A STOCHASTIC SMOOTHING ALGORITHM FOR SEMIDEFINITE PROGRAMMING

ALEXANDRE D'ASPREMONT AND NOUREDDINE EL KAROUI

ABSTRACT. We use rank one Gaussian perturbations 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, and detail a variant of this stochastic algorithm with monotonic line search. Overall, compared to classical smooth algorithms, this method runs a larger number of significantly cheaper iterations and, in certain precision/dimension regimes, its total complexity is lower than that of deterministic smoothing algorithms.

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. we 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 [Helmberg and Rendl, 2000]. 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 minimize directly $\lambda_{\max}(X)$. 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 $O(n^2 \log n)$ (see the Appendix for more details) 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 $O(D_Q\sqrt{\log n}/\epsilon)$ 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

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

Date: March 4, 2014.

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

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

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 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 recent independent results similar in spirit to those presented here, 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 [Baes et al., 2011] 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. However, Baes et al. [2011] use stochastic techniques to reduce the cost of computing classical smoothing steps (matrix exponential, etc.) and Juditsky et al. [2008] use them to reduce the cost of linear algebra operations. In this work, we directly use stochastic methods for smoothing.

In this paper, we use stochastic smoothing results, combined with an optimal accelerated algorithm for stochastic optimization recently developed by Lan [2012], to derive a stochastic algorithm for solving (1). The algorithm detailed below 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 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, we briefly outline the stochastic smoothing algorithm for maximum eigenvalue minimization and compare its complexity with existing first-order algorithms. Section 3 details 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, 2012] where the scale of the step size is allowed to vary adaptively (but monotonically). Section 6.4 informally discusses extensions of our results to other smoothing techniques, together with their impact on complexity. Section 5 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, technical details about various functions that play a central role in our analysis, 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 $\stackrel{\mathcal{L}}{=}$ and $\Longrightarrow$

stands for convergence in law. We use the notation O_P with the standard probabilistic meaning (see [van der Vaart, 1998], p.12). When we compute local Lipschitz constants, they are always computed with respect to Euclidian or Frobenius norms, unless otherwise noted. We call $L\left[\Gamma(X)\right]$ the local Lipschitz constant of the function Γ at X.

2. Stochastic smoothing algorithm

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

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

in the variable $X \in \mathbf{S}_n$, where $Q \subset \mathbf{S}_n$ is a compact convex set, $z_i \overset{i.i.d}{\sim} \mathcal{N}(0, \mathbf{I}_n)$, $\epsilon \geq 0$ is in $\mathbb{R}$ and $k > 0$ is a small constant (typically 3). We call F_k^* the optimal value of this problem. We also define $F_k(X)$ as the random valued function inside the expectation, with

$$F_k(X) \triangleq \max_{i=1,\ldots,k} \lambda_{\max}\left(X + \frac{\epsilon}{n} z_i z_i^T\right) \tag{4}$$

so that $\mathsf{F}_k(X) = \mathbf{E}\left[F_k(X)\right]$. We have the following approximation result.

Lemma 2.1. *$\mathsf{F}_k(X)$ is a $c_k\epsilon$-uniform approximation of $\lambda_{\max}(X)$, where*

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

In other words, for all $X \in \mathbf{S}_n$

$$\lambda_{\max}(X) + \frac{\epsilon}{n} \leq \mathsf{F}_k(X) \leq \lambda_{\max}(X) + c_k\epsilon\ . \tag{5}$$

Proof. We first establish the upper bound. The fact that $\lambda_{\max}(\cdot)$ is subadditive on $\mathbf{S}_n$ gives

$$\lambda_{\max}\left(X + \frac{\epsilon}{n} z_i z_i^T\right) \leq \lambda_{\max}(X) + \lambda_{\max}\left(\frac{\epsilon}{n} z_i z_i^T\right) = \lambda_{\max}(X) + \epsilon\frac{\|z_i\|^2}{n}\ ,$$

since $\lambda_{\max}(z_i z_i^T) = \|z_i\|^2$. It follows that

$$\max_{1\leq i\leq k} \lambda_{\max}\left(X + \frac{\epsilon}{n} z_i z_i^T\right) \leq \max_{1\leq i\leq k} \lambda_{\max}(X) + \epsilon \max_{1\leq i\leq k} \frac{\|z_i\|^2}{n}\ ,$$

and

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

Let us now prove the lower bound. The mapping $M \mapsto \lambda_{\max}(X + M)$ is convex from $\mathbf{S}_n$ to $\mathbb{R}$ when $X \in \mathbf{S}_n$. Therefore, Jensen's inequality applied to this mapping with the random varible $z_i z_i^T$ gives

$$\lambda_{\max}(X + \frac{\epsilon}{n}\mathbf{E}\left[z_i z_i^T\right]) \leq \mathbf{E}\left[\lambda_{\max}\left(X + \frac{\epsilon}{n} z_i z_i^T\right)\right]\ .$$

Using $\mathbf{E}[z_i z_i^T] = \mathbf{I}_n$, we conclude that

$$\begin{aligned} \forall 1 \leq i \leq k,\ \lambda_{\max}(X + \frac{\epsilon}{n}\mathbf{I}_n) &\leq \mathbf{E}\left[\lambda_{\max}\left(X + \frac{\epsilon}{n} z_i z_i^T\right)\right]\ , \text{ hence} \\ \lambda_{\max}(X) + \frac{\epsilon}{n} &\leq \mathbf{E}\left[\max_{i=1,\ldots,k} \lambda_{\max}\left(X + \frac{\epsilon}{n} z_i z_i^T\right)\right], \end{aligned}$$

which is the lower bound above. ■

We begin by briefly introducing the smoothing results on (3) detailed in Section 3, then describe our main algorithm.

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.

2.1. **Smoothness of $\mathsf{F}_k(X)$.** In Section 3, we will show that the function F_k has a Lipschitz continuous gradient w.r.t. the Frobenius norm, i.e.

$$\|\nabla \mathsf{F}_k(X) - \nabla \mathsf{F}_k(Y)\|_F \leq L\|X - Y\|_F \,,$$

with (uniform) constant L satisfying

$$L \leq C_k \frac{n}{\epsilon} \,, \tag{6}$$

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 3.9).

2.2. **Gradient variance.** Section 3 also produces an explicit expression for the gradient of F_k. Let ϕ_{i_0} be a leading eigenvector of the matrix $X + \frac{\epsilon}{n} z_{i_0} z_{i_0}^T$ where

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

We will see that i_0 is unique with probability one and that we have

$$\nabla \mathsf{F}_k(X) = \mathbf{E}\left[\phi_{i_0}\phi_{i_0}^T\right] \quad \text{and} \quad \mathbf{E}\left[\left\|\phi_{i_0}\phi_{i_0}^T - \nabla \mathsf{F}_k(X)\right\|_F^2\right] \leq 1 \,. \tag{7}$$

Therefore the variance of the stochastic gradient oracle $\phi_{i_0}\phi_{i_0}^T$ is bounded by one. Once again, we will see in Section 3 that this bound too is often quite conservative.

2.3. **Stochastic algorithm.** Given an unbiased estimator for $\nabla \mathsf{F}_k$ with unit variance, the optimal algorithm for stochastic optimization derived in [Lan, 2012] will produce a (random) matrix X_N such that

$$\mathbf{E}[\mathsf{F}_k(X_N) - \mathsf{F}_k^*] \leq \frac{4LD_Q^2}{\alpha N^2} + \frac{4D_Q}{\sqrt{Nq}} \tag{8}$$

after N iterations [Lan, 2012, Corollary 1], where $L \leq C_k n/\epsilon$ is the Lipschitz constant of $\nabla \mathsf{F}_k$ discussed in the previous section, α is the strong convexity constant of the prox function, and q is the number of independent 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 $q = \lceil \max\{1, D_Q/(\epsilon\sqrt{n})\} \rceil$, the approximation bounds in Proposition 3.7 will then ensure $\mathbf{E}[\mathsf{F}_k(X_N) - \mathsf{F}_k^*] \leq 5\epsilon$. We compare in Table 1 the computational cost of the smooth stochastic algorithm in [Lan, 2012, Corollary 1] in this setting with that of the smoothing technique in [Nesterov, 2007a] and the nonsmooth stochastic averaging method. Recall that the cost of computing one leading eigenvector of $X + vv^T$ is of order $O(n^2 \log n)$ (cf. Appendix) 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 algorithm 1 detailed on page 17 is lower than that of both smooth and nonsmooth methods. This is the case for instance when $D_Q \geq \sqrt{n}\epsilon$ and

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

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 start by recalling a classical argument about Gaussian regularization and then improve smoothing performance by using explicit structural results on the spectrum of rank one updates of symmetric matrices.

3.1. Gaussian smoothing.

The following is 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* $\mathsf{f} : \mathbb{R}^m \to \mathbb{R}$ *is Lipschitz continuous with constant* μ *with respect to the Euclidean norm. The function* sf *such that*

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

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

$$\|\nabla \mathsf{sf}(x) - \nabla \mathsf{sf}(y)\|_2 \leq \frac{2\sqrt{m}\mu}{\epsilon}\|x - y\|_2.$$

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

Let us consider the function $\mathsf{F}_{GUE}(X)$ taking values

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

where $U \in \mathbf{S}_n$ is a symmetric matrix with standard normal upper triangle coefficients. Using convexity and positive homogeneity of the $\lambda_{\max}(X)$ function, together with the fact that it 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 $\lambda_{\max}(X)$.

Lemma 3.1 above shows that $\mathsf{F}_{GUE}(X)$ has a Lipschitz continuous gradient with constant bounded by $O\left(n^{3/2}/\epsilon\right)$, since, with the notation of Lemma 3.1, $m = n^2$. 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. We present a short discussion on a finer bound on the Lipschitz-constant of this function in Section 6.4.

3.2. Gradient smoothness.

We recall the following classical result (which can be derived from results in [Kato, 1995] and [Lewis and Sendov, 2001] 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 also that* $\lambda_{\max}(X)$*, the largest eigenvalue of* X*, has multiplicity one. Let* Y *be a symmetric matrix with* $\|Y\|_F = 1$ *and call*

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

Call $\lambda_{\max}$ the mapping $X \mapsto \lambda_{\max}(X)$. Then the local Lipschitz constant - with respect to the Frobenius norm - of the gradient of the mapping $\lambda_{\max}$ is given by

$$L\left[\nabla\lambda_{\max}(X)\right] = \sup_{Y\in\mathbf{S}_n, \|Y\|_F=1} \gamma(X,Y) = \frac{1}{\lambda_{\max}(X) - \lambda_2(X)}. \tag{9}$$

This result shows that 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.** The following proposition summarizes the information we will need about the impact of rank-one updates on the largest eigenvalue of a symmetric matrix. Equation (10) below will prove useful later to control the smoothness of $\nabla\mathsf{F}_k(X)$.

Proposition 3.3. *Suppose $X \in \mathbf{S}_n$ and has spectral decomposition $X = \mathcal{O}_X^T D_X \mathcal{O}_X$. Let $v \neq 0$ be a vector in $\mathbb{R}^n$ which is not an eigenvector of X. Let $\epsilon > 0$ be in $\mathbb{R}$. Then, $\lambda_{\max}(X + (\epsilon/n)vv^T)$ has multiplicity 1 and $\lambda_{\max}(X + (\epsilon/n)vv^T) - \lambda_{\max}(X) > 0$. Let us call λ_2 the second largest eigenvalue of a symmetric matrix. Then, if $(\mathcal{O}_X v)_1$ is the first coordinate of the vector $\mathcal{O}_X v$, we have*

$$\frac{\epsilon(\mathcal{O}_X v)_1^2}{n} \leq \lambda_{\max}\left(X + \frac{\epsilon}{n}vv^T\right) - \lambda_{\max}(X) \leq \lambda_{\max}\left(X + \frac{\epsilon}{n}vv^T\right) - \lambda_2\left(X + \frac{\epsilon}{n}vv^T\right) . \tag{10}$$

Proof. For $X \in \mathbf{S}_n$, we call $\lambda(X) \in \mathbb{R}^n$ the spectrum of the matrix X, in decreasing algebraic order. Whenever $v \neq 0$ is not an eigenvector of X and $\epsilon > 0$, the leading eigenvalue l_1 of the matrix $X+(\epsilon/n)vv^T$, is always strictly larger than $\lambda_1(X)$ [see Golub and Van Loan, 1990, §8.5.3] and we write $l_1 = \lambda_1(X) + t$, $t \geq 0$. Our aim is now to characterize t and understand its properties. We note that

$$X + (\epsilon/n)vv^T = \mathcal{O}_X^T\left[D_X + (\epsilon/n)(\mathcal{O}_X v)(\mathcal{O}_X v)^T\right]\mathcal{O}_X .$$

Since we are interested in eigenvalues, we assume without loss of generality that X is diagonal. If X were not diagonal, we would just need to replace v by $(\mathcal{O}_X v)$ in what follows for all the statements to hold.

