag} - x_t^{md}\|^2 + 2\mathcal{M}\|x_{t+1}^{ag} -$
      $x_t^{md}\|$ or $\gamma \leq \gamma^{min}$. If exit condition fails, set $\gamma = \gamma\gamma^d$ and go back to step 5.
10:   Set $\gamma = \max\{\gamma^{min}, \gamma\}$.
11: end for
Output: A point $x_{N+1}^{ag}$.
```

In this section, we first modify the convergence lemma in [Lan, 2009, Lemma 5] to adapt it to the line search strategy detailed in Algorithm 1. Note that a particularity of our method is that testing the line search exit condition uses *two* oracle calls, the current one in ξ_t and the next one in ξ_{t+1}. This last oracle call is of course recycled at the next iteration.

Lemma 4.2. *Assume that $\Psi(\cdot, \xi_t)$ is convex for any given sample of the r.v. ξ_t. Let x_t, x_t^{md}, x_t^{ag} be computed as in Algorithm 1. Suppose also that γ and these points satisfy the line search exit condition in line 9, i.e.*

$$\Psi(x_{t+1}^{ag}, \xi_{t+1}) \leq \Psi(x_t^{md}, \xi_t) + \langle G(x_t^{md}, \xi_t), x_{t+1}^{ag} - x_t^{md}\rangle + \frac{\alpha}{4\gamma_t\beta_t}\|x_{t+1}^{ag} - x_t^{md}\|^2 + 2\mathcal{M}\|x_{t+1}^{ag} - x_t^{md}\|$$

then, for every x in the feasible set, we have

$$\begin{aligned}\beta_t\gamma_t[\Psi(x_{t+1}^{ag}, \xi_{t+1}) - \Psi(x)] + V(x_{t+1}, x) \quad &\leq \quad (\beta_t - 1)\gamma_t[\Psi(x_t^{ag}, \xi_t) - \Psi(x)] + V(x_t, x) + \frac{4\mathcal{M}^2\gamma_t^2}{\alpha} \\ &\quad +\gamma_t(\Psi(x, \xi_t) - \Psi(x))\end{aligned}$$

Proof. As in [Lan, 2009, Lemma 5], we write $d_t = x_{t+1} - x_t$ and use the parameter $\beta_t = (t+1)/2$ for step sizes so that $x_{t+1}^{ag} - x_t^{md} = d_t/\beta_t$. If the current iterates satisfy the line search exit condition, the fact that $\alpha\|d_t\|^2/2 \leq V(x_t, x_{t+1})$ by construction yields

$$\begin{aligned}\beta_t\gamma_t\Psi(x_{t+1}^{ag}, \xi_{t+1}) \quad &\leq \quad \beta_t\gamma_t[\Psi(x_t^{md}, \xi_t) + \langle G(x_t^{md}, \xi_t), x_{t+1}^{ag} - x_t^{md}\rangle] + \frac{\alpha}{4}\|d_t\|^2 + 2\gamma_t\mathcal{M}\|d_t\| \\ &\leq \quad \beta_t\gamma_t[\Psi(x_t^{md}, \xi_t) + \langle G(x_t^{md}, \xi_t), x_{t+1}^{ag} - x_t^{md}\rangle] + V(x_t, x_{t+1}) - \frac{\alpha}{4}\|d_t\|^2 + 2\gamma_t\mathcal{M}\|d_t\|.\end{aligned}$$

Using the convexity of $\Psi(\cdot, \xi_t)$ we then get

$$\begin{aligned}&\beta_t\gamma_t[\Psi(x_t^{md}, \xi_t) + \langle G(x_t^{md}, \xi_t), x_{t+1}^{ag} - x_t^{md}\rangle] \\ = \quad &(\beta_t - 1)\gamma_t[\Psi(x_t^{md}, \xi_t) + \langle G(x_t^{md}, \xi_t), x_t^{ag} - x_t^{md}\rangle] + \gamma_t[\Psi(x_t^{md}, \xi_t) + \langle G(x_t^{md}, \xi_t), x_{t+1} - x_t^{md}\rangle] \\ \leq \quad &(\beta_t - 1)\gamma_t\Psi(x_t^{ag}, \xi_t) + \gamma_t[\Psi(x_t^{md}, \xi_t) + \langle G(x_t^{md}, \xi_t), x_{t+1} - x_t^{md}\rangle].\end{aligned}$$

Combining these last two results and using the fact that $bu - au^2/2 \leq b^2/(2a)$ whenever $a > 0$, we obtain

$$\begin{aligned}\beta_t\gamma_t\Psi(x_{t+1}^{ag}, \xi_{t+1}) \quad &\leq \quad (\beta_t - 1)\gamma_t\Psi(x_t^{ag}, \xi_t) + \gamma_t[\Psi(x_t^{md}, \xi_t) + \langle G(x_t^{md}, \xi_t), x_{t+1} - x_t^{md}\rangle] \\ &\quad +V(x_t, x_{t+1}) - \frac{\alpha}{4}\|d_t\|^2 + 2\gamma_t\mathcal{M}\|d_t\| \\ &\leq \quad (\beta_t - 1)\gamma_t\Psi(x_t^{ag}, \xi_t) + \gamma_t[\Psi(x_t^{md}, \xi_t) + \langle G(x_t^{md}, \xi_t), x_{t+1} - x_t^{md}\rangle] \\ &\quad +V(x_t, x_{t+1}) + \frac{4\gamma_t^2\mathcal{M}^2}{\alpha}.\end{aligned}$$

For any x in the feasible set, we can then use the properties of the prox mapping detailed in [Lan, 2009, Lemma 1], with $p(\cdot) = \gamma_t\langle G(x_t^{md}, \xi_t), \cdot - x_t^{md}\rangle$ together with the convexity of $\Psi(\cdot, \xi_t)$ and the definition of x_{t+1} in Algorithm 1 to show that

$$\begin{aligned}&\gamma_t[\Psi(x_t^{md}, \xi_t) + \langle G(x_t^{md}, \xi_t), x_{t+1} - x_t^{md}\rangle] + V(x_t, x_{t+1}) \\ \leq \quad &\gamma_t\Psi(x_t^{md}, \xi_t) + \gamma_t\langle G(x_t^{md}, \xi_t), x - x_t^{md}\rangle + V(x_t, x) - V(x_{t+1}, x) \\ \leq \quad &\gamma_t\Psi(x, \xi_t) + V(x_t, x) - V(x_{t+1}, x),\end{aligned}$$

and combining these last results shows that

$$\beta_t\gamma_t\Psi(x_{t+1}^{ag}, \xi_{t+1}) \leq (\beta_t - 1)\gamma_t\Psi(x_t^{ag}, \xi_t) + \gamma_t\Psi(x, \xi_t) + V(x_t, x) - V(x_{t+1}, x) + \frac{4\gamma_t^2\mathcal{M}^2}{\alpha}$$

and subtracting $\beta_t\gamma_t\Psi(x)$ from both sides yields the desired result. ■

We are now ready to prove the main convergence result, adapted from [Lan, 2009, Corollary 1]. We simply stitch together the convergence results we obtained in Lemma 4.2 for the line search phase of the algorithm, with that of [Lan, 2009, Lemma 5] for the second phase where $\gamma = \gamma^{min}$. Note that the step size is still increasing in the second phase of the algorithm because $\gamma_t = \gamma^{min}(t+1)/2$.