It is a standard result (see e.g Theorem 8.5.3 in [see Golub and Van Loan, 1990, §8.5.3]) that the variable t is the unique positive solution of the *secular equation*

$$s(t) \triangleq \frac{n}{\epsilon} - \frac{\mathsf{v}_1^2}{t} - \sum_{i=2}^{n} \frac{\mathsf{v}_i^2}{(\lambda_1(X) - \lambda_i(X)) + t} = 0, \tag{11}$$

where v_i are the coefficients of the vector v; we give an elementary derivation of this result in Subsection 6.3 in the Appendix. We plot the function $s(\cdot)$ for a sample matrix X in Figure 1.

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.

$$\lambda_n(X) \leq \lambda_n(X + \frac{\epsilon}{n}vv^T) \leq \ldots \leq \lambda_2(X + \frac{\epsilon}{n}vv^T) \leq \lambda_{\max}(X) < \lambda_{\max}(X + \frac{\epsilon}{n}vv^T).$$

This implies in particular that $\lambda_{\max}(X + \frac{\epsilon}{n}vv^T)$ has multiplicity 1. 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 $(0, \infty)$. Since both functions are non-decreasing, the positive root of the equation $s^+(t) = 0$ is a lower bound on the positive root t^* of the equation $s(t) = 0$. We therefore have

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

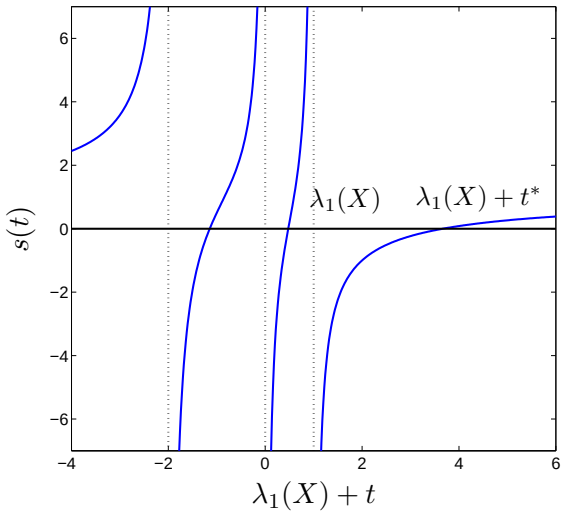

FIGURE 1. Plot of $s(t)$ versus $\lambda_1(X)+t$. The matrix has dimension four and its spectrum is here $\{-2,-2,0,1\}$. The three leading eigenvalues of $X+\epsilon vv^T$ are the roots of $s(t)$, the fourth eigenvalue is at -2.

Using interlacing, we also 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

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

which yields (10) and will allow us to control the smoothness of $\nabla \mathsf{F}_k(X)$. ■

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

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

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 $\mathsf{F}_k(X)$. This requires a few preliminaries which we now present.

Lemma 3.4. *Let $\lambda_1(X)+T$ be the largest eigenvalue of the matrix $X+(\epsilon/n)zz^T$, where $X\in\mathbf{S}_n$ is a given deterministic matrix and $z\sim\mathcal{N}(0,\mathbf{I}_n)$. Then the random variable T has a density on $[0,\infty)$.*

The proof of this lemma is in the Appendix in §6.2.2. Two corollaries immediately follow. The first one shows that two perturbed eigenvalues obtained from independent rank one perturbations are different with probability one.

Corollary 3.5. *Suppose $l_{1,1}=\lambda_{\max}(X+(\epsilon/n)z_1z_1^T)$ and $l_{1,2}=\lambda_{\max}(X+(\epsilon/n)z_2z_2^T)$, 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. The result follows from Lemma 3.4 since $l_{1,1}-\lambda_{\max}(X)$ and $l_{1,2}-\lambda_{\max}(X)$ are two independent draws from a distribution with a density on $[0,\infty)$ and $P(l_{1,1}-\lambda_{\max}(X)=0)=P(l_{1,2}-\lambda_{\max}(X)=0)=0$. ■

The second corollary shows that the maximum of (independent) perturbed eigenvalues is differentiable with probability one and bounds its Lipschitz constant.

Corollary 3.6. *Let $X \in \mathbf{S}_n$ and suppose $l_{1,i} = \lambda_{\max}(X + (\epsilon/n) z_i z_i^T)$, where $z_i \overset{i.i.d}{\sim} \mathcal{N}(0, \mathbf{I}_n)$ for $i = 1, \ldots, k$. The mapping $F_k : X \to \max_{i=1,\ldots,k} l_{1,i}$ is differentiable with probability one. Then, if $i_0 = \operatorname{argmax}_{1 \leq i \leq k} l_{1,i}$ and ϕ_{i_0} is an eigenvector associated with the eigenvalue l_{1,i_0}, its gradient is*

$$\nabla F_k(X) = \phi_{i_0} \phi_{i_0}^T \,. \tag{12}$$

Also, with probability 1, the local Lipschitz constant of ∇F_k is bounded by

$$L\left[\nabla F_k(X)\right] \leq \frac{1}{F_k(X) - \lambda_{\max}(X)} \,. \tag{13}$$

Proof. We first recall that it is well-known (and indeed follows from results in [Kato, 1995]) that if a matrix M_0 has a unique largest eigenvalue, the gradient of $M \mapsto \lambda_{\max}(M)$ at M_0 is simply $\phi_0 \phi_0^T$, where ϕ_0 is an eigenvector associated with $\lambda_{\max}(M_0)$.

Corollary 3.5 shows that with probability 1, there exists a unique i_0 such that $l_{1,i_0} = F_k(X)$. Furthermore, since with probability 1, z_{i_0} is not an eigenvector of X, Proposition 3.3 shows that the multiplicity of the largest eigenvalue of $X + (\epsilon/n) z_{i_0} z_{i_0}^T$ is one. This implies that $X \mapsto \lambda_{\max}(X + (\epsilon/n) z_{i_0} z_{i_0}^T)$ is differentiable at X with probability 1. Lemma 6.2 then applies and shows that F_k is differentiable at X with probability 1. Our reminder on the gradient of $M \mapsto \lambda_{\max}(M)$ gives the value of the differential. The last part of the corollary follows from Lemma 6.3, whose assumptions are clearly satisfied with probability 1. ■

We now use these preliminary results to prove the main result of this section, namely a bound on the Lipschitz constant of the gradient of $\mathsf{F}_k(X)$ defined above, using the spectral gap bound in (10).

Proposition 3.7. *Let $\{z_i\}_{i=1}^k$ be i.i.d. $\mathcal{N}(0, \mathbf{I}_n)$, $k \geq 3$ be an integer and $X \in \mathbf{S}_n$. The function F_k such that*

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

is smooth. The Lipschitz constant L of its gradient w.r.t. the Frobenius norm satisfies

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

Proof. The fact that F_k is smooth follows from Equation (12) and the fact that we can interchange expectation and differentiation here. Details about the validity of this interchange - whose proof requires care - are in the Appendix in Lemma 6.4. We assume without loss of generality that X is diagonal - Lemma 6.1 proving that we can do so. Let us call z_i the first coordinate of the vector z_i. The spectral gap bound in (10) gives

$$\forall\, i, 1 \leq i \leq k \,, \quad \lambda_{\max}(X + \frac{\epsilon}{n} z_i z_i^T) - \lambda_{\max}(X) \geq \frac{\epsilon}{n} \mathsf{z}_i^2 \,.$$

It follows that

$$F_k(X) - \lambda_{\max}(X) \geq \frac{\epsilon}{n} \max_{1 \leq i \leq k} \mathsf{z}_i^2 \,, \text{ and}$$
$$\frac{1}{F_k(X) - \lambda_{\max}(X)} \leq \frac{n}{\epsilon} \frac{1}{\max_{i=1,\ldots,k} \mathsf{z}_i^2} \,.$$

The results of Corollary 3.6 then guarantee that with probability 1, we have

$$L\left[\nabla F_k(X)\right] \leq \frac{n}{\epsilon} \frac{1}{\max_{i=1,\ldots,k} \mathsf{z}_i^2} \,,$$

and therefore

$$L\left[\nabla \mathsf{F}_k(X)\right] \leq \mathbf{E}\left[\frac{n}{\epsilon} \frac{1}{\max_{i=1,\ldots,k} \mathsf{z}_i^2}\right] \leq \mathbf{E}\left[\frac{n}{\epsilon} \frac{1}{\sum_{i=1}^k \mathsf{z}_i^2 / k}\right] = \mathbf{E}\left[\frac{n}{\epsilon} \frac{k}{\chi_k^2}\right]$$

where χ^2_k is χ^2 distributed with k degrees of freedom. The fact that

$$\mathbf{E}\left[\frac{1}{\chi^2_k}\right] = \frac{1}{k-2}$$

whenever $k \geq 3$ - see e.g. [Mardia et al., 1979, p. 487] - yields

$$\forall X \in \mathbf{S}_n \ , \ \ L\left[\nabla \mathsf{F}_k(X)\right] \leq C_k \frac{n}{\epsilon} \ .$$

The function $\nabla \mathsf{F}_k$ is thus Lipschitz with Lipschitz constant

$$L \leq \sup_{X \in \mathbf{S}_n} L\left[\nabla \mathsf{F}_k(X)\right] \leq C_k \frac{n}{\epsilon} \ ,$$

which concludes the proof. ■

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

$$\mathbf{E}\left[1/\max\{\mathsf{z}_1^2, \mathsf{z}_2^2, \mathsf{z}_3^2\}\right] = 1.5...$$

while $C_3 = 3$, for example. We could of course use the density of the minimum above to get a more accurate bound, but then C_k would not have a simple closed form.

3.5. **Gradient variance.** In this section, we will bound the variance of ∇F_k, the stochastic gradient oracle approximating $\nabla \mathsf{F}_k$.

Lemma 3.8. *Let $X \in \mathbf{S}_n$ and $z_i \overset{i.i.d}{\sim} \mathcal{N}(0, \mathbf{I}_n)$, the gradient of $\mathsf{F}_k(X)$ is given by*

$$\nabla \mathsf{F}_k(X) = \mathbf{E}[\phi_{i_0} \phi_{i_0}^T] \tag{14}$$

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

$$\mathbf{E}\left[\|\phi_{i_0}\phi_{i_0}^T - \mathbf{E}[\phi_{i_0}\phi_{i_0}^T]\|_F^2\right] = 1 - \mathbf{Tr}\left(\nabla \mathsf{F}_k(X)^2\right) \leq 1, \tag{15}$$

where $\mathbf{Tr}\left(\nabla \mathsf{F}_k(X)\right) = 1$ by construction.

Proof. Equation (14) follows from Equation (12) and the fact that we can interchange expectation and differentiation here (see Lemma 6.4 for details). We now focus on the variance $\mathbf{E}\left[\|\phi_{i_0}\phi_{i_0}^T - \mathbf{E}\left[\phi_{i_0}\phi_{i_0}^T\right]\|_F^2\right]$. Recall that for any symmetric matrix M, $\|M\|_F^2 = \mathbf{Tr}(M^T M) = \mathbf{Tr}\, M^2$. The matrix $\phi_{i_0}\phi_{i_0}^T - \mathbf{E}\left[\phi_{i_0}\phi_{i_0}^T\right]$ is symmetric. So we can rewrite

$$\|\phi_{i_0}\phi_{i_0}^T - \mathbf{E}\left[\phi_{i_0}\phi_{i_0}^T\right]\|_F^2 = \mathbf{Tr}\left(\phi_{i_0}\phi_{i_0}^T - \mathbf{E}\left[\phi_{i_0}\phi_{i_0}^T\right]\right)^2 .$$

Using the fact that $\phi_{i_0}^T \phi_{i_0} = 1$, we see that $(\phi_{i_0}\phi_{i_0}^T)^2 = \phi_{i_0}\phi_{i_0}^T$. Therefore,

$$\mathbf{E}\left[\mathbf{Tr}\left(\phi_{i_0}\phi_{i_0}^T\right)^2\right] = \mathbf{E}\left[\mathbf{Tr}\,\phi_{i_0}\phi_{i_0}^T\right] = \mathbf{E}\left[\phi_{i_0}^T\phi_{i_0}\right] = 1 \ .$$

Recalling that $\mathbf{E}\left[\phi_{i_0}\phi_{i_0}^T\right] = \nabla \mathsf{F}_k(X)$, we have shown that $\mathbf{Tr}\left(\nabla \mathsf{F}_k(X)\right) = 1$. We also see that

$$\mathbf{E}\left[\mathbf{Tr}\left(\phi_{i_0}\phi_{i_0}^T - \mathbf{E}\left[\phi_{i_0}\phi_{i_0}^T\right]\right)^2\right] = \mathbf{Tr}\left(\nabla \mathsf{F}_k(X)\right) - \mathbf{Tr}\left(\nabla \mathsf{F}_k(X)^2\right) = 1 - \mathbf{Tr}\left(\nabla \mathsf{F}_k(X)^2\right) \leq 1 \ .$$

which is the desired result. ■

Furthermore, we show in Lemma 6.5 in the Appendix that $\nabla \mathsf{F}_k$ is diagonalizable in the same basis as X. In particular, when X is diagonal, so is $\nabla \mathsf{F}_k$. Simply using the fact that ϕ_{i_0} is an eigenvector, we have of course

$$\|\phi_{i_0}\phi_{i_0}^T - \mathbf{E}[\phi_{i_0}\phi_{i_0}^T]\|_F^2 \leq 4 \tag{16}$$

which means that the gradient will naturally satisfy the "light-tail" condition A2 in [Lan, 2012] for $\sigma^2 = 4$. The bound in (15) together with the proof above (in particular Equation (33)) show that when the spectral gaps $\lambda_1(X) - \lambda_i(X)$ are large, the diagonal of $\nabla \mathsf{F}_k(X)$ is approximately sparse. In that scenario, $\mathbf{Tr}(\nabla \mathsf{F}_k(X)^2)$ is close to $\mathbf{Tr}(\nabla \mathsf{F}_k(X))$, hence close to one, and the variance of the gradient oracle is small.

3.6. **A phase transition.** We can push our analysis if the impact of the low rank perturbation a little bit further. We focus again on the properties of a random rank one perturbation of a deterministic matrix X, specifically $X(\epsilon) = X + (\epsilon/n) z z^T$, 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 next theorem is asymptotic in nature but is informative in practice even for moderate size matrices. We make the dependence on n, the dimension of the matrices we are working with, explicit everywhere. This undoubtedly makes for somewhat cumbersome notations but also makes the statement of the results less ambiguous. We will work under the following assumptions:

- A1 $X_n \in \mathbf{S}_n$. Its eigenvalues are denoted $\lambda_1(n) \geq \lambda_2(n) \geq \ldots \geq \lambda_n(n)$. $\lambda_1(n)$ has multiplicity $l_n \in \mathbb{N}$. There exists a constant $l \in \mathbb{N}$ such that $l_n \leq l$ for all n. We call $\gamma_n = \lambda_1(n) - \lambda_{l_n+1}(n)$ and assume that there exists a constant γ such that $\gamma_n \geq \gamma > 0$. We call $\lambda_1(n) - \lambda_i(n) = \gamma_n + \delta_{i,n}$, for $i > l(n)$. Of course, $\delta_{i,n} \geq 0$.
- A2 ϵ_n is a sequence in $\mathbb{R}$. We assume that $\epsilon_n \asymp 1$, i.e $\liminf_{n\to\infty} \epsilon_n > 0$ and $\limsup_{n\to\infty} \epsilon_n < \infty$.
- A3 We assume that there exists a constant C, independent of n such that

$$\frac{1}{\gamma^2} > \frac{1}{n} \sum_{j=l_n+1}^{n} \frac{1}{(\gamma_n + \delta_{j,n})^2} > \frac{1}{n} \sum_{j=l_n+1}^{n} \frac{1}{(\epsilon_n + \gamma_n + \delta_{j,n})^2} > C \, .$$

Theorem 3.9 (**Phase transition for the largest eigenvalue: rank one perturbation**). *Assume that Assumptions A1-A3 above are satisfied and consider the matrix*

$$X_n(\epsilon_n) = X_n + \frac{\epsilon_n}{n} z z^T \ , \ \textit{where } z \sim \mathcal{N}(0, \mathbf{I}_n) \, .$$

Define $\epsilon_{0,n}$ *by*

$$\frac{1}{\epsilon_{0,n}} = \frac{1}{n} \sum_{j=l_n+1}^{n} \frac{1}{\gamma_n + \delta_{j,n}} \, .$$

Call, for i.i.d $\mathcal{N}(0,1)$ *random variables* $\{\mathsf{z}_{j,n}\}_{j=1}^n$, $\chi^2_{l_n} = \sum_{j=1}^{l_n} \mathsf{z}^2_{j,n}$,

$$\xi_{1,n} = \frac{1}{\sqrt{n}} \sum_{j=l_n+1}^{n} \frac{\mathsf{z}^2_{j,n} - 1}{\gamma_n + \delta_{j,n}} = O_P(1) \ \textit{and} \ \zeta_{1,n} = \frac{1}{n} \sum_{j=l_n+1}^{n} \frac{\mathsf{z}^2_{j,n}}{(\gamma_n + \delta_{j,n})^2} = O_P(1) \, .$$

We have the following three situations:

(1) *If* $0 < \epsilon_n < \epsilon_{0,n}$ *and* $\liminf_{n\to\infty} [\epsilon_{0,n} - \epsilon_n] > 0$, *as* $n \to \infty$,

$$\lambda_{\max}[X_n(\epsilon_n)] = \lambda_{\max}[X_n] + \frac{W_{1,n}}{n} + \frac{W_{2,n}}{n^{3/2}} + O_P\left(\frac{1}{n^2}\right),$$

where

$$W_{1,n} = \frac{\chi^2_{l_n}}{1/\epsilon_n - 1/\epsilon_{0,n}} \quad \textit{and} \quad W_{2,n} = \frac{W_{1,n}\xi_{1,n}}{1/\epsilon_n - 1/\epsilon_{0,n}} \, .$$

(2) *If* $\epsilon_n = \epsilon_{0,n}$*, as* $n \to \infty$*,*

$$\lambda_{\max}[X_n(\epsilon_n)] = \lambda_{\max}[X_n] + \frac{W_{1,n}}{\sqrt{n}} + O_P\left(\frac{1}{n}\right),$$

where

$$W_{1,n} = \frac{\xi_{1,n} + \sqrt{\xi_{1,n}^2 + 4\chi_{l_n}^2 \zeta_{1,n}}}{2\zeta_{1,n}}.$$

(3) *If* $\epsilon_n > \epsilon_{0,n}$ *and* $\liminf_{n\to\infty}[\epsilon_n - \epsilon_{0,n}] > 0$*, call* $t_{0,n} > 0$*, the (unique) positive solution of*

$$\frac{1}{\epsilon_n} = \frac{1}{n}\sum_{j=l_n+1}^{n} \frac{1}{t_{0,n} + \gamma_n + \delta_{j,n}}.$$

Note that $t_{0,n} \leq (1 - l_n/n)\epsilon_n$*. Then, as* $n \to \infty$*,*

$$\lambda_{\max}[X_n(\epsilon_n)] = \lambda_{\max}[X_n] + t_{0,n} + \frac{W_{1,n}}{\sqrt{n}} + O_P\left(\frac{1}{n}\right).$$

Here, $W_{1,n} = \frac{\xi(t_{0,n})}{\zeta(t_{0,n})}$ *, where*

$$\xi(t_{0,n}) = \frac{1}{\sqrt{n}}\sum_{j=l_n+1}^{n} \frac{\mathsf{z}_{j,n}^2 - 1}{t_{0,n} + \gamma_n + \delta_{j,n}} = O_P(1) \text{ , and}$$

$$\zeta(t_{0,n}) = \frac{1}{n}\sum_{j=l_n+1}^{n} \frac{1}{(t_{0,n} + \gamma_n + \delta_{j,n})^2} = O(1) .$$

Proof. The strategy is the following. We are looking for the zeros of a certain random function - defined in the secular equation - which can be seen as a perturbation of a deterministic function. Hence, it is natural to use ideas from 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 now turn to the proof.

3.6.1. *Preliminaries.* Let us call

$$g_{l_n,n}(t) = \frac{1}{n}\sum_{j=l_n+1}^{n} \frac{1}{t + \gamma_n + \delta_{j,n}} ,$$

$$h_{l_n,n}(t) = \frac{1}{n}\sum_{j=l_n+1}^{n} \frac{\mathsf{z}_{j,n}^2}{t + \gamma_n + \delta_{j,n}} ,$$

$$h_n(t) = \frac{\sum_{j=1}^{l_n} \mathsf{z}_{j,n}^2}{n}\frac{1}{t} + h_{l_n,n}(t) .$$

Recall that if $\lambda_{\max}[X_n(\epsilon_n)] = \lambda_{\max}[X_n] + T$, T is the unique positive solution of the equation

$$\frac{1}{\epsilon_n} = h_n(T) = \frac{\sum_{j=1}^{l_n} \mathsf{z}_{j,n}^2}{n}\frac{1}{T} + \frac{1}{n}\sum_{j=l_n+1}^{n} \frac{\mathsf{z}_{j,n}^2}{T + \gamma_n + \delta_{j,n}} . \tag{17}$$

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

$$\mathbf{var}\left(\frac{1}{n}\sum_{j=l_n+1}^{n} \frac{\mathsf{z}_{j,n}^2 - 1}{t + \gamma_n + \delta_{j,n}}\right) = \frac{1}{n}\left[\frac{1}{n}\sum_{j=l_n+1}^{n} \frac{2}{(t + \gamma_n + \delta_{j,n})^2}\right] \leq \frac{2}{n}\frac{1}{\gamma^2} = O\left(\frac{1}{n}\right) .$$

It therefore follows from Chebyshev's inequality that the error made when replacing $h_{l_n,n}$ by $g_{l_n,n}$ when seeking the root of Equation (17) is $O_P(1/\sqrt{n})$.

Our strategy is to expand T in powers (possibly non-integer) of $1/n$. If we can find an approximate solution $t(m)$ of Equation (17), such that

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

we claim that

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

This is because h_n is, at $\mathsf{z}_{j,n}$ fixed, a Lipschitz function on $(\frac{\epsilon_n \sum_{j=1}^{l_n} \mathsf{z}_{j,n}^2}{n}, \infty)$, and its Lipschitz constant is bounded below with high-probability on any compact subinterval of this interval. Hence, we have, if $\|h_n^{-1}\|_{\mathrm{L},t(m),T}$ is the Lipschitz constant of h_n^{-1} over an interval to which both $t(m)$ and T belong,

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

Note that if we can show that $|h_n'(y)| > Cn^b$ in an interval containing both $t(m)$ and T, then $\|h_n^{-1}\|_{\mathrm{L},t(m),T} \leq n^{-b}/C$ and we get by the same token

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

More details about these estimates are given in 3.6.2, where we carry out a detailed proof.

To summarize, if we can come up with $t(m)$ which is a near solution of the equation $h_n(t) = 1/\epsilon_n$, it will be a good approximation of T. The quality of the approximation is detailed in the estimates above. In the proof below, we will exhibit such $t(m)$'s and, from them, get fine approximations of T. That is our strategy.

We finally recall that by definition

$$\frac{1}{\epsilon_{0,n}} = g_{l_n,n}(0) = \frac{1}{n} \sum_{j=l_n+1}^{n} \frac{1}{\gamma_n + \delta_{j,n}} \text{ .}$$

Intuitive explanations The following might help in giving the reader a sense of where the results come from. We wish to find an approximation of the root T of the equation $1/\epsilon_n = h_n(T)$. The overall strategy is to write a Laurent-series expansion of h_n around x_n, a real such that $h_n(x_n) - 1/\epsilon_n$ is small. Practically, calling $A_{k,x_n}(h_n)$ our expansion to order k of h_n around x_n, we solve exactly the equation $A_{k,x_n}(h_n)(t) = 1/\epsilon_n$. This strategy amounts practically to dropping various O_P terms in our expansions of h_n and solving the corresponding equations. Let us call x_n^* the solution of $A_{k,x_n}(h_n)(t) = 1/\epsilon_n$. Our proof shows that T is indeed close to x_n^*, to various order of accuracy.

More specifically, we break h_n into a component that stays bounded when $t \to 0$ - this is what $h_{l_n,n}$ is - and a component that behaves like $1/(nt)$ as $t \to 0$.

Cases 1) and 2) of the Theorem In these cases, it is clear that if $c > 0$, $\lim_{t \to c} h_n(t) < 1/\epsilon_n$ with high-probability. This suggests that $T \to 0$ with high-probability. Hence our strategy is to write $h_{l_n,n}(t) = P_{l_n,n}(t) + O_P(t^\alpha)$, where $P_{l_n,n}$ is a polynomial and α an integer, i.e expand $h_{l_n,n}$ in powers of t for t close to 0, and instead of solving $h_n(T) = 1/\epsilon_n$, solve the approximating equation $h_n(x) - h_{l_n,n}(x) + P_{l_n,n}(x) = 1/\epsilon_n$. This simply amounts to dropping the $O_P(t^\alpha)$ term from our (Laurent-series) expansion of $h_n(t)$ in a neighborhood of 0. This latter equation is a polynomial equation - hence it is easy to solve. Call x_n^* its solution. By construction, it is fairly clear that x_n^* is such that $h_n(x_n^*)$ is close to $1/\epsilon_n$. The proof makes this statement fully rigorous and pushes further to give rigorous statements concerning $T - x_n^*$, which is really the quantity we are interested in.