Proposition 4.3. *Let $N > 0$, and write $\Psi(x^*)$ the optimal value of problem (12). Suppose that the sequences x_t, x_t^{md}, x_t^{ag} are computed as in Algorithm 1, with line search parameter γ initially set to $\gamma = \gamma^{max}$ with*

$$\gamma^{max} \leq \frac{\sqrt{6\alpha} D_{\omega,Q}}{(N+2)^{3/2}(4\mathcal{M}^2+\sigma^2)^{1/2}} \quad \textit{and} \quad \gamma^{min} = \min\left\{\frac{\alpha}{2L}, \gamma^{max}\right\} \tag{20}$$

with $\gamma^d < 1$. After N iterations of Algorithm 1, we have

$$\mathbf{E}[\Psi(x_{N+1}^{ag}) - \Psi^*] \leq \frac{8LD_{\omega,Q}^2}{\alpha N^2} + \frac{4}{N(N+2)\gamma^{min}} \mathbf{E}\left[\frac{2(4\mathcal{M}^2+\sigma^2)}{\alpha}\sum_{t=1}^{N}\gamma_t^2\right] \tag{21}$$

and a simpler, but coarser bound is given by

$$\mathbf{E}\left[\Psi(x_{N+1}^{ag}) - \Psi^*\right] \leq \frac{8LD_{\omega,Q}^2}{N^2} + \frac{4D_{\omega,Q}\sqrt{4\mathcal{M}^2+\sigma^2}}{\sqrt{N}}\left(\frac{\gamma^{max}}{\gamma^{min}}\rho(T_\gamma, N) + 1 - \rho(T_\gamma, N)\right), \tag{22}$$

where $\rho(T_\gamma, N) = (T_\gamma + 2)^3/(N+2)^3$.

Proof. Lemma 4.2 applied at x^* shows

$$\begin{aligned}\beta_t\gamma_t[\Psi(x_{t+1}^{ag}, \xi_{t+1}) - \Psi(x^*)] + V(x_{t+1}, x^*) \quad &\leq \quad (\beta_t - 1)\gamma_t[\Psi(x_t^{ag}, \xi_t) - \Psi(x^*)] + V(x_t, x^*) + \frac{4\mathcal{M}^2\gamma_t^2}{\alpha} \\ &\quad\quad +\gamma_t(\Psi(x^*, \xi_t) - \Psi(x^*))\end{aligned}$$

hence, having assumed $\Psi(x, \xi_t) - \Psi(x) \geq 0$ a.s.,

$$\begin{aligned}(\beta_{t+1} - 1)\gamma_t[\Psi(x_{t+1}^{ag}, \xi_{t+1}) - \Psi(x^*)] \quad &\leq \quad \beta_t\gamma_t[\Psi(x_{t+1}^{ag}, \xi_t) - \Psi(x^*)] \\ &\leq \quad (\beta_t - 1)\gamma_t[\Psi(x_t^{ag}, \xi_{t+1}) - \Psi(x^*)] + \frac{4\mathcal{M}^2\gamma_t^2}{\alpha} \\ &\quad\quad +\gamma_t(\Psi(x^*, \xi_t) - \Psi(x^*)) + V(x_t, x) - V(x_{t+1}, x)\end{aligned}$$

whenever the line search successfully terminates, with the last term satisfying

$$\mathbf{E}[\gamma_t(\Psi(x^*, \xi_t) - \Psi(x^*))] \leq \frac{\gamma^{max}(t+1)}{2}\mathbf{E}[\Psi(x^*, \xi_t) - \Psi(x^*)] = 0$$

using again $\Psi(x^*, \xi_t) - \Psi(x^*) \geq 0$ a.s.. When the line search fails $\gamma_t = \gamma^{min}(t+1)/2$ is deterministic and [Lan, 2009, Lem. 5 & Th. 2] show that

$$(\beta_{t+1} - 1)\gamma_t[\Psi(x_{t+1}^{ag}) - \Psi(x^*)] \leq (\beta_t - 1)\gamma_t[\Psi(x_t^{ag}) - \Psi(x^*)] + V(x_t, x^*) - V(x_{t+1}, x^*) + \Delta(x^*)$$

where

$$\Delta(x^*) \leq \gamma_t\langle\delta_t, x^* - x_t\rangle + \frac{2(4\mathcal{M}^2 + \|\delta_t\|_*^2)\gamma_t^2}{\alpha}.$$

with $\delta_t = G(x_t^{md}, \xi_t) - g(x_t^{md})$ and $\gamma_t \langle \delta_t, x^* - x_t \rangle \leq \gamma_t \|\delta_t\|_* \|x^* - x_t\|$. We call $t = T_\gamma$ the iteration where the line search first fails. Combining these last results, using $\beta_1 = 1$, we obtain

$$\begin{aligned}
(\beta_{N+1} - 1)\gamma_N \, \mathbf{E}[\Psi(x_{N+1}^{ag}) - \Psi(x^*)] \quad &\leq \quad D_{\omega,Q}^2 + \sum_{t=1}^{T_\gamma} \mathbf{E}\left[\frac{4\mathcal{M}^2\gamma_t^2}{\alpha}\right] \\
&\quad\; +(\beta_{T_\gamma} - 1)\gamma_{T_\gamma} \, \mathbf{E}[\Psi(x_{T_\gamma}^{ag}) - \Psi(x_{T_\gamma}^{ag}, \xi_t)] \\
&\quad\; + \sum_{T_\gamma}^{N} \mathbf{E}\left[\gamma_t \langle \delta_t, x^* - x_t \rangle + \frac{2(4\mathcal{M}^2 + \|\delta_t\|_*^2)\gamma_t^2}{\alpha}\right] \\
&\leq \quad D_{\omega,Q}^2 + \sum_{t=1}^{T_\gamma} \mathbf{E}\left[\frac{4\mathcal{M}^2\gamma_t^2}{\alpha}\right] + \sum_{T_\gamma}^{N} \mathbf{E}\left[\frac{2(4\mathcal{M}^2 + \|\delta_t\|_*^2)\gamma_t^2}{\alpha}\right] \\
&\leq \quad D_{\omega,Q}^2 + \mathbf{E}\left[\frac{2(4\mathcal{M}^2 + \sigma^2)}{\alpha} \sum_{t=1}^{N} \gamma_t^2\right]
\end{aligned}$$

because $\mathbf{E}[\Psi(x_{T_\gamma}^{ag}) - \Psi(x_{T_\gamma}^{ag}, \xi_t)] = 0$. Using the fact that $\sum_{t=1}^{N}(t+1)^q \leq (N+q)^{q+1}/(q+1)$ for $q = 1, 2$ then yields the coarser bound. ■

We observe that, as in [Nesterov, 2007b], allowing a line search slightly increases the complexity bound, by a factor

$$\left(\frac{\gamma^{max}}{\gamma^{min}}\rho(T_\gamma, N) + 1 - \rho(T_\gamma, N)\right),$$

where $\rho(T_\gamma, N) = (T_\gamma + 2)^3/(N+2)^3$. We will see however that overall numerical performance can significantly improve because the algorithm takes longer steps.