Case 3) of the Theorem In this case, it is clear that T has to remain bounded away from 0, since $h_{l_n,n}(0) > 1/\epsilon_n$ with probability going to 1. Hence, we employ the same strategy as the one described above, except that we expand $h_n(t)$ around $t_{0,n}$, a non-random sequence bounded away from 0 picked such

that $h_n(t_{0,n}) - 1/\epsilon_n \to 0$ in probability. h_n is linearized around $t_{0,n}$ to yield an approximating polynomial $P_{n,t_{0,n}}(t)$ of degree 1 and a remainder of the form $|t - t_{0,n}|^\alpha$. Our approximation x_n^* of T is simply the root of the equation $P_{n,t_{0,n}}(t) = 1/\epsilon_n$. Once again, this amounts to dropping the $O_P(|t - t_{0,n}|^\alpha)$ from our expansion of $h_n(t)$. The proof ensures that x_n^* has all the properties announced in the Theorem - in particular that it is close to T.

3.6.2. *Case* $\epsilon_n < \epsilon_{0,n}$. We treat this case in full detail and go faster on the two other ones, since the ideas are similar. Recall that the equation defining T is

$$\frac{1}{\epsilon_n} = h_n(T) = \frac{\sum_{j=1}^{l_n} \mathsf{z}_{j,n}^2}{n} \frac{1}{T} + h_{l_n,n}(T) \ .$$

In this case, $g_{l_n,n}(0) = \frac{1}{\epsilon_{0,n}} < \frac{1}{\epsilon_n}$. Let us first localize T. Denoting $\chi_{l_n}^2 = \sum_{j=1}^{l_n} \mathsf{z}_{j,n}^2$, and using $h_n(t) \geq \chi_{l_n}^2/(nt)$ as well as the fact that h_n is decreasing, we see that $T \geq (\chi_{l_n}^2/n)\epsilon_n$. On the other hand, $h_n(t) \leq \chi_{l_n}^2/(nt) + h_{l_n,n}(0)$. Recall that $h_{l_n,n}(0) = g_{l_n,n}(0) + O_P(n^{-1/2}) = 1/\epsilon_{0,n} + O_P(n^{-1/2})$. Simple algebra then gives that $T \leq (\chi_{l_n}^2/n)\epsilon_n/(1 - \epsilon_n h_{l_n,n}(0))$. Of course, in the situation we are investigating, $\epsilon_n h_{l_n,n}(0)$ is bounded away from 1 with probability going to 1 as $n \to \infty$.

Let us now expand the last term above, i.e $h_{l_n,n}(t)$, in powers of t's. Because $h'_{l_n,n}$ is uniformly bounded in probability for t in a neighborhood of 0, we have, for small t,

$$h_{l_n,n}(t) = \frac{1}{n} \sum_{j=l_n+1}^{n} \frac{\mathsf{z}_{j,n}^2}{\gamma_n + \delta_{j,n}} + O_P(t) = \frac{1}{\epsilon_{0,n}} + \frac{1}{\sqrt{n}} \xi_{1,n} + O_P(t) \ ,$$

where $\xi_{1,n} = \frac{1}{\sqrt{n}} \sum_{j=l_n+1}^{n} \frac{\mathsf{z}_{j,n}^2 - 1}{\gamma_n + \delta_{j,n}} = O_P(1)$. We see that by taking

$$t(2) = \frac{W_1}{n} \left(1 + \frac{1}{\sqrt{n}} \frac{\xi_{1,n}}{\frac{1}{\epsilon_n} - \frac{1}{\epsilon_{0,n}}} \right) \ , \text{ with } W_1 = \frac{\chi_{l_n}^2}{\frac{1}{\epsilon_n} - \frac{1}{\epsilon_{0,n}}} \ ,$$

we have

$$h_n(t(2)) - h_n(T) = h_n(t(2)) - \frac{1}{\epsilon_n} = O_P(1/n) \ .$$

It is clear that both $t(2)$ and T are contained in the interval $I = \left(\chi_{l_n}^2 \epsilon_n/n, 2\epsilon_n \chi_{l_n}^2 \upsilon_n/n\right)$, where $\upsilon_n = \max[1/(1 - \epsilon_n h_{l_n,n}(0)), 1/(1 - \epsilon_n/\epsilon_{0,n})]$. The mean value theorem gives

$$|t(2) - T| \leq \frac{|h_n(t(2)) - h_n(T)|}{\inf_{t \in I} |h'_n(t)|} \ .$$

Of course, $|h'_n(t)| \geq \chi_{l_n}^2/(nt^2) + |h'_{l_n,n}(0)| \geq \chi_{l_n}^2/(nt^2)$. So we see that $\inf_{t \in I} |h'_n(t)| \geq nR_n$, where R_n is a positive random variable bounded away from 0 with probability going to 1, since we assume that ϵ_n and l_n remain bounded. We conclude that

$$|t(2) - T| = O_P(|h_n(t(2)) - h_n(T)|/n) = O_P(n^{-2}) \ , \text{ as announced in Theorem 3.9.}$$

3.6.3. *Case* $\epsilon_n = \epsilon_{0,n}$. We now have, for small t, using the fact that $g_{l_n,n}(0) = 1/\epsilon_{0,n} = 1/\epsilon_n$,

$$h_n(t) = \frac{\chi_{l_n}^2}{n} \frac{1}{t} + \frac{1}{\epsilon_n} + \frac{1}{n} \sum_{j=l_n+1}^{n} \frac{\mathsf{z}_{j,n}^2 - 1}{\gamma_n + \delta_{j,n}} - t\zeta_{1,n} + O_P(t^2) \ ,$$

where $\zeta_{1,n} = \frac{1}{n}\sum_{j=l_n+1}^{n} \frac{\mathsf{z}_{j,n}^2}{(\gamma_n+\delta_{j,n})^2} = O_P(1)$. Because $\xi_{1,n} = \frac{1}{\sqrt{n}}\sum_{j=l_n+1}^{n} \frac{\mathsf{z}_{j,n}^2-1}{\gamma_n+\delta_{j,n}} = O_P(1)$, we see that now, T has to be of order $1/\sqrt{n}$, since $h_n(T) = 1/\epsilon_n$. Using the ansatz $t(1) = \alpha/\sqrt{n}$, we see that

$$h_n(t(1)) - \frac{1}{\epsilon_n} = O_P\left(\frac{1}{n}\right) \text{ , if } \alpha = \frac{\xi_{1,n} + \sqrt{\xi_{1,n}^2 + 4\chi_{l_n}^2\zeta_{1,n}}}{2\zeta_{1,n}} \text{ .}$$

In a neighborhood of $\alpha/\sqrt{n}$, h_n is Lipschitz with Lipschitz constant bounded away from 0, with probability going to 1. Hence, as argued in 3.6.1 and detailed in 3.6.2, we can conclude that

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

3.6.4. *Case* $\epsilon_n > \epsilon_{0,n}$. Recall that the equation defining T is

$$\frac{1}{\epsilon_n} = \frac{\chi_{l_n}^2}{n}\frac{1}{T} + \frac{1}{n}\sum_{j=l_n+1}^{n} \frac{\mathsf{z}_{j,n}^2-1}{T+\gamma_n+\delta_{j,n}} + \frac{1}{n}\sum_{j=l_n+1}^{n} \frac{1}{T+\gamma_n+\delta_{j,n}} \text{ .}$$

When $\epsilon_n > \epsilon_{0,n}$, we can find $t_{0,n}$ bounded away from 0 such that

$$\frac{1}{\epsilon_n} = \frac{1}{n}\sum_{j=l_n+1}^{n} \frac{1}{t_{0,n}+\gamma_n+\delta_{j,n}} \text{ .}$$

$t_{0,n}$ is furthermore bounded - in n - under our assumptions.

By writing $t = t_{0,n} + \eta$, for η small, after expanding h_n around $t_{0,n}$, we see that we have

$$h_n(t) = \frac{\chi_{l_n}^2}{nt_{0,n}} + \frac{1}{\sqrt{n}}\xi(t_{0,n}) + \frac{1}{\epsilon_n} - \eta\zeta(t_{0,n}) + O_P\left(\max\left[\frac{\eta}{\sqrt{n}}, \eta^2\right]\right) \text{ ,}$$

where

$$\xi(t_{0,n}) = \frac{1}{\sqrt{n}}\sum_{j=l_n+1}^{n} \frac{\mathsf{z}_{j,n}^2-1}{t_{0,n}+\gamma_n+\delta_{j,n}} = O_P(1) \text{ , and } \zeta(t_{0,n}) = \frac{1}{n}\sum_{j=l_n+1}^{n} \frac{1}{(t_{0,n}+\gamma_n+\delta_{j,n})^2} = O_P(1) \text{ .}$$

Let us call

$$t(1) = t_{0,n} + \frac{1}{\sqrt{n}}\frac{\xi(t_{0,n})}{\zeta(t_{0,n})} \text{ .}$$

Our assumptions and the fact that $t_{0,n} \leq \epsilon_n$ guarantee that $\zeta(t_{0,n})$ is bounded below as n becomes large. The expansion above shows that

$$\frac{1}{\epsilon_n} - h_n(t(1)) = O_P(1/n) \text{ .}$$

Because, with probability going to 1, h_n is Lipschitz with Lipschitz constant bounded below in a neighborhood of $t_{0,n}$, we conclude as in 3.6.1 that

$$T = t_{0,n} + \frac{1}{\sqrt{n}}\frac{\xi(t_{0,n})}{\zeta(t_{0,n})} + O_P\left(\frac{1}{n}\right) \text{ ,}$$

which concludes the proof. ■

The phase transition can be further explored in the situation where $\epsilon_n - \epsilon_{0,n}$ is infinitesimal in n but not exactly zero. We are especially concerned in this paper with random variables of the type

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

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_n and the whole spectrum of X_n. 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: in case (1) of the previous theorem, the gap between the two largest eigenvalues of $X_n(\epsilon_n)$ scales like $1/n$, the rate we obtained in our non-asymptotic bounds. However, in many situations, *the gap is much greater than* $1/n$, usually of order at least $1/\sqrt{n}$, and the worst case bound on the Lipschitz constant of $\mathsf{F}_k(X)$ is very conservative.

4. Stochastic composite optimization

In this section, we will develop a variant of the algorithm in [Lan, 2012] which allows for adaptive (monotonic) scaling of the step size parameter. For the sake of completeness, we first recall the key definitions in [Lan, 2012], 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

$$\min_{x\in Q} \Psi(x) := f(x) + h(x), \tag{18}$$

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 we only observe noisy oracles for $f(x)$ and $h(x)$ written

$$f(x,\xi) \quad \text{and} \quad h(x,\xi),$$

for some random variable $\xi \in \mathbb{R}^d$; we write $\Psi(x,\xi) := f(x,\xi) + h(x,\xi)$ with $\Psi(x) = \mathbf{E}[\Psi(x,\xi)]$. We also assume that $\Psi(\cdot,\xi)$ is convex for any $\xi \in \mathbb{R}^d$, and that $\Psi(x,\xi) \geq \Psi(x,0)$ a.s. with

$$\mathbf{E}[\Psi(x^*,\xi)] - \Psi(x^*,0) \leq \mu$$

for some $\mu > 0$ at the optimum of problem (18), with μ typically of order ϵ. The value of μ is typically controlled by the magnitude of the noise ξ. The function $f(x)$ is assumed to be convex with Lipschitz continuous gradient

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

and $h(x)$ is also assumed to be 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

$$\mathbf{E}[G(x,\xi)] = g(x) \in \partial\Psi(x), \tag{19}$$

$$\mathbf{E}[\|G(x,\xi) - g(x)\|_*^2] \leq \sigma^2. \tag{20}$$

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. The function $\omega(x)$ 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:

$$V(x,y) \equiv \omega(y) - [\omega(x) + \nabla\omega(x)^T(y-x)]. \tag{21}$$

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

$$P_x^{Q,\omega}(y) \equiv \operatorname*{argmin}_{z\in Q}\{y^T(z-x) + V(x,z)\}. \tag{22}$$

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

$$D_{\omega,Q} \equiv (\max_{z\in Q} \omega(z) - \min_{z\in Q} \omega(z))^{1/2}, \tag{23}$$

and let

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

which satisfies

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

[Lan, 2012, Corollary 1] implies the following result on the complexity of solving (3) using the AC-SA algorithm in [Lan, 2012, §3].