5. Extensions

In this section, we discuss possible extensions of the stochastic regularization techniques, their efficiency and regularity.

5.1. GUE smoothing.

We have chosen to analyze the rank one perturbation because of its numerical efficiency and mathematical simplicity. However, many other random smoothing algorithms are possible and modern random matrix theory offers tools to understand their properties. We expect that some of them will lead to better worst case bounds than the rank one perturbation methods we have considered here.

A case in point is the following. Consider a matrix U from the Gaussian Unitary Ensemble (GUE). Matrices from GUE are Hermitian random matrices with complex Gaussian entries, i.i.d $\mathcal{N}_\mathbb{C}(0,1)$ above the diagonal and i.i.d $\mathcal{N}(0,1)$ on the diagonal. Recall that if $\mathsf{z}_\mathbb{C}$ is $\mathcal{N}_\mathbb{C}(0,1)$, $\mathsf{z}_\mathbb{C} = (\mathsf{z}_1 + i\mathsf{z}_2)/\sqrt{2}$, where z_1 and z_2 are independent with distribution $\mathcal{N}(0,1)$.

In what follows, X is a deterministic matrix and U is a random GUE matrix. We assume, without loss of generality, that the largest eigenvalue of X is bounded (if not, we can always shift X by a multiple of $\mathbf{I}_n$, which takes care of the problem).

A natural smoothing of $\lambda_{\max}(X)$ is $f_{GUE}(X) = \mathbf{E}[\lambda_{\max}(X + (\epsilon/\sqrt{n})U)]$, where U is a GUE matrix. This type of matrices belong to the so-called "deformed GUE". Johansson [2007] is an important paper in this area and contains a result, Theorem 1.12, that is not exactly suited to our problem but quite close, perhaps despite the appearances. Before we proceed, we note that showing that $f_{GUE}(X)$ is an ϵ-approximation of $\lambda_{\max}(X)$ is immediate from standard results on GUE matrices (see Trotter [1984], Davidson and Szarek [2001]).

In a nutshell, random matrix theory indicates that $\lambda_{\max}(X + (\epsilon/\sqrt{n})U)$ undergoes a phase transition as ϵ changes when X is not a multiple of $\mathbf{I}_n$. If ϵ is sufficiently large (more details follow), the behavior of

$\lambda_{\max}(X+(\epsilon/\sqrt{n})U)$ is driven by the GUE component and the spacing between the two largest eigenvalues is of order $n^{-2/3}$. On the other hand, if ϵ is not large enough, we remain essentially in a perturbative regime and the spacing between the two largest eigenvalues is larger than $n^{-2/3}$. A very detailed study of the phase transition should be possible, too. However, all these results are asymptotic. Non-asymptotic results could be obtained (the machinery to obtain results such as Johansson's is non-asymptotic) but would be hard to interpret and exploit. We therefore keep this discussion at an informal level.

Smoothing by a GUE matrix should therefore give a worst case bound on $\|\nabla f\|_L$ of order $n^{2/3}$, which is better than the worst case bound of n we have when we smooth with rank one matrices (but requires generating $O(n^2)$ random numbers instead of $O(n)$). GUE smoothing might therefore improve the performance of the algorithm since the cost of generating these random variables is typically dominated by the cost of computing a leading eigenvector of the perturbed matrix.

Let us give a bit more quantitative details. Based on Johansson's work and the solution to a similar problem in a different context (El Karoui [2007]), it is clear that the condition for the spacings to be of order $n^{-2/3}$ is the following (this result might be available in the literature but we have not found a reference). Call F_n the spectral distribution of X, i.e the probability distribution that puts mass $1/n$ at each of the n eigenvalues of X. Call w_c the solution in $(\lambda_{\max}(X),\infty)$ of

$$\int \frac{dF_n(t)}{(w_c-t)^2} = \frac{1}{\epsilon^2} \ .$$

Call $\mathcal{G}$ the class of matrices for which

$$\liminf_{n\to\infty} \left[w_c - \lambda_{\max}(X)\right] > 0 \ .$$

Then, looking carefully at Johansson's work, it should be possible to show that: if the sequence of matrices X is in $\mathcal{G}$, then, if $X(\epsilon) = X + \epsilon/\sqrt{n}U$,

$$n^{2/3}\frac{\lambda_{\max}(X(\epsilon)) - \alpha_n}{\beta_n} \Longrightarrow \mathrm{TW}_2 \ ,$$

where

$$\alpha_n = w_c + \epsilon^2 \int \frac{dF_n(t)}{w_c - t} \quad \text{and} \quad \beta_n = \epsilon^2 \left(\int \frac{dF_n(t)}{(w_c-t)^3}\right)^{1/3}$$

and TW_2 is the Tracy-Widom distribution appearing in the study of GUE [see Tracy and Widom, 1994]. The same is true for the joint distribution of the k largest eigenvalues, where k is a fixed integer, and TW_2 is replaced by the corresponding limiting joint distribution for the k largest eigenvalues of a GUE matrix.

When the matrix X is not in $\mathcal{G}$, then the top two eigenvalues should have spacing greater than $n^{-2/3}$. We expect that if X has some sufficiently separated eigenvalues with multiplicity higher than one, the spacings there are at least $n^{-1/2}$, by analogy with Capitaine et al. [2009] and Baik et al. [2005]. To quantify what "sufficiently separated" means, we could suppose that X is a completion of a $(n-k_0)\times(n-k_0)$ matrix X_0 which is in $\mathcal{G}$, to which we add k_0 eigenvalues $\lambda_{\max}(X)$, all equal and greater than $\lambda_{\max}(X_0)$, with $\lambda_{\max}(X)$ greater than and bounded away from $w_c(X_0)$. Call $F_{n-k_0,0}$ the spectral distribution of X_0. Then, we should have

$$n^{1/2}\frac{\lambda_{\max}(X(\epsilon)) - \widetilde{\alpha}_n}{\widetilde{\beta}_n} \Longrightarrow \lambda_{\max}\left(\mathrm{GUE}_{k_0\times k_0}\right) \ ,$$

where $\widetilde{\alpha}_n = \lambda_{\max}(X) + \epsilon^2 \int \frac{dF_{n-k_0,0}(t)}{\lambda_{\max}(X)-t}$ and $\widetilde{\beta}_n = \epsilon\left(1 - \epsilon^2 \int \frac{dF_{n-k_0,0}(t)}{(\lambda_{\max}(X)-t)^2}\right)^{1/2}$.

The same is true for the k_0 largest eigenvalues of $X(\epsilon)$ and $\lambda_{\max}(\mathrm{GUE}_{k_0\times k_0})$ is replaced by the corresponding joint distribution for the $k_0 \times k_0$ GUE.

In light of the integrability problems we had in the rank one perturbation case for the inverse spectral gap $1/(l_1(X(\epsilon)) - l_2(X(\epsilon)))$, it is natural to ask whether such problems would arise with a GUE smoothing.