Proposition 4.1. *Let $N > 0$, and write F_k^* the optimal value of problem (3). Suppose that the sequences X_t, X_t^{md}, X_t^{ag} are computed as in [Lan, 2012, Corollary 1] using the stochastic gradient oracle in (28). After N iterations of the AC-SA algorithm in [Lan, 2012, §3], we have*

$$\mathbf{E}[\mathsf{F}_k(X_{N+1}^{ag}) - \mathsf{F}_k^*] \le \frac{8nC_kD_{\omega,Q}^2}{\epsilon N(N+2)} + \frac{4\sqrt{2}D_{\omega,Q}}{\sqrt{Nq}}. \tag{24}$$

Proof. Using the bound on the variance of the stochastic oracle $G(X, z)$, we know that G satisfies (19) with $\sigma^2 = 1/q$. Section 3 also shows that the Lipschitz constant of the gradient is bounded by C_kn/ϵ. If we pick $\|\cdot\|_F^2/2$ as the prox function, [Lan, 2012, Corollary 1] yields the desired result. ■

Setting $N = 2D_{\omega,Q}\sqrt{n}/\epsilon$ and $q = \max\{1, D_{\omega,Q}/(\epsilon\sqrt{n})\}$ in the convergence bound above will then ensure $\mathbf{E}[\mathsf{F}_k(X_N) - \mathsf{F}_k^*] = O(\epsilon)$. Because our bounds on the Lipschitz constant are usually very conservative, in the section that follows, we detail a version of the AC-SA algorithm with adaptive (but monotonically decreasing) step-size scaling parameter.

4.1. **Stochastic composite optimization with line search.** The algorithm in [Lan, 2012, §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 iteration to determine the optimal step size while guaranteeing convergence. The algorithm in [Lan, 2012] 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, 2012, §3]. Here, we will study the convergence properties of an adaptive variant of the algorithm for stochastic composite optimization in [Lan, 2012, §3], with monotonic line search.

In this section, we first modify the convergence lemma in [Lan, 2012, Lemma 5] to adapt it to the line search strategy detailed in Algorithm 1. Note that our method requires testing the line search exit condition using *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, with $\beta_t = (t+1)/2$. Suppose also that γ and these points satisfy the line search exit condition in line 9, i.e.*

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

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 $\gamma^{min}, \gamma^{max}$, γ^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}.

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

$$\begin{aligned}\beta_t\gamma_t[\Psi(x_{t+1}^{ag}, \xi_{t+1}) - \Psi(x,0)] + V(x_{t+1}, x) &\leq (\beta_t - 1)\gamma_t[\Psi(x_t^{ag}, \xi_t) - \Psi(x,0)] + V(x_t, x) \\ &\quad + \gamma_t(\Psi(x, \xi_t) - \Psi(x,0)) + \frac{4\mathcal{M}^2\gamma_t^2}{\alpha}.\end{aligned}$$

Proof. As in [Lan, 2012, 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

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

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

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

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

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

For any x in the feasible set, we can then use the properties of the prox mapping detailed in [Lan, 2012, 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

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

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 $\beta_t\gamma_t\Psi(x,0)$ from both sides yields the desired result. ■

We are now ready to prove the main convergence result, adapted from [Lan, 2012, 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, 2012, Lemma 5] for the second phase where $\gamma=\gamma^{min}$, writing the switch time T_γ. 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 $\Psi(x^*,0)$ the optimal value of problem (18). 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*

$$\gamma^{max}\leq\frac{\sqrt{6\alpha}D_{\omega,Q}}{(N+2)^{3/2}(4\mathcal{M}^2+\sigma^2)^{1/2}}\quad\text{and}\quad\gamma^{min}=\min\left\{\frac{\alpha}{2L},\gamma^{max}\right\}\tag{25}$$

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

$$\mathbf{E}[\Psi(x_{N+1}^{ag})-\Psi(x^*,0)]\leq\frac{8LD_{\omega,Q}^2}{\alpha N^2}+\frac{8}{N^2\gamma^{min}}\,\mathbf{E}\left[\frac{2(4\mathcal{M}^2+\sigma^2)}{\alpha}\sum_{t=1}^N\gamma_t^2\right]+\frac{(T_\gamma+2)^2\gamma^{max}\mu}{N^2 2\gamma^{min}}\tag{26}$$

and a simpler, but coarser bound is given by

$$\mathbf{E}\left[\Psi(x_{N+1}^{ag})-\Psi(x^*,0)\right]\leq\frac{8LD_{\omega,Q}^2}{\alpha N^2}+\frac{8D_{\omega,Q}\sqrt{4\mathcal{M}^2+\sigma^2}}{\sqrt{N}}\left(\frac{\gamma^{max}}{\gamma^{min}}\rho_N+1-\rho_N\right)+\frac{(T_\gamma+2)^2\gamma^{max}\mu}{N^2 2\gamma^{min}},\tag{27}$$

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

Proof. Lemma 4.2 applied at x^* shows

$$\begin{aligned}\beta_t\gamma_t[\Psi(x_{t+1}^{ag},\xi_{t+1})-\Psi(x^*,0)]+V(x_{t+1},x^*)\ \leq\ & (\beta_t-1)\gamma_t[\Psi(x_t^{ag},\xi_t)-\Psi(x^*,0)]+V(x_t,x^*) \\ & +\frac{4\mathcal{M}^2\gamma_t^2}{\alpha}+\gamma_t(\Psi(x^*,\xi_t)-\Psi(x^*,0))\end{aligned}$$

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

$$\begin{aligned}(\beta_{t+1}-1)\gamma_t[\Psi(x_{t+1}^{ag},\xi_{t+1})-\Psi(x^*,0)]\ \leq\ & \beta_t\gamma_t[\Psi(x_{t+1}^{ag},\xi_{t+1})-\Psi(x^*,0)] \\ \leq\ & (\beta_t-1)\gamma_t[\Psi(x_t^{ag},\xi_t)-\Psi(x^*,0)]+\frac{4\mathcal{M}^2\gamma_t^2}{\alpha} \\ & +\gamma_t(\Psi(x^*,\xi_t)-\Psi(x^*,0))+V(x_t,x^*)-V(x_{t+1},x^*)\end{aligned}$$

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

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

using again $\Psi(x^*,\xi_t) - \Psi(x^*,0) \geq 0$ a.s. When the line search fails, $\gamma_t = \gamma^{min}(t+1)/2$ is deterministic and [Lan, 2012, Lem. 5 & Th. 2] show that

$$(\beta_{t+1}-1)\gamma_t[\Psi(x_{t+1}^{ag}) - \Psi(x^*,0)] \leq (\beta_t - 1)\gamma_t[\Psi(x_t^{ag}) - \Psi(x^*,0)] + V(x_t,x^*) - V(x_{t+1},x^*) + \Delta(x^*)$$

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 $t = T_\gamma + 1$ the iteration where the line search first fails. Combining these last results, using $\beta_1 = 1$, we obtain

$$\begin{aligned}
& (\beta_{N+1}-1)\gamma_N\, \mathbf{E}[\Psi(x_{N+1}^{ag}) - \Psi(x^*,0)] \\
\leq\ & D_{\omega,Q}^2 + \sum_{t=1}^{T_\gamma} \mathbf{E}\left[\frac{4\mathcal{M}^2\gamma_t^2}{\alpha}\right] + \sum_{t=1}^{T_\gamma}\frac{\gamma^{max}(t+1)}{2}\mu + (\beta_{T_\gamma+1}-1)\gamma_{T_\gamma}\,\mathbf{E}[\Psi(x_{T_\gamma+1}^{ag}) - \Psi(x_{T_\gamma+1}^{ag},\xi_{T_\gamma+1})] \\
& + \sum_{T_\gamma+1}^{N} \mathbf{E}\left[\gamma_t\langle \delta_t, x^* - x_t\rangle + \frac{2(4\mathcal{M}^2 + \|\delta_t\|_*^2)\gamma_t^2}{\alpha}\right] \\
\leq\ & D_{\omega,Q}^2 + \sum_{t=1}^{T_\gamma} \mathbf{E}\left[\frac{4\mathcal{M}^2\gamma_t^2}{\alpha}\right] + \sum_{T_\gamma+1}^{N} \mathbf{E}\left[\frac{2(4\mathcal{M}^2 + \|\delta_t\|_*^2)\gamma_t^2}{\alpha}\right] + \frac{(T_\gamma+2)^2\gamma^{max}\mu}{4} \\
\leq\ & D_{\omega,Q}^2 + \mathbf{E}\left[\frac{2(4\mathcal{M}^2+\sigma^2)}{\alpha}\sum_{t=1}^{N}\gamma_t^2\right] + \frac{(T_\gamma+2)^2\gamma^{max}\mu}{4}
\end{aligned}$$

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

4.2. **Stochastic composite optimization for semidefinite optimization.** We can use the results above to solve problem (3). In this case,

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

and by construction $\Psi(X,z) \geq \Psi(X,0) = \lambda_{\max}(X)$. Recall, that with this choice of oracle, Section 2 shows

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

so $\mu = k\epsilon$ in Proposition 4.3. We use the following gradient oracle

$$G(X,z) = \frac{1}{q}\sum_{l=1}^{q}\phi_l\phi_l^T \tag{28}$$

where each ϕ_l is a leading eigenvector of the matrix $X + \frac{\epsilon}{n}z_{i_0}z_{i_0}^T$, with

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

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) and q is used to control the variance. The Lipschitz constant of the gradient is bounded by (6) with

$$L \leq \frac{n}{\epsilon} \frac{k}{(k-2)}$$

and the variance of the gradient oracle is bounded by $1/q$ with $\mathcal{M} = 0$ in the results above.

5. 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

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}(A + X) \\ \text{subject to} & -\rho \leq X_{ij} \leq \rho, \quad \text{for } i, j = 1, \ldots, n \end{array} \tag{29}$$

which is a semidefinite program in the matrix $X \in \mathbf{S}_n$. Since randomly generated matrices A have a 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 change the problem dimension n (we select the n genes with the highest variance) and normalizing the matrix A so that its spectral norm is one. We set $\rho = \max\{\mathbf{diag}(A)\}/2$ in (29).

We also test performance on the classical *MaxCut* relaxation. The primal semidefinite program is written

$$\begin{array}{ll} \text{maximize} & \mathbf{Tr}(CX) \\ \text{subject to} & \mathbf{diag}(X) = 1, \; X \succeq 0, \end{array}$$

in the variable $X \in \mathbf{S}_n$. The objective matrix C is sampled from the Wishart distribution with $C = G^T G / \|G\|_2^2$ where G is a standard Gaussian matrix. Here, we solve the dual, written

$$\min \; \lambda_{\max}(C + \mathbf{diag}(w)) - \mathbf{1}^T w \tag{30}$$

in the variable $w \in \mathbb{R}^n$. The problem is unconstrained, and we add a bound on the Euclidean norm of the vector w. The prox function used in both examples (where the feasible sets are an hypercube and an Euclidean ball) is the square Euclidean norm, which means that the prox is simply an Euclidean projection of the matrix X in (29) (with the projection taken elementwise) and of the vector w in (30).

We first compare the performance of Algorithm 1 with that of the corresponding deterministic algorithms, ACSA as detailed in Lan [2012] and the accelerated first-order method (with line search) in [Nesterov, 2007b, §4] after smoothing problem (29) 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 $\epsilon = 5 \times 10^{-2}$, $q = 0.1/\epsilon$, $k = 3$ and the maximum number of iterations to $O(\sqrt{n})$ in the stochastic algorithm. In line with the discussion of Section 3.6, we scale down the Lipschitz constant by a factor 100 in both stochastic and deterministic algorithms. This significantly speeds up the algorithms with no apparent effect on convergence, thus confirming that the worst case bounds are indeed somewhat conservative.

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 Tables 2 and 3 for DSPCA (29) and *MaxCut* respectively. We observe that, for the DSPCA tests, the total number of eigenvectors computed is significantly lower, while the number of iterations is much higher for the stochastic code. The tradeoff is much less favorable for the *MaxCut* experiments. In Figure 2 we plot the objective value reached as a function of the number of eigenvectors computed for both experiments, when $n = 1000$. We again see that the behavior of the stochastic algorithm is much better for DSPCA than for *MaxCut*. In Figure 4, we plot the spectrum of the solution matrices for both problems. We notice that the leading eigenvalues are much more

separated in the DSPCA problem which at least partly explains the difference in performance. More importantly, the deterministic implementation of the ACSA algorithm in [Lan, 2012] seems to be significantly slower than that of the smooth algorithm in [Nesterov, 2007a]. Improving the numerical performance of the ACSA algorithm itself thus seems to be the key to a competitive implementation of the results detailed here.

n	Stoch. # iters.	Stoch. # eigvs.	ACSA # iters.	ACSA # eigvs.	Det. # iters.	Det. # eigvs.
50	707	1266	51	2550	16	3700
100	1000	1806	50	5000	12	5800
200	1414	2532	55	11000	28	24800
500	2236	8016	60	30000	12	29000
1000	3162	18990	65	65000	12	56000
2000	4472	21444	66	132000	14	132000

TABLE 2. Number of iterations and total number of eigenvectors computed by Algorithm 1 (Stoch.), the ACSA algorithm in Lan [2012] and the algorithm in [Nesterov, 2007b, §4] (Det.) (both with exponential smoohting) to reach identical objective values when solving the DSPCA relaxation in (29).

n	Stoch. # iters.	Stoch. # eigvs.	ACSA # iters.	ACSA # eigvs.	Det. # iters.	Det. # eigvs.
50	3536	9534	217	10850	2	400
100	5000	30024	353	35300	4	1600
200	7071	42438	537	107400	6	4400
500	11180	67086	545	272500	6	9000
1000	15811	94872	601	601000	6	16000
2000	22361	134178	377	754000	4	20000

TABLE 3. Number of iterations and total number of eigenvectors computed by Algorithm 1 (Stoch.), the ACSA algorithm in Lan [2012] and the algorithm in [Nesterov, 2007b, §4] (Det.) (both with exponential smoohting) to reach identical objective values when solving the *MaxCut* relaxation in (30).