For this informal discussion, we limit ourselves to answering this question for the GUE. We recall that the joint density of the eigenvalues $\{l_{i,GUE}\}_{i=1}^n$ of a $n \times n$ GUE matrix is

$$C \exp(-\sum_{i=1}^n l_{i,GUE}^2/2) \prod_{1 \leq i < j \leq n} |l_{i,GUE} - l_{j,GUE}|^2 \ ,$$

where C is a normalizing constant. So we see immediately that $1/(l_{1,GUE} - l_{2,GUE})$ is integrable in the GUE setting. (The formula above is often stated for the unordered eigenvalues of a GUE matrix. The functional form of the density is unchanged by ordering, because of the symmetry. The domain of definition and the constant change when considering ordered eigenvalues, but this has no bearing on the question of integrability.)

The smoothing could also be done by a matrix from the Gaussian Orthogonal Ensemble (GOE), where the entries above the diagonal are i.i.d $\mathcal{N}(0,1)$ and the entries on the diagonal are i.i.d $\mathcal{N}(0,2)$ - the different normalization on and off the diagonal yields rotational invariance. We do not know of a result corresponding to Johansson's in that case, though we would expect that the behavior of the top eigenvalues is the same as described above, with TW_2 replace by TW_1, the Tracy-Widom distribution appearing in the study of GOE. From an algorithmic point of view, the two methods should therefore be equivalent.

6. Numerical Experiments

We test the algorithm detailed above on a maximum eigenvalue minimization problem over a hypercube, a problem used in approximating sparse eigenvectors [d'Aspremont et al., 2007]. We seek to solve

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}(A+X) \\ \text{subject to} & -\rho \leq X_{ij} \leq \rho, \quad \text{for } i,j = 1,\ldots,n \end{array} \tag{23}$$

which is a semidefinite program in the matrix $X \in \mathbf{S}_n$. Since randomly generated matrices A have highly structured spectrum, we use a covariance matrix from the gene expression data set in [Alon et al., 1999] to generate the matrix $A \in \mathbf{S}_n$, varying the number of genes to vary the problem dimension n (we select the n genes with the highest variance). We set $\rho = \max\{\mathbf{diag}(A)\}/2$ in (23).

We first compare the performance of Algorithm 1 with that of the corresponding deterministic algorithm detailed in [Nesterov, 2007a,b], using the accelerated first-order method in [Nesterov, 2007b, §4] after smoothing problem (23) as in [Nesterov, 2007a; d'Aspremont et al., 2007]. We set a fixed number of outer iterations for Algorithm 1 and record the number of iterations (and eigenvector evaluations, these numbers differ because of line search steps) required by the algorithm in [Nesterov, 2007b, §4] to reach the best objective value attained by the stochastic method. We set $q = 5$, $k = 3$ and the maximum number of iterations to $20\sqrt{n}$ in the stochastic algorithm. To provide a complexity benchmark that is both hardware and implementation independent, we record the total number of eigenvectors used by each algorithm to reach a given objective value (the matrix exponential thus counts as n eigenvectors). We report these results in Table 2.

n	# Iters. (Stoch.)	# Eigvs. (Stoch.)	# Iters. (Det.)	# Eigvs. (Det.)
100	200	6120	100	40400
200	283	8565	100	81200
500	447	13470	100	203000

TABLE 2. Number of iterations and total number of eigenvectors computed by Algorithm 1 (Stoch.) and the algorithm in [Nesterov, 2007b, §4] (Det.) to reach identical objective values.

In both algorithms, the cost of each iteration is dominated by that of computing gradients. The cost of each gradient computation in Algorithm 1 is dominated by the cost of computing the leading eigenvector of q

perturbed matrices. The cost of each gradient computation in [Nesterov, 2007b, §4] is dominated by the cost of computing a matrix exponential. This means that the ratio between these costs grows as $O(n/(q \log n))$.

In Figure 2 we plot the sequence of line search parameters γ for the stochastic algorithm together with the values of the Lipschitz constant L used in the deterministic smoothing algorithm, when solving problem (23) with $n = 500$. We observe that both algorithms initially make longer steps, then slow down as they get closer to the optimum (where the leading eigenvalues are clustered).

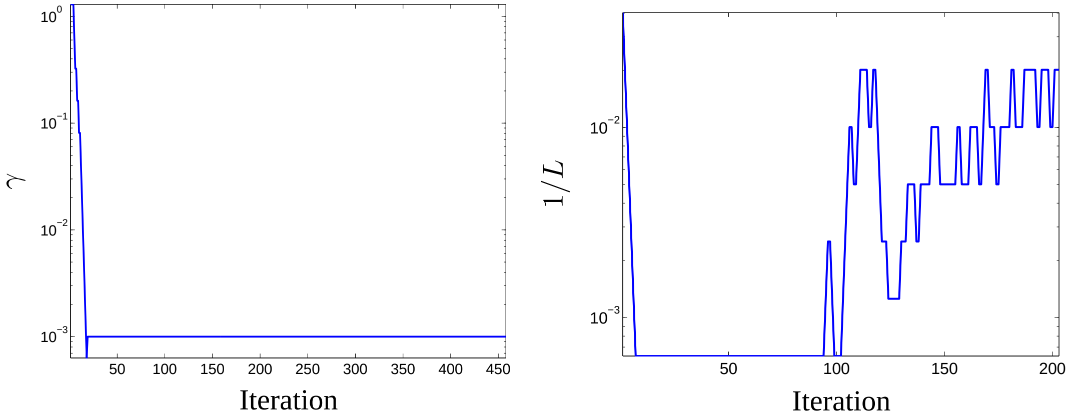

FIGURE 2. Line search parameters γ for the stochastic algorithm (left) together with the values of the inverse of the Lipschitz constant L used in the deterministic smoothing algorithm (right).

7. APPENDIX

In this appendix, we recall several useful results related to the algorithm presented here. The first summarizes the complexity of computing *one* leading eigenvector of a symmetric matrix (versus computing the entire spectrum). We then prove Theorem 3.2 linking the local Lipschitz constant of the gradient and the spectral gap. Finally, we detail the proof of the phase transition result in Theorem 3.8 and show how the secular equation can be generalized to perturbations of higher rank.

7.1. **Computing one leading eigenvector of a symmetric matrix.**

The complexity results detailed above heavily rely on the fact that extracting *one* leading eigenvector of a symmetric matrix $X \in \mathbf{S}_n$ can be done by computing a few matrix vector products. This simple fact is easy to prove using the power method when the eigenvalues of X are well separated, and Krylov subspace methods making full use of the matrix vector products converge even faster. However, the problem becomes more delicate when the spectrum of X is clustered. The section that follows briefly summarizes how modern numerical methods produce eigenvalue decompositions in practice.

We start by recalling how packages such as LAPACK Anderson et al. [1999] form a full eigenvalue (or Schur) decomposition of a symmetric matrix $X \in \mathbf{S}_n$. The algorithm is strikingly stable and, despite its $O(n^3)$ complexity, often competitive with more advanced techniques when the matrix X is small. We then discuss the problem of approximating one leading eigenpair of X using Krylov subspace methods with complexity growing as $O(n^2 \log n)$ with the dimension (or less when the matrix is structured). In practice, we will see that the constants in these bounds differ significantly, with the cost of a full eigenvalue decompositions (and matrix exponentials) growing as $4n^3/3$ while computing one leading eigenpair has cost cn^2, with c in the hundreds.