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, which is $O(qn^2 \log n)$. The cost of each gradient computation in [Nesterov, 2007b, §4] is dominated by the cost of computing a matrix exponential, which is $O(n^3)$. This means that the ratio between these costs grows as $O(n/(q \log n))$.

In Figure 3 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 (29) 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).

6. APPENDIX

In this Appendix, we recall several useful results related to the algorithm presented here. Subsection 6.1 summarizes the complexity of computing *one* leading eigenvector of a symmetric matrix (versus computing the entire spectrum). In Subsection 6.2, we prove a number of technical results concerning the function F_k and its components. In particular, we prove Theorem 3.2 linking the local Lipschitz constant of the gradient

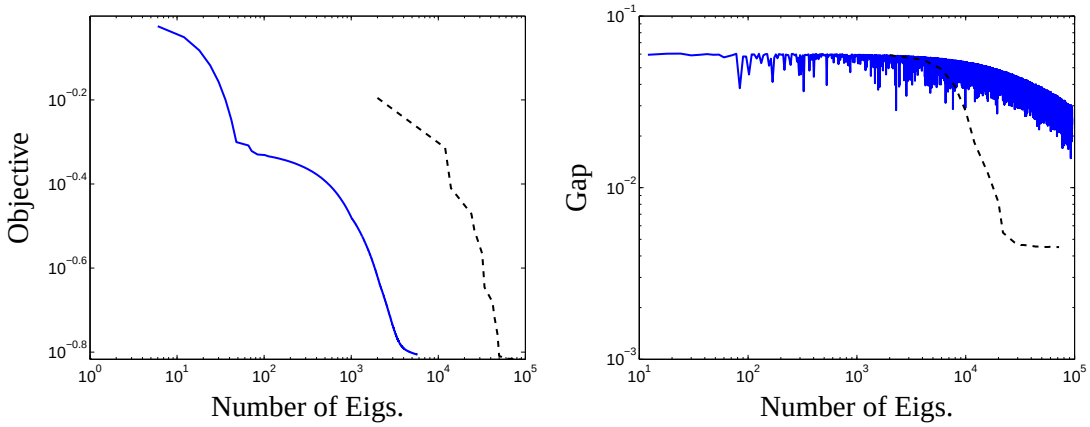

FIGURE 2. Objective value (or gap) versus number of eigenvectors computed by Algorithm 1 (solid blue) and the algorithm in [Nesterov, 2007b, §4] (dashed black) for the DSPCA *(left)* and *MaxCut (right)* relaxations.

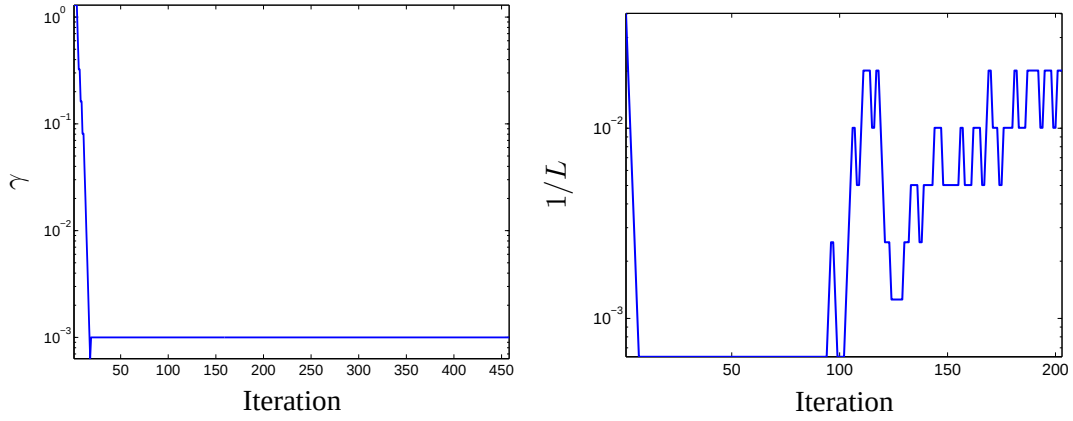

FIGURE 3. Line search parameters γ for the stochastic algorithm *(left)* together with the values of the inverse of the Lipschitz constant L used in the deterministic smoothing algorithm *(right)*.

and the spectral gap. Finally, we show in Subsection 6.3 how the secular equation can be generalized to perturbations of higher rank, and we discuss extensions of our smoothing argument using GUE matrices.

6.1. **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

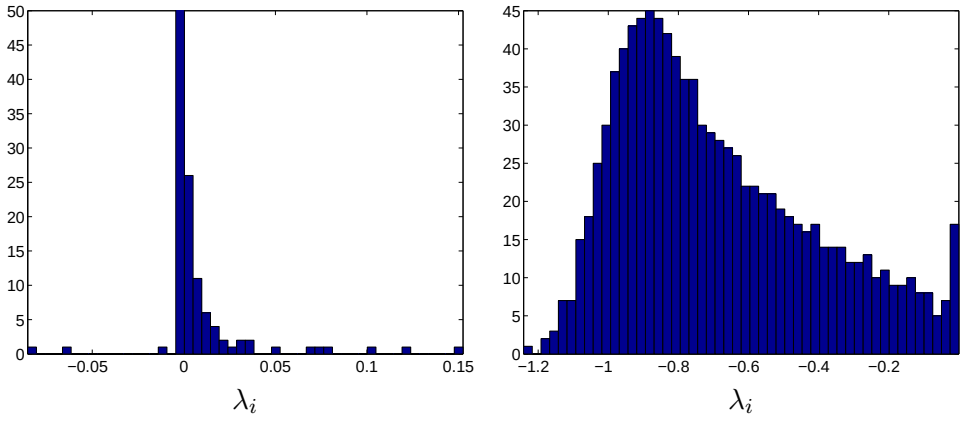

FIGURE 4. Histogram of eigenvalues for the matrix solutions to the sparse PCA *(left)* and *MaxCut (right)* problems for $n = 1000$. For clarity, the graph on the left is truncated above 50.

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.

6.1.1. *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.

6.1.2. *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, i.e. such that $|\lambda - \lambda_{\max}| \leq \epsilon\lambda_{\max}$, 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 compute efficiently 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 $T_k \in \mathbf{S}_k$.

6.2. Technical results concerning F_k and its components.

6.2.1. *General remarks on rotational invariance.* We use repeatedly in the paper the fact that the type of smoothing we devised has some rotational invariance properties, which allows us to perform our computations on diagonal matrices without losing generality. We summarize the results we need and use in the following statement.

Lemma 6.1. *Let X be a deterministic matrix in $\mathbf{S}_n$. X is diagonalizable in an orthonormal basis and we write $X = \mathcal{O}_X^T D_X \mathcal{O}_X$, where D_X is a diagonal matrix containing the eigenvalues of X and $\mathcal{O}_X$ is a matrix of eigenvectors of X. Let $\{z_i\}_{i=1}^k$ be k i.i.d $\mathcal{N}(0, \mathbf{I}_n)$ random vectors. Let ν be in $\mathbb{R}$ and call*

$$F_k(X) = \max_{1 \le i \le k} \lambda_{\max}(X + \nu z_i z_i^T) \ .$$

Then

$$F_k(X) \stackrel{\mathcal{L}}{=} F_k(D_X) \ .$$

Furthermore, if $\phi[F_k(X)]$ is an eigenvector associated with $F_k(X)$, we have

$$\phi[F_k(X)] \stackrel{\mathcal{L}}{=} \mathcal{O}_X^T \phi[F_k(D_X)] \ .$$

Proof. We observe that

$$X + \nu z_i z_i^T = \mathcal{O}_X^T \left[D_X + \nu (\mathcal{O}_X z_i)(\mathcal{O}_X z_i)^T\right] \mathcal{O}_X \ .$$

Now it is a standard property of the normal distribution that if $\{z_i\}_{i=1}^k$ are i.i.d $\mathcal{N}(0, \mathbf{I}_n)$, then $\{\mathcal{O}_X z_i\}_{i=1}^k$ are i.i.d $\mathcal{N}(0, \mathbf{I}_n)$, for any (deterministic) orthonormal matrix $\mathcal{O}_X$. The results we announced follow immediately. ■

6.2.2. *Existence of a density for T.* In Lemma 3.4, we were interested in $T = \lambda_{\max}(X + \epsilon/nzz^T) - \lambda_{\max}(X)$. We prove Lemma 3.4 here, showing that T has a density on $[0, \infty)$ when $z \sim \mathcal{N}(0, \mathbf{I}_n)$.

Proof. [of Lemma 3.4] 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, since then X is proportional to $\mathbf{I}_n$). By rotational invariance of the standard Gaussian distribution, we can and do assume that X is diagonal in what follows (see Lemma 6.1 for details, if needed). As we have seen before, T is therefore the only 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 when $t > 0$. Hence, for any given $t > 0$,

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

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. As a matter of fact, recall that if X has density p and $t > 0$, X/t has density $tp(t\cdot)$. Recall also that a random variable with χ_l^2 distribution has density (see e.g Mardia et al. [1979], p. 487)

$$p_l(x) = \frac{2^{-l/2}}{\Gamma(l/2)} x^{l/2-1} \exp(-x/2) \ 1_{x \in (0,\infty)} \ .$$

So it is clear that for any l, any $x > 0$, any $t > 0$, and any $\alpha \geq 0$, $t \to (t + \alpha) p_l((t + \alpha)x)$ 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 π such that $\pi(t) = P(T \geq t)$ 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 $[0, \infty)$. ■

6.2.3. *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 the local Lipschitz constant. To do so, we compute

$$\gamma(X, Y) = \lim_{t \to 0} \partial^2 F_0(X + tY)/\partial t^2 \ ,$$

where $\|Y\|_F = 1$, and Y is symmetric. It is standard that the local Lipschitz constant - with respect to Frobenius norm - of ∇F_0 is

$$L\left[\nabla F_0(X)\right] = \sup_{Y \in \mathbf{S}_n : \|Y\|_F = 1} \gamma(X, Y) \ .$$

Let us call $\lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n$ the ordered eigenvalues of X. Very importantly we assume that λ_1 has multiplicity one. If not, it is easy to see that the function $\lambda_{\max}(X)$ is continuous but not differentiable. We refer the reader to [Kato, 1995; Overton and Womersley, 1995; Lewis and Sendov, 2002] for a more complete discussion. Recall that in this situation Theorem 3.2 stated that

$$L\left[\nabla F_0(X)\right] = \frac{1}{\lambda_{\max}(X) - \lambda_2(X)}. \tag{31}$$

We now prove this statement.

Proof. [of Theorem 3.2] The strategy is to first exhibit a matrix Y_c in $\mathbf{S}_n$ that will give us the right-hand side of Equation (31) 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] or [Lewis and Sendov, 2002] to get, for small t

$$F_0(X+tY) = \lambda_{\max}(X) + t\phi_1^T Y \phi_1 + t^2 \sum_{j=2}^{n} \frac{1}{\lambda_1(X) - \lambda_j(X)} (\phi_1^T Y \phi_j)^2 + o(t^2) \ ,$$

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

$$\gamma(X,Y) = \lim_{t\to 0} \frac{\partial^2 F_0(X+tY)}{\partial t^2} = 2 \sum_{j=2}^{n} \frac{1}{\lambda_1(X) - \lambda_j(X)} (\phi_1^T Y \phi_j)^2 \ , \tag{32}$$

Finding a lower bound for $L\left[\nabla F_0(X)\right]$. Let $\mathcal{O}$ 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, $\mathcal{O}e_i = \phi_i$ and hence $\mathcal{O}^T\phi_i = e_i$. 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

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

Note that $Y_c \in \mathbf{S}_n$. Since $\mathcal{O}^T\phi_i = e_i$, 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, $\|Y_c\|_F^2 = \mathbf{Tr}\, Y_c^T Y_c = \mathbf{Tr}\, Y_c^2 = \mathbf{Tr}\, \mathcal{O} P_0^2 \mathcal{O}^T/2 = \|P_0\|_F^2/2 = 1$. Now, $\phi_1^T Y_c \phi_j = \delta_{2,j}\|\phi_1\|^2/\sqrt{2}$. Hence,

$$\gamma(X, Y_c) = \lim_{t\to 0} \frac{\partial^2 F_0(X+tY_c)}{\partial t^2} = \frac{2}{2} \frac{1}{\lambda_1(X) - \lambda_2(X)} \ ,$$

and therefore,

$$L\left[\nabla F_0(X)\right] \geq \frac{1}{\lambda_1(X) - \lambda_2(X)} \ .$$

Finding an upper bound for $L\left[\nabla F_0(X)\right]$. 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

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

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 $(\phi_1^TY\phi_1)^2=\tilde{y}_{1,1}^2$. Hence,

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

We conclude that

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

and therefore

$$L\left[\nabla F_0(X)\right]=\sup_{Y\in\mathbf{S}_n,\|Y\|_F=1}\gamma(X,Y)\leq\frac{1}{\lambda_1(X)-\lambda_2(X)}\ .$$

Since we have matching upper and lower bounds for $L\left[\nabla F_0(X)\right]$, we have established Theorem 3.2. ■

6.2.4. *Differentials of maximum of several differentiable functions.* We need the following elementary and well-known results at several points in the paper. We put them in this Appendix for the convenience of the reader.

Lemma 6.2. *Consider the function $\Psi_k=\max_{1\leq i\leq k}\psi_k$, where $\psi_1,\ldots,\psi_k$ are Gâteaux-differentiable functions from $\mathcal{D}\subset\mathbb{R}^d$ to $\mathbb{R}$ and k is an integer. Let int$(\mathcal{D})$ be the interior of $\mathcal{D}$. Let $x_0\in$ int$(\mathcal{D})$ be such that there exists $i_0\in\{1,\ldots,k\}$ such that $\psi_{i_0}(x_0)>\psi_j(x_0)$ for all $j\neq i_0$. Then, Ψ_k is Gâteaux-differentiable at x_0 with*

$$\nabla_G\Psi_k(x_0)=\nabla_G\psi_{i_0}(x_0)\ .$$

Furthermore, when ψ_j's are Fréchet-differentiable, so is Ψ_k at x_0.

The proof shows that the result extends to higher order derivatives when they exist.

Proof. This is simply a restatement of the results of Proposition 7.2.7 in [Schirotzek, 2007], or [Hiriart-Urruty and Lemaréchal, 2001] Theorem 4.4.2 and Corollary 4.4.4. We give the key idea and a proof of this easy fact for the sake of completeness.

Indeed, let I_j be the set of points y such that $\psi_j(y)\geq\psi_l(y)$ for all $l\neq j$. We call 1_{I_j} the function taking value 1 on I_j and 0 elsewhere. Let $N(x)$ be equal to card $\{j,1\leq j\leq k:\psi_j(x)=\Psi_k(x)\}$. Note that $N(x)=\sum_{j=1}^k1_{I_j}(x)$. It is clear that $1\leq N(x)\leq k$. We also have

$$\Psi_k(x)=\frac{\sum_{j=1}^k\psi_j(x)1_{I_j}(x)}{N(x)}\ .$$

Under our assumptions on x_0, it is clear that $N(x_0)=1$. Furthermore, in a neighborhood $V(x_0)$ of x_0, we have $N(x)=1$ by continuity of the functions ψ_j's. Of course, $V(x_0)$ is open, by definition of a neighborhood. It follows that for all x in $V(x_0)$, we have $\Psi_k(x)=\psi_{i_0}(x)$. It now follows from the definition of Gâteaux-differentiability that Ψ_k is Gâteaux-differentiable at x_0 with the same Gâteaux-differential as ψ_{i_0}. The result in the case of Fréchet differentiable functions ψ_j's holds for the same reasons and is established by the same analysis. ■

We have the following corollary.

Lemma 6.3. *Suppose $X \in \mathbf{S}_n$ and $v_1, \ldots, v_k$ are vectors in $\mathbb{R}^n$. Denote by $F_k(X) = \max_{1 \le i \le k} \lambda_{\max}(X + v_i v_i^T)$. Suppose that X and $\{v_i\}_{i=1}^k$ are such that there exists a unique i_0 such that*

$$\lambda_{\max}(X + v_{i_0} v_{i_0}^T) = F_k(X) \ .$$

Suppose further that the largest eigenvalue of $X + v_{i_0} v_{i_0}^T$ has multiplicity one. Then

$$L\left[\nabla F_k(X)\right] = \frac{1}{\lambda_{\max}(X + v_{i_0} v_{i_0}^T) - \lambda_2(X + v_{i_0} v_{i_0}^T)} \ .$$

It follows that for $i_0 = \operatorname{argmax}_{1 \le i \le k} \lambda_{\max}(X + v_i v_i^T)$, if $F_k(X) \neq \lambda_{\max}(X)$,

$$L\left[\nabla F_k(X)\right] \le \frac{1}{\lambda_{\max}(X + v_{i_0} v_{i_0}^T) - \lambda_{\max}(X)} = \frac{1}{F_k(X) - \lambda_{\max}(X)} \ .$$

Proof. The proof of Lemma 6.2 shows that under our assumptions, F_k coincides locally with $\lambda_{\max}(X + v_{i_0} v_{i_0}^T)$. Hence the local Lipschitz constant of ∇F_k is the same as that of $X \mapsto \lambda_{\max}(X + v_{i_0} v_{i_0}^T)$. One of our assumptions is that the largest eigenvalue of $X + v_{i_0} v_{i_0}^T$ has multiplicity 1. In that situation, Theorem 3.2 guarantees that the local Lipschitz constant of $X \mapsto \lambda_{\max}(X + v_{i_0} v_{i_0}^T)$ is

$$\frac{1}{\lambda_{\max}(X + v_{i_0} v_{i_0}^T) - \lambda_2(X + v_{i_0} v_{i_0}^T)} \ .$$

So we have established that

$$L\left[F_k(X)\right] = \frac{1}{\lambda_{\max}(X + v_{i_0} v_{i_0}^T) - \lambda_2(X + v_{i_0} v_{i_0}^T)} \ .$$

We now recall that by Cauchy's interlacing theorem (Theorem 4.3.4 in Horn and Johnson [1990]), $\lambda_2(X + v_{i_0} v_{i_0}^T) \le \lambda_{\max}(X)$. We can therefore conclude that

$$L\left[F_k(X)\right] = \frac{1}{\lambda_{\max}(X + v_{i_0} v_{i_0}^T) - \lambda_{\max}(X)} \ ,$$

since we have assumed that $\lambda_{\max}(X) \neq \lambda_{\max}(X + v_{i_0} v_{i_0}^T) = F_k(X)$. ■

6.2.5. *Interchanging expectation and differentiation for F_k.*

Lemma 6.4. *We can interchange expectation and differentiation for F_k so that*

$$\nabla \mathsf{F}_k(X) = \mathbf{E}\left[\phi_{i_0} \phi_{i_0}^T\right]$$

using the notation of Lemma 3.8.

Proof. F_k is convex as an average of convex functions. To show that it is differentiable, it is therefore enough to show that it is Gâteaux-differentiable (see Hiriart-Urruty and Lemaréchal [2001], Corollary D.2.1.4). Let X_0 be given. We use the notation

$$\mathsf{F}_k(X_0) = \int F_k(X_0; z) \mu(dz)$$

to make things simpler in this proof. Of course, $F_k(X_0; z) = \max_{1 \le i \le k} \lambda_{\max}(X_0 + \frac{\epsilon}{n} z_i z_i^T)$. (Compared to the main text, we have now made the dependence on z explicit as it is needed below to address a potential measure theoretic problem.) $\mu(dz)$ is just the joint distribution of z_i's, for $i = 1, \ldots, k$. To make notations simple in this proof, we use z to denote $(z_i)_{i=1}^k$.
We know that $F_k(X_0; z)$ has a subdifferential for all X_0 and all z, since it is the maximum of k functions with a subdifferential (see Hiriart-Urruty and Lemaréchal [2001], Theorem D.4.4.2). The spectral norm of the elements of this subdifferential is bounded by 1, since they are a convex combination of matrices of

spectral norm at most 1 (see Hiriart-Urruty and Lemaréchal [2001], Theorem D.4.4.2 and Equation (5.1.3) p. 195 in that book, which characterizes the subdifferential of the largest eigenvalue mapping of a symmetric matrix).
Suppose Y_0 is a fixed matrix, with $\|Y_0\|_2 = 1$ without loss of generality, where $\|Y\|_2$ is the spectral norm of the symmetric matrix Y. By the mean value theorem for functions with a subdifferential (see Hiriart-Urruty and Lemaréchal [2001], Theorem D.2.3.4) we have, if we call $X_{0,t} = X_0 + tY_0$

$$F_k(X_0 + tY_0; z) - F_k(X_0; z) = t \int_0^1 < \partial F_k(X_{0,tu}; z), Y_0 > du \ ,$$

where $\partial F_k(X_{0,tu}; z)$ is any choice of subgradient of $F_k(X_{0,tu}; z)$ and $< A, B >= \mathbf{Tr}\, A^T B$ for the symmetric matrices we are working with. (We use the same notation in this proof for subgradients and subdifferentials since it does not create confusion.)
Because $\|\partial F_k(X_{0,tu})\|_2 \leq 1$, we can apply Fubini's theorem to get

$$\frac{\mathsf{F}_k(X_0 + tY_0) - \mathsf{F}_k(X_0)}{t} = \int_0^1 du \int < \partial F_k(X_{0,tu}; z), Y_0 > \mu(dz) \ .$$

Now let $\eta > 0$ be given and let

$$\begin{aligned}
\mathcal{A}_\eta &= \{z : \ \forall i \ , 1 \leq i \leq k \ , \lambda_{\max}(X + \frac{\epsilon}{n} z_i z_i^T) - \lambda_{\max}(X) > \eta\} \ , \\
\mathcal{B}_\eta &= \{z : \ \exists i_0 : \forall j \neq i_0 \ , 1 \leq j \leq k \ , \lambda_{\max}(X + \frac{\epsilon}{n} z_{i_0} z_{i_0}^T) \geq \lambda_{\max}(X + \frac{\epsilon}{n} z_j z_j^T) + \eta\} \ , \\
\mathcal{E}_\eta &= \mathcal{A}_\eta \cap \mathcal{B}_\eta \ .
\end{aligned}$$

By continuity of the maps involved, $\mathcal{E}_\eta$ is clearly measurable with respect to Lebesgue measure and therefore μ. Equation (10) in Proposition 3.3 implies that $\lim_{\eta \to 0} \mu(\mathcal{A}_\eta) = 1$. Lemma 3.4 also implies that $\lim_{\eta \to 0} \mu(\mathcal{B}_\eta) = 1$. We conclude that $\lim_{\eta \to 0} \mu(\mathcal{E}_\eta) = 1$.