7.1.1. *Full eigenvalue decomposition.* Full eigenvalue decompositions are computed by first reducing the matrix X to symmetric tridiagonal form using Householder transformations, then diagonalizing the tridiagonal factor using iterative techniques such as the QR or divide and conquer methods for example (see [Stewart, 2001, Chap. 3] for an overview). The classical QR algorithm (see [Golub and Van Loan, 1990, §8.3]) implicitly relied on power iterations to compute the eigenvalues and eigenvectors of a symmetric tridiagonal matrix with complexity $O(n^3)$, however more recent methods such as the MRRR algorithm by Dhillon and Parlett [2003] solve this problem with complexity $O(n^2)$. Starting with the third version of LAPACK, the MRRR method is part of the default routine for diagonalizing a symmetric matrix and is implemented in the `STEGR` driver (see Dhillon et al. [2006]).

Overall, the complexity of forming a *full* Schur decomposition of a symmetric matrix $X \in \mathbf{S}_n$ is then $4n^3/3$ flops for the Householder tridiagonalization, followed by $O(n^2)$ flops for the Schur decomposition of the tridiagonal matrix using the MRRR algorithm.

7.1.2. *Computing one leading eigenpair.* We now give a brief overview of the complexity of computing leading eigenpairs using Krylov subspace methods and we refer the reader to [Stewart, 2001, §4.3], [Golub and Van Loan, 1990, §8.3, §9.1.1] or Saad [1992] for a more complete discussion. Successful termination of a *deterministic* power or Krylov method can never be guaranteed since in the extreme case where the starting vector is orthogonal to the leading eigenspace, the Krylov subspace contains no information about leading eigenpairs, so the results that follow are stochastic. [Kuczynski and Wozniakowski, 1992, Th.4.2] show that, for any matrix $X \in \mathbf{S}_n$ (including matrices with clustered spectrum), starting the algorithm at a random u_1 picked uniformly over the sphere means the Lanczos decomposition will produce a leading eigenpair with *relative* precision ϵ in

$$k^{\mathrm{Lan}} \leq \frac{\log(n/\delta^2)}{4\sqrt{\epsilon}}$$

iterations, with probability at least $1-\delta$. This is of course a highly conservative bound and in particular, the worst case matrices used to prove it vary with k^{Lan}.

This means that computing one leading eigenpair of the matrix X requires computing at most k^{Lan} matrix vector products (we can always restart the code in case of failure) plus $4nk^{\mathrm{Lan}}$ flops. When the matrix is dense, each matrix vector product costs n^2 flops, hence the total cost of computing one leading eigenpair of X is

$$O\left(\frac{n^2\log(n/\delta^2)}{4\sqrt{\epsilon}}\right)$$

flops. When the matrix is sparse, the cost of each matrix vector product is $O(s)$ instead of $O(n^2)$, where s is the number of nonzero coefficients in X. Idem when the matrix X has rank $r < n$ and an explicit factorization is known, in which case each matrix vector product costs $O(nr)$ which is the cost of two $n \times r$ matrix vector products, and the complexity of the Lanczos procedure decreases accordingly.

The numerical package ARPACK by Lehoucq et al. [1998] implements the Lanczos procedure with a reverse communication interface allowing the user to efficiently compute the matrix vector product Xu_j. However, it uses the implicitly shifted QR method instead of the more efficient MRRR algorithm to compute the Ritz pairs of the matrix $T_k \in \mathbf{S}_k$.

7.2. **Controlling the Hessian of $\lambda_{\max}(X)$.** Consider the map $f_0 : \mathbf{S}_n \to \mathbb{R}$ such that $f_0(X) = \lambda_{\max}(X)$. We want to show that its gradient is Lipschitz continuous, when the largest eigenvalue of X has multiplicity one and control its constant. To do so, we compute $\partial^2 f_0(X+tY)/\partial t^2$, where $\|Y\|_F = 1$, and Y is symmetric. Let us call $\lambda_1 > \lambda_2 \geq \lambda_3 \geq \lambda_n$ the ordered eigenvalues of X. Very importantly we assume that λ_1 has multiplicity one. If not, it is easy to see that the map we are looking at is continuous but not Lipschitz. We refer the reader to [Kato, 1995; Overton and Womersley, 1995; Lewis and Sendov, 2002] for a more complete discussion. We have the following theorem.

Theorem 7.1. *Suppose X is an $n \times n$ symmetric matrix with decreasingly ordered eigenvalues $\{\lambda_i\}_{i=1}^n$. Call $f_0(X) = \lambda_{\max}(X)$ and suppose that $\lambda_{\max}(X)$ has multiplicity one. Let Y be a symmetric matrix with $\|Y\|_F = 1$. Let us call*

$$g(X,Y) = \lim_{t\to 0} \frac{\partial^2 f_0(X+tY)}{\partial t^2} \ .$$

Then we have

$$\|\nabla f_0(X)\|_L = \sup_{Y\in \mathbf{S}_n, \|Y\|_F=1} g(X,Y) = \frac{1}{2}\frac{1}{\lambda_1(X)-\lambda_2(X)} \ . \tag{24}$$

Proof. The strategy is to first exhibit a matrix Y_c in $\mathbf{S}_n$ that will give us the right-hand side of Equation (24) as a lower bound. And then we will show that indeed this bound is the best one can do. We will rely heavily on the following classical result from the analytic perturbation theory of matrices. We can use [Kato, 1995, p.81] to get

$$\lim_{t\to 0} \frac{\partial^2 f_0(X+tY)}{\partial t^2} = \sum_{j=2}^n \frac{1}{\lambda_1(X)-\lambda_j(X)} (\phi_1^T Y \phi_j)^2 \ , \tag{25}$$

where ϕ_1 is an eigenvector corresponding to the eigenvalue λ_1 and ϕ_j is an eigenvector corresponding to the eigenvalue λ_j. Here we have crucially used the fact that $\lambda_1(X)$ has multiplicity one.