This implies that

$$\left| \int < \partial F_k(X_{0,tu}; z), Y_0 > \mu(dz) - \int_{\mathcal{E}_\eta} < \partial F_k(X_{0,tu}; z), Y_0 > \mu(dz) \right| \leq n\mu(\mathcal{E}_\eta^c) \to 0 \text{ as } \eta \to 0 \ ,$$

since the absolute value of the integrand is bounded by n. For the same reasons, as $\eta \to 0$,

$$\left| \int_0^1 du \int < \partial F_k(X_{0,tu}; z), Y_0 > \mu(dz) - \int_0^1 du \int_{\mathcal{E}_\eta} < \partial F_k(X_{0,tu}; z), Y_0 > \mu(dz) \right| \leq n\mu(\mathcal{E}_\eta^c) \to 0.$$

When $|t| < \eta/4$, it is clear that for all $z \in \mathcal{A}_\eta$, and all $u \in [0, 1]$, we have

$$\forall i \ , 1 \leq i \leq k \ , \lambda_{\max}(X + tuY_0 + \frac{\epsilon}{n} z_i z_i^T) - \lambda_{\max}(X + tuY_0) > \eta/2 \ .$$

As matter of fact, we have

$$\begin{aligned}
\lambda_{\max}(X + tuY_0 + \frac{\epsilon}{n} z_i z_i^T) - \lambda_{\max}(X + tuY_0) = {} & \lambda_{\max}(X + tuY_0 + \frac{\epsilon}{n} z_i z_i^T) - \lambda_{\max}(X + \frac{\epsilon}{n} z_i z_i^T) \\
& + \lambda_{\max}(X + \frac{\epsilon}{n} z_i z_i^T) - \lambda_{\max}(X) \\
& + \lambda_{\max}(X) - \lambda_{\max}(X + tuY_0)
\end{aligned}$$

By Weyl's inequality, $|\lambda_{\max}(X + tuY_0 + \frac{\epsilon}{n} z_i z_i^T) - \lambda_{\max}(X + \frac{\epsilon}{n} z_i z_i^T)| \leq |t|u\|Y_0\| \leq \eta/4$ and $|\lambda_{\max}(X) - \lambda_{\max}(X + tuY_0)| \leq \eta/4$ for the same reason. By the same reasoning, if $z \in \mathcal{B}_\eta$, when $|t| < \eta/4$, for all $u \in [0, 1]$,

$$\lambda_{\max}(X + tuY_0 + \frac{\epsilon}{n} z_{i_0} z_{i_0}^T) \geq \lambda_{\max}(X + tuY_0 + \frac{\epsilon}{n} z_j z_j^T) + \frac{\eta}{2} \ .$$

This shows that for $z \in \mathcal{E}_\eta$, when $|t| < \eta/4$, $F_k(X_{0,tu}; z)$ is differentiable for all u and so its subdifferential is reduced to a singleton. Therefore,

$$\forall z \in \mathcal{E}_\eta \, , 0 \leq u \leq 1, \;\; \text{and } |t| < \eta/4 \; , \;\; \partial F_k(X_{0,tu}; z) = \nabla F_k(X_{0,tu}; z) \; .$$

We know in fact that under the aforementioned conditions $F_k(X_{0,tu}; z)$ is twice differentiable as a function of tu (see Kato [1995], pp.80-81) and therefore the gradient $\nabla F_k(X_{0,tu}; z)$ is continuous (as a function of tu). So we have, pointwise in $z \in \mathcal{E}_\eta$ and $u \in [0, 1]$,

$$\lim_{t \to 0} \nabla F_k(X_{0,tu}; z) = \nabla F_k(X_0; z) \; .$$

Using the fact that $\nabla F_k(X_{0,tu}; z)$ is bounded (it is a rank 1 matrix of norm 1), we can use the dominated convergence theorem and Fubini's theorem to conclude that

$$\begin{aligned} \lim_{t \to 0} \int_0^1 du \int_{\mathcal{E}_\eta} < \partial F_k(X_{0,tu}; z), Y_0 > \mu(dz) &= \lim_{t \to 0} \int_0^1 du \int_{\mathcal{E}_\eta} < \nabla F_k(X_{0,tu}; z), Y_0 > \mu(dz) \\ &= \int_{\mathcal{E}_\eta} < \nabla F_k(X_0; z), Y_0 > \mu(dz) \; . \end{aligned}$$

Finally, because $< \partial F_k(X_0; z), Y_0 >$ is bounded and because $\mathcal{E}_\eta$ is a decreasing family of sets, we see that

$$\lim_{\eta \to 0} \int_{\mathcal{E}_\eta} < \nabla F_k(X_0; z), Y_0 > \mu(dz) = \int < \partial F_k(X_0; z), Y_0 > \mu(dz) \; ,$$

by the dominated convergence theorem. Naturally, the previous equality is true for any choice of subgradients on the set of (μ-)measure 0 where the subdifferential $F_k(X_0; z)$ is not reduced to a singleton. Furthermore,

$$\lim_{t \to 0} \frac{\mathsf{F}_k(X_0 + tY_0) - \mathsf{F}_k(X_0)}{t} =< \int \partial F_k(X_0; z) \mu(dz), Y_0 > \; .$$

So we conclude that F_k is Gâteaux-differentiable at X_0 and hence differentiable, since F_k is convex. The previous expression is valid for any subgradient of $F_k(X_0; z)$. Since with probability 1 the subdifferential is a singleton, we can also write

$$\nabla \mathsf{F}_k(X_0) = \int \nabla (F_k(X_0; z)) \mu(dz)$$

In the particular situation we are considering, this can also be re-written as

$$\nabla \mathsf{F}_k(X_0) = \mathbf{E} \left[\phi_{i_0} \phi_{i_0}^T \right] \; .$$

which is the desired result. ■

We now show that the gradient is diagonalizable in the same basis as X.

Lemma 6.5. *The matrix $\nabla \mathsf{F}_k(X)$ is diagonalizable in the same basis as X. In particular, when X is diagonal, so is $\nabla \mathsf{F}_k$.*

Proof. Call $\lambda_i(X)$ the eigenvalues of X in decreasing order. As above, $z_i \overset{i.i.d}{\sim} \mathcal{N}(0, \mathbf{I}_n)$ implies that no z_i is an eigenvector of X with probability one. We call $l_{1,i} = \lambda_{\max}(X + \frac{\epsilon}{n} z_i z_i^T)$ and ϕ_i the corresponding eigenvector. With probability 1, ϕ_i is uniquely defined, up to sign, since $l_{1,i}$ has multiplicity 1 with probability 1.

(i) X diagonal. We first focus on the case where X is diagonal. Our strategy is to show that the off-diagonal entries of $\phi_{i_0} \phi_{i_0}^T$ have a (marginal) distribution that is symmetric around 0.

In what follows we use the notation $z_i(j)$ to denote the j-th coordinate of the vector z_i. It is well-known ([see Golub and Van Loan, 1990, §8.5.3], Theorem 8.5.3) that when X is diagonal, the j-th coordinate of ϕ_i is given by

$$\phi_i(j) = c\frac{z_i(j)}{l_{1,i} - \lambda_j}, \tag{33}$$

where $c > 0$ is a normalizing factor. Recall that $l_{1,i}$ is the largest root of $\chi(\lambda) = 0$, where

$$\chi(\lambda) = 1 + \frac{n}{\epsilon}\frac{\sum_{j=1}^{l}[z_i(j)]^2}{\lambda_1(X) - \lambda} + \frac{n}{\epsilon}\sum_{j=l+1}^{n}\frac{[z_i(j)]^2}{\lambda_j(X) - \lambda} .$$

This equation shows in particular that $l_{1,i}$'s depend on z_i's only through the absolute values of the coordinates of these vectors. Let us now pick j_0, an integer such that $1 \leq j_0 \leq n$. Suppose that we change the j_0-th coordinates of the vectors z_i's to their opposites. Call $\tilde{l}_{1,i}$ and $\tilde{\phi}_i$ the corresponding eigenvalue and eigenvectors. As we have just seen,

$$\tilde{l}_{1,i} = l_{1,i} \ , \forall i \ .$$

In particular, i_0 is unaffected by this sign change operation.
On the other hand,

$$\forall i \ , \quad \left[\tilde{\phi}_i\tilde{\phi}_i^T\right](l,m) = \begin{cases} -\left[\phi_i\phi_i^T\right](l,m) & \text{if } l \neq m \text{ and } m = j_0 \text{ or } l = j_0 \ , \\ \left[\phi_i\phi_i^T\right](l,m) & \text{otherwise .} \end{cases}$$

However, since z_i's have a symmetric distribution, their distribution is unaffected by a change of sign to one of the coefficients. So it is clear that

$$\forall i \ , \quad \tilde{\phi}_i\tilde{\phi}_i^T \stackrel{\mathcal{L}}{=} \phi_i\phi_i^T \ .$$

So for all i, the distribution of the off-diagonal entries of the matrix $\phi_i\phi_i^T$ is symmetric around 0, since it is equal in law to its opposite. (We have just shown it for the off-diagonal entries for the j_0-th row and columns, but since there was nothing special about j_0, it is true for all the off-diagonal entries.) Furthermore, since the value of i_0 is unaffected by the sign change operation we discussed, we have shown that the off-diagonal entries of the matrix $\phi_{i_0}\phi_{i_0}^T$ have a symmetric distribution. Since $\phi_{i_0}^T\phi_{i_0} = 1$, the entries of the matrix $\phi_{i_0}\phi_{i_0}^T$ are bounded and therefore have a mean. This mean must be zero for the off-diagonal entries since they have a symmetric distribution. So we have shown that $\mathbf{E}\left[\phi_{i_0}\phi_{i_0}^T\right]$ is a diagonal matrix when X is diagonal.

(ii) X **not diagonal.** When X is not diagonal, we simply diagonalize X into $X = \mathcal{O}_X^T D_X \mathcal{O}_X$ and use rotational invariance of the distribution of the z_i's to see that

$$\left[\phi_{i_0}(X)\phi_{i_0}(X)^T\right] \stackrel{\mathcal{L}}{=} \mathcal{O}_X^T\left[\phi_{i_0}(D_X)\phi_{i_0}(D_X)^T\right]\mathcal{O}_X \ ,$$

where by a slight abuse of notation we have denoted by $\phi_{i_0}(X)$ an eigenvector associated with $\max_{1\leq j\leq k} l_{1,j}$. Since we have already seen that $\mathbf{E}\left[\phi_{i_0}(D_X)\phi_{i_0}(D_X)^T\right]$ is diagonal, we have shown that $\mathbf{E}\left[\phi_{i_0}(X)\phi_{i_0}(X)^T\right]$ is diagonal in the basis that diagonalizes X. ■

6.3. **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 possibly well-known, the advantage of our derivation is that it extends easily to higher rank perturbation. More precisely, let us consider the matrix

$$M_1 = M + U \ , \tag{34}$$

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. Our aim is to compute the characteristic polynomial of M_1 and relate it to that of M. We call

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

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 case where U has rank one, that is $U = vv^T$. 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,

$$P_{M_1}(\lambda) = \left[\prod_{i=1}^n (\lambda_i - \lambda) \right] \left(1 + \sum_{i=1}^n \frac{\mathsf{v}_i^2}{\lambda_i - \lambda} \right) \ , \tag{35}$$

from which the secular equation follows. From Equation (35), it is also clear that if λ_i is an eigenvalue of M with multiplicity $m > 1$, λ_i is also an eigenvalue of M_1 with multiplicity $m - 1$.

6.4. **GUE smoothing.** In this section, we discuss possible extensions of the stochastic regularization techniques, their efficiency and regularity. We have chosen to analyze the rank one perturbation scheme - and slight variants of it - 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 order n bound on the Lipschitz constant of the gradient for the rank one Gaussian perturbations 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, for instance $\lambda_{\max(X)} = 1$ (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 $\mathsf{F}_{GUE}(X) = \mathbf{E}\left[\lambda_{\max}(X + (\epsilon/\sqrt{n})U)\right]$, 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 $\mathsf{F}_{GUE}(X)$ 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 give a worst case bound on $L\left[\nabla \mathsf{F}_{GUE}\right]$ 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 Λ_n the spectral distribution of the $n \times n$ matrix X_n, i.e the probability distribution that puts mass $1/n$ at each of the n eigenvalues of X_n. Call w_c the solution in $(\lambda_{\max}(X_n), \infty)$ of

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

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

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

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

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

where

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

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_n is not in $\mathcal{G}$, then the top two eigenvalues should have spacing greater than $n^{-2/3}$. We expect that if X_n 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_n is a completion of a $(n - k_0) \times (n - k_0)$ matrix $X_{n-k_0,0}$ which is in $\mathcal{G}$, to which we add k_0 eigenvalues $\lambda_{\max}(X_n)$, all equal and greater than $\lambda_{\max}(X_{n-k_0,0})$, with $\lambda_{\max}(X_n)$ greater than and bounded away from $w_c(X_{n-k_0,0})$. Calling $\Lambda_{n-k_0,0}$ the spectral distribution of $X_{n-k_0,0}$, we should have

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

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

The same is true for the k_0 largest eigenvalues of $X_n(\epsilon)$ 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 $1/(l_1(X_n(\epsilon)) - l_2(X_n(\epsilon)))$, 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 GUE (and not deformed

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 $\mathcal{N}(0,2)$. 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.

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.

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.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Grundlehren Text Editions. Springer-Verlag, Berlin, 2001. ISBN 3-540-42205-6. Abridged version of ıt Convex analysis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and ıt II [ibid.; MR1295240 (95m:90002)].

Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1990. ISBN 0-521-38632-2. Corrected reprint of the 1985 original.

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. *Mathematical Programming*, 133(1):365–397, 2012.

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.

Adrian S. Lewis and Hristo S. Sendov. Twice differentiable spectral functions. *SIAM J. Matrix Anal. Appl.*, 23(2): 368–386 (electronic), 2001. ISSN 0895-4798.

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.

Kantilal Varichand Mardia, John T. Kent, and John M. Bibby. *Multivariate analysis*. Academic Press [Harcourt Brace Jovanovich Publishers], London, 1979. ISBN 0-12-471250-9. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

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.

Winfried Schirotzek. *Nonsmooth analysis*. Universitext. Springer, Berlin, 2007. ISBN 978-3-540-71332-6; 3-540-71332-8.

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.

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.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. ISBN 0-521-49603-9; 0-521-78450-6.

CNRS & D.I., UMR 8548,
ÉCOLE NORMALE SUPÉRIEURE, PARIS, FRANCE.
E-mail address: aspremon@ens.fr

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