Finding a lower bound for $\|\nabla f_0(X)\|_L$. Let O be an orthonormal matrix that transforms the canonical basis $(e_1,\ldots,e_n)$ into the orthonormal basis $(\phi_1,\ldots,\phi_n)$. In other words, $Oe_i = \phi_i$ and hence $O^T\phi_i = e_i$. Let us call P_0 the matrix that exchanges e_1 and e_2 and send the other e_j's to 0. In other words, the 2×2 upper left block of P_0 is the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and P_0 is zero everywhere else. Now call

$$Y_c = \frac{1}{\sqrt{2}} O P_0 O^T \ .$$

Note that $Y_c \in \mathbf{S}_n$. Since $O^T\phi_i = e_i$, we see that $Y_c\phi_1 = \phi_2/\sqrt{2}$, $Y_c\phi_2 = \phi_1/\sqrt{2}$, and $Y_c\phi_j = 0$ if $j > 2$. Further, $\|Y_c\|_F^2 = \mathbf{Tr}\, Y_c^T Y_c = \mathbf{Tr}\, Y_c^2 = \mathbf{Tr}\, O P_0^2 O^T/2 = \|P_0\|_F^2/2 = 1$. Now, $\phi_1^T Y_c \phi_j = \delta_{2,j}\|\phi_1\|^2/\sqrt{2}$. Hence,

$$g(X,Y_c) = \frac{1}{2}\frac{1}{\lambda_1(X)-\lambda_2(X)} \ ,$$

and therefore,

$$\|\nabla f_0(X)\|_L \geq \frac{1}{2}\frac{1}{\lambda_1(X)-\lambda_2(X)} \ .$$

Finding an upper bound for $\|\nabla f_0(X)\|_L$. On the other hand, we clearly have, for $j \geq 2$, $0 \leq 1/(\lambda_1(X) - \lambda_j(X)) \leq 1/(\lambda_1(X)-\lambda_2(X))$. Therefore,

$$\sum_{j=2}^n \frac{1}{\lambda_1(X)-\lambda_j(X)} (\phi_1^T Y \phi_j)^2 \leq \frac{1}{\lambda_1(X)-\lambda_2(X)} \sum_{j=2}^n (\phi_1^T Y \phi_j)^2 \ .$$

Since $\{\phi_j\}_{j=1}^n$ form an orthonormal basis, and Y is symmetric,

$$\sum_{j=1}^n (\phi_1^T Y \phi_j)^2 = \|Y\phi_1\|_2^2 \ .$$

As a matter of fact $\phi_1^T Y \phi_j$ is just the coefficient of the vector $Y^T\phi_1 = Y\phi_1$ in its representation in the basis of the ϕ_i's. We therefore have

$$\sum_{j=2}^n \frac{1}{\lambda_1(X)-\lambda_j(X)} (\phi_1^T Y \phi_j)^2 \leq \frac{1}{\lambda_1(X)-\lambda_2(X)} \left(\|Y\phi_1\|_2^2 - (\phi_1^T Y \phi_1)^2\right) \ .$$

Now let us call $\tilde{y}_{i,j}$ the (i,j)-th entry of the matrix that represents Y in the basis of the ϕ_i's. Since $\|Y\|_F^2 = 1$,

$$\sum_{i,j} \tilde{y}_{i,j}^2 = 1 \ .$$

Using the symmetry of Y, we therefore see that

$$2 \sum_{j=2}^{n} \tilde{y}_{1,j}^2 + \tilde{y}_{1,1}^2 \leq 1 \ .$$

Now, $\|Y\phi_1\|_2^2 = \sum_{j=1}^n \tilde{y}_{1,j}^2$ and $(\phi_1^T Y \phi_1)^2 = \tilde{y}_{1,1}^2$. Hence,

$$\left(\|Y\phi_1\|_2^2 - (\phi_1^T Y \phi_1)^2\right) = \sum_{j=2}^{n} \tilde{y}_{1,j}^2 \leq \frac{1 - \tilde{y}_{1,1}^2}{2} \leq \frac{1}{2} \ .$$

We conclude that

$$\forall Y \in \mathbf{S}_n, \ \|Y\|_F = 1 \ , \ \ g(X,Y) \leq \frac{1}{2} \frac{1}{\lambda_1(X) - \lambda_2(X)} \ ,$$

and therefore

$$\|\nabla f_0(X)\|_L = \sup_{Y \in \mathbf{S}_n, \|Y\|_F = 1} g(X,Y) \leq \frac{1}{2} \frac{1}{\lambda_1(X) - \lambda_2(X)} \ .$$

Since we have matching upper and lower bounds for $\|\nabla f_0(X)\|_L$, we have established the theorem. ■

7.3. **Phase transition.** We prove Theorem 3.8 in this section.

7.3.1. *Preliminaries.* Let us call

$$\begin{aligned} g_l(t) &= \frac{1}{n} \sum_{j=l+1}^{n} \frac{1}{t + \gamma + \delta_j} \ , \\ h_l(t) &= \frac{1}{n} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2}{t + \gamma + \delta_j} \ , \\ g(t) &= \frac{\sum_{j=1}^{l} \mathsf{z}_j^2}{n} \frac{1}{t} + \frac{1}{n} \sum_{j=l+1}^{n} \frac{1}{t + \gamma + \delta_j} = \frac{\sum_{j=1}^{l} \mathsf{z}_j^2}{n} \frac{1}{t} + g_l(t) \ , \\ h(t) &= \frac{\sum_{j=1}^{l} \mathsf{z}_j^2}{n} \frac{1}{t} + h_l(t) \ . \end{aligned}$$

Recall that $l_1 = \lambda_1 + T$ is the root of the equation

$$\frac{1}{\epsilon} = h(T) = \frac{\sum_{j=1}^{l} \mathsf{z}_j^2}{n} \frac{1}{T} + \frac{1}{n} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2}{T + \gamma + \delta_j} \ . \tag{26}$$

It is clear that $T \geq (\epsilon/n) \sum_{j=1}^{l} \mathsf{z}_j^2$. Also, $h'(t) < 0$ on $(0, \infty)$, so h is invertible. We note that

$$\mathbf{var}\left(\frac{1}{n} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2 - 1}{t + \gamma + \delta_j} \right) = \frac{1}{n} \left[\frac{1}{n} \sum_{j=l+1}^{n} \frac{2}{(t + \gamma + \delta_j)^2} \right] \leq \frac{1}{n} \frac{1}{\gamma^2} = O\left(\frac{1}{n}\right) \ .$$

So the error made when replacing h_l by g_l when seeking the root of Equation (26) is $O_P(1/\sqrt{n})$.

Our strategy is to expand T in powers (possibly non-integer) of $1/n$. We call $t(m)$ an approximation of T to order m. If we can find an approximate solution $t(m)$, such that

$$|h(t(m)) - \frac{1}{\epsilon}| = O_P(n^{-\beta}) \text{ , for some } \beta \text{ ,}$$

we claim that

$$|t(m) - T| = O_P(n^{-\beta}) \text{ .}$$

This is because h is, at z_i fixed, a Lipschitz function on $(\frac{\epsilon \sum_{j=1}^{l} \mathsf{z}_j^2}{n}, \infty)$, and its Lipschitz constant is bounded below with high-probability on any compact interval of this interval. Hence, we have

$$|t(m) - T| = |h^{-1}(h(t(m))) - h^{-1}(h(T))| \leq \|h^{-1}\|_{\mathrm{L}} |h(t(m)) - \frac{1}{\epsilon}| = O_P(n^{-\beta}) \text{ .}$$

Note that if we can show that $|h'(y)| > Cn^b$ in a neighborhood of $t(m)$, then we get by the same token

$$|h(t(m)) - \frac{1}{\epsilon}| = O_P(n^{-\beta}) \Longrightarrow |t(m) - T| = O_P(n^{-(\beta+b)}) \text{ .}$$

We finally recall that

$$\frac{1}{\epsilon_0} = g_l(0) = \frac{1}{n} \sum_{j=l+1}^{n} \frac{1}{\gamma + \delta_j} \text{ .}$$

7.3.2. *Case* $\epsilon < \epsilon_0$. Recall that the equation defining T is

$$\frac{1}{\epsilon} = h(T) = \frac{\sum_{j=1}^{l} \mathsf{z}_j^2}{n} \frac{1}{T} + \frac{1}{n} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2}{T + \gamma + \delta_j} = \frac{\sum_{j=1}^{l} \mathsf{z}_j^2}{n} \frac{1}{T} + h_l(T) \text{ .}$$

In this case,

$$g_l(0) = \frac{1}{\epsilon_0} < \frac{1}{\epsilon} \text{ ,}$$

so it is clear that the term $\frac{\sum_{j=1}^{l} \mathsf{z}_j^2}{nT}$ needs to enter into play to "saturate" the equality. In particular, T is going to be of order $1/n$. But we can expand it further.

Let us now expand the last term above, i.e $h_l(t)$, in powers of t's. Because h_l is uniformly bounded in probability for t in a neighborhood of 0, we have

$$h_l(t) = \frac{1}{n} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2}{\gamma + \delta_j} - t \frac{1}{n} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2}{(\gamma + \delta_j)^2} + t^2 \frac{1}{n} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2}{(\gamma + \delta_j)^3} + O_P(t^3) \text{ .}$$

So calling $\zeta_1 = \frac{1}{n} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2}{(\gamma+\delta_j)^2}$, and $\zeta_2 = \frac{1}{n} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2}{(\gamma+\delta_j)^3}$ the equation defining T becomes

$$\frac{1}{\epsilon} = \frac{\chi_l^2}{n} \frac{1}{T} + \frac{1}{\epsilon_0} + \frac{1}{n} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2 - 1}{\gamma + \delta_j} - T\zeta_1 + T^2 \zeta_2 + O(T^3) \text{ .}$$

We see that by taking

$$t(2) = \frac{\alpha_1}{n} + \frac{\alpha_1^2}{n^{3/2}} \frac{1}{\sqrt{n}} \sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2 - 1}{\gamma + \delta_j} \text{ ,}$$

with

$$\alpha_1 = \frac{\chi_l^2}{\frac{1}{\epsilon} - \frac{1}{\epsilon_0}} \text{ ,}$$

we have

$$h(t(2)) - h(T) = O_P(1/n) \text{ .}$$

Now, we note that in a neighborhood of $1/n$, the derivative of h is bounded below in absolute value and in probability by $O_P(n)$. Our argument in the previous subsection therefore allows us to conclude that

$$T - t(2) = O_P\left(\frac{1}{n^2}\right) .$$

7.3.3. *Case* $\epsilon = \epsilon_0$. We can use the same expansion in t as above, but Equation (26) defining T becomes

$$\frac{1}{\epsilon} = \frac{\chi_l^2}{n}\frac{1}{T} + \frac{1}{\epsilon} + \frac{1}{n}\sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2 - 1}{\gamma + \delta_j} - T\zeta_1 + O(T^2) ,$$

where, as above,

$$\zeta_1 = \frac{1}{n}\sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2}{(\gamma + \delta_j)^2} .$$

Because $\xi_1 = \frac{1}{\sqrt{n}}\sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2 - 1}{\gamma + \delta_j} = O_P(1)$, we see that now, T is of order $1/\sqrt{n}$. Using the ansatz $t(1) = \alpha/\sqrt{n}$, we see that α should equal (recall that $\alpha > 0$),

$$\alpha = \frac{\xi_1 + \sqrt{\xi_1^2 + 4\chi_l^2\zeta_1}}{2\zeta_1} .$$

Now

$$h(t(1)) - \frac{1}{\epsilon} = O_P\left(\frac{1}{n}\right) ,$$

and in a neighborhood of $\alpha/\sqrt{n}$, h is Lipschitz with Lipschitz constant bounded away from 0. Hence, as argued in 7.3.1

$$T = \frac{\alpha}{\sqrt{n}} + O_P(\frac{1}{n}) .$$

7.3.4. *Case* $\epsilon > \epsilon_0$. Recall that the equation defining T is

$$\frac{1}{\epsilon} = \frac{\chi_l^2}{n}\frac{1}{T} + \frac{1}{n}\sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2 - 1}{T + \gamma + \delta_j} + \frac{1}{n}\sum_{j=l+1}^{n} \frac{1}{T + \gamma + \delta_j} .$$

When $\epsilon > \epsilon_0$, we can find t_0 bounded away from 0 such that

$$\frac{1}{\epsilon} = \frac{1}{n}\sum_{j=l+1}^{n} \frac{1}{t_0 + \gamma + \delta_j} .$$

t_0 is furthermore bounded. So T is going to converge to t_0 and the question is to understand how far away it is. By writing $T = t_0 + \eta$, after expanding the equation characterizing T around t_0, we see that we have

$$\frac{\chi_l^2}{nt_0}\left(1 - \frac{\eta}{t_0}\right) + \frac{1}{\sqrt{n}}\xi(t_0) - \eta\zeta(t_0) = O(\eta^2) ,$$

where

$$\xi(t_0) = \frac{1}{\sqrt{n}}\sum_{j=l+1}^{n} \frac{\mathsf{z}_j^2 - 1}{t_0 + \gamma + \delta_j} = O_P(1) ,$$

and

$$\zeta(t_0) = \frac{1}{n}\sum_{j=l+1}^{n} \frac{1}{(t_0 + \gamma + \delta_j)^2} .$$

We conclude that, informally, $\eta = O_P(\frac{1}{\sqrt{n}})$. Now let us verify it properly. Let us call

$$t(1) = t_0 + \frac{1}{\sqrt{n}} \frac{\xi(t_0)}{\zeta(t_0)} \ .$$

The expansion above shows that

$$\frac{1}{\epsilon} - h(t(1)) = O_P(1/n) \ .$$

Because h is Lipschitz with Lipschitz constant bounded below in a neighborhood of t_0, we conclude as in 7.3.1 that

$$T = t_0 + \frac{1}{\sqrt{n}} \frac{\xi(t_0)}{\zeta(t_0)} + O_P(\frac{1}{n}) \ .$$

7.4. **On the secular equation and higher-order perturbations.** We give an elementary proof of the validity of the secular equation, which avoids matrix representations. Though simple and likely well-known, the advantage of our derivation is that it extends easily to higher rank perturbation. More precisely, let us consider the matrix

$$M_1 = M + U \ , \tag{27}$$

where U is a symmetric matrix. We assume without loss of generality that M is diagonal. We write $U = \sum_{j=1}^k v_j v_j^T$. We do not require the v_j to be orthogonal and they could also be complex valued in what follows.

Let us call $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ the eigenvalues of M and compute the characteristic polynomial of M_1 and relate it to that of M. We call

$$\begin{aligned} P_{M_1}(\lambda) &= \det(M_1 - \lambda \mathbf{I}_n) \ , \\ P_M(\lambda) &= \det(M - \lambda \mathbf{I}_n) \ , \\ M_\lambda &= M - \lambda \mathbf{I}_n \ . \end{aligned}$$

Assuming for a moment that λ is not an eigenvalue of M, we clearly have $M_1 - \lambda \mathbf{I}_n = M_\lambda(\mathbf{I}_n + M_\lambda^{-1} U)$. We call $G(\lambda)$ the $k \times k$ matrix with (i,j) entry $v_j^T M_\lambda^{-1} v_i$.

We have

$$P_{M_1}(\lambda) = \det(M_\lambda) \det(\mathbf{I}_n + M_\lambda^{-1} U) = P_M(\lambda) \det(\mathbf{I}_k + G(\lambda)) \ ,$$

since $\det(\mathbf{I}_n + AB) = \det(\mathbf{I}_k + BA)$ for rectangular matrices A and B whenever AB is $n \times n$ and BA is $k \times k$. The previous formula can be used to study the eigenvalues of finite rank perturbations of M, since they are the zeros of the characteristic polynomial P_{M_1}.

Let us focus on the rank one case. Since we assume wlog that M is diagonal, we have, when $k = 1$,

$$\det(\mathbf{I}_k + G(\lambda)) = \det(1 + v^T M_\lambda^{-1} v) = 1 + \sum_{i=1}^n \frac{\mathsf{v}_i^2}{\lambda_i - \lambda} \ .$$

We therefore get , when λ is not an eigenvalue of M,

$$P_{M_1}(\lambda) = \left[\prod_{i=1}^n (\lambda_i - \lambda) \right] \left(1 + \sum_{i=1}^n \frac{\mathsf{v}_i^2}{\lambda_i - \lambda} \right) \ , \tag{28}$$

from which the secular equation follows. From Equation (28), it is also clear that if λ_i is an eigenvalue of M with multiplicity k, λ_i is also an eigenvalue of M_1 with multiplicity $k - 1$.

Acknowledgments

Alexandre d'Aspremont would like to acknowledge partial support from NSF grants SES-0835550 (CDI), CMMI-0844795 (CAREER), CMMI-0968842, a starting grant from the European Research Council (project SIPA), a Peek junior faculty fellowship, a Howard B. Wentz Jr. award and a gift from Google. Noureddine El Karoui acknowledges support from an Alfred P. Sloan research Fellowship and NSF grant DMS-0847647 (CAREER).

References

A. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biology*, 96:6745–6750, 1999.

E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, et al. *LAPACK Users' guide*. Society for Industrial Mathematics, 1999.

M. Baes, M. Bürgisser, and A. Nemirovski. A randomized mirror-prox method for solving structured large-scale matrix saddle-point problems. *Arxiv preprint arXiv:1112.1274*, 2011.

J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for non-null complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697, 2005.

A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization : analysis, algorithms, and engineering applications*. MPS-SIAM series on optimization. Society for Industrial and Applied Mathematics : Mathematical Programming Society, Philadelphia, PA, 2001.

S. Burer and R.D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Mireille Capitaine, Catherine Donati-Martin, and Delphine Féral. The largest eigenvalues of finite rank deformation of large Wigner matrices: convergence and nonuniversality of the fluctuations. *Ann. Probab.*, 37(1):1–47, 2009. ISSN 0091-1798. doi: 10.1214/08-AOP394. URL http://dx.doi.org/10.1214/08-AOP394.

A. d'Aspremont. Subsampling algorithms for semidefinite programming. *arXiv:0803.1990Version3*, 2008.

A. d'Aspremont. Subsampling algorithms for semidefinite programming. *Stochastic Systems*, 2(1):274–305, 2011.

A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007.

Kenneth R. Davidson and Stanislaw J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pages 317–366. North-Holland, Amsterdam, 2001.

Inderjit S. Dhillon and Beresford N. Parlett. Orthogonal eigenvectors and relative gaps. *SIAM Journal on Matrix Analysis and Applications*, 25(3):858–899, 2003.

I.S. Dhillon, B.N. Parlett, and C. Vömel. The design and implementation of the MRRR algorithm. *ACM Transactions on Mathematical Software (TOMS)*, 32(4):560, 2006.

R. Durrett. *Probability: theory and examples*. Cambridge Univ Pr, 2010.

Noureddine El Karoui. Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability*, 35(2):663–714, March 2007.

G.H. Golub and C.F. Van Loan. Matrix computation. *North Oxford Academic*, 1990.

C. Helmberg and F. Rendl. A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10(3):673–696, 2000.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer, 1993.

K. Johansson. From Gumbel to Tracy-Widom. *Probab. Theory Related Fields*, 138(1-2):75–112, 2007. ISSN 0178-8051.

M. Journée, F. Bach, P.A. Absil, and R. Sepulchre. Low-rank optimization for semidefinite convex problems. *Arxiv preprint arXiv:0807.4423*, 2008.

A. Juditsky, A.S. Nemirovskii, and C. Tauvel. Solving variational inequalities with Stochastic Mirror-Prox algorithm. *Arxiv preprint arXiv:0809.0815*, 2008.

T. Kato. *Perturbation theory for linear operators*. Springer, 1995.

J. Kuczynski and H. Wozniakowski. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM J. Matrix Anal. Appl*, 13(4):1094–1122, 1992.

G. Lan. An optimal method for stochastic composite optimization. *Technical report, School of Industrial and Systems Engineering, Georgia Institute of Technology, 2009*, 2009.

R.B. Lehoucq, D.C. Sorensen, and C. Yang. *ARPACK: Solution of Large-scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. Society for Industrial & Applied Mathematics, 1998.

A.S. Lewis and H.S. Sendov. Quadratic expansions of spectral functions. *Linear algebra and its applications*, 340(1): 97–121, 2002.

Z. Lu, A. Nemirovski, and R.D.C. Monteiro. Large-scale semidefinite programming via a saddle point Mirror-Prox algorithm. *Mathematical Programming*, 109(2):211–237, 2007.

Peter D. Miller. *Applied asymptotic analysis*, volume 75 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2006. ISBN 0-8218-4078-9.

C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

Boaz Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817, December 2008.

A. Nemirovski. Prox-method with rate of convergence O(1/T) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1): 229–251, 2004.

A. Nemirovskii and D. Yudin. Problem complexity and method efficiency in optimization. *Nauka (published in English by John Wiley, Chichester, 1983)*, 1979.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Y. Nesterov. *Introductory Lectures on Convex Optimization*. Springer, 2003.

Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming*, 110 (2):245–259, 2007a.

Y. Nesterov. Gradient methods for minimizing composite objective function. *CORE DP2007/96*, 2007b.

Y. Nesterov. Random gradient-free minimization of convex functions. *CORE Discussion Papers*, 2011.

M. L. Overton and Robert S. Womersley. Second derivatives for optimizing eigenvalues of symmetric matrices. *SIAM J. Matrix Anal. Appl.*, 16(3):697–718, 1995.

Y. Saad. *Numerical methods for large eigenvalue problems*. Manchester Univ Press, 1992. URL http://www-users.cs.umn.edu/$\sim$saad/books.html.

G.W. Stewart. *Matrix Algorithms Vol. II: Eigensystems*. Society for Industrial Mathematics, 2001.

Craig A. Tracy and Harold Widom. Level-spacing distributions and the Airy kernel. *Comm. Math. Phys.*, 159(1): 151–174, 1994. ISSN 0010-3616. URL http://projecteuclid.org/getRecord?id=euclid.cmp/1104254495.

Hale F. Trotter. Eigenvalue distributions of large Hermitian matrices; Wigner's semicircle law and a theorem of Kac, Murdock, and Szegő. *Adv. in Math.*, 54(1):67–82, 1984. ISSN 0001-8708.

CMAP, École Polytechnique, UMR CNRS 7641
E-mail address: alexandre.daspremont@m4x.org

Statistics, U.C. Berkeley. Berkeley, CA 94720.
E-mail address: nkaroui@stat.berkeley.edu