The Gradient Complexity of Linear Regression

Mark Braverman MBRAVERM@CS.PRINCETON.EDU
Princeton University

Elad Hazan EHAZAN@PRINCETON.EDU
Princeton University and Google AI Princeton

Max Simchowitz[*] MSIMCHOW@BERKELEY.EDU
UC Berkeley

Blake Woodworth[†] BLAKE@TTIC.EDU
Toyota Technological Institute at Chicago

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We investigate the computational complexity of several basic linear algebra primitives, including largest eigenvector computation and linear regression, in the computational model that allows access to the data via a matrix-vector product oracle. We show that for polynomial accuracy, $\Theta(d)$ calls to the oracle are necessary and sufficient even for a randomized algorithm.

Our lower bound is based on a reduction to estimating the least eigenvalue of a random Wishart matrix. This simple distribution enables a concise proof, leveraging a few key properties of the random Wishart ensemble.

1. Introduction

Solving linear systems and computing eigenvectors are fundamental problems in numerical linear algebra, and have widespread applications in numerous scientific, mathematical, and computational fields. Due to their simplicity, parallelizability, and limited computational overhead, first-order methods based on iterative gradient updates have become increasingly popular for solving these problems. Moreover, in many settings, the complexity of these methods is currently well understood: tight upper and lower bounds are known for gradient methods, accelerated gradient methods and related algorithms.

First order methods for regression and eigenvector computation. As an example, consider the problem of computing the largest eigenvector for a given matrix $M \in \mathbb{R}^{d\times d}$. The power method finds an ϵ-approximate solution in $\mathcal{O}\left(\frac{\log d}{\epsilon}\right)$ iterations, each involving a matrix-vector product that can be computed in time proportional to the number of non-zeros in the matrix. A variant of the Lanczos algorithm improves this complexity to $\mathcal{O}\left(\frac{\log d}{\sqrt{\epsilon}}\right)$ (Kuczyński and Woźniakowski, 1992; Musco and Musco, 2015). Alternatively, if the matrix has an inverse-eigengap $\frac{\lambda_1(M)}{\lambda_1(M)-\lambda_2(M)}$ bounded by κ, the above running times can be improved to $\mathcal{O}\left(\kappa \log \frac{d}{\epsilon}\right)$ and $\mathcal{O}\left(\sqrt{\kappa} \log \frac{d}{\epsilon}\right)$. In a low accuracy regime, where $\epsilon \gg 1/d$, these upper bounds attained by Lanczos are known to be

[*] Work done while visiting Princeton University.
[†] Work done while visiting Google AI Princeton.

information-theoretically tight in the number of matrix-vector products required to compute a solution to the given precision (Simchowitz et al., 2018). The optimal complexities are nearly identical for solving linear systems, except that these do not incur a $\log d$ dependence on the ambient dimension d (Simchowitz, 2018). More generally, these upper and lower bounds extend to convex optimization with first order methods more broadly (Nemirovskii et al., 1983).

The blessing of data sparsity. One major advantage of first order methods is that they benefit from *sparsity*. Each iteration of first order methods computes $\mathcal{O}(1)$ matrix-vector multiplies, and if the matrix in question has $\#\text{nnz}$ non-zero entries, then these multiplications can be performed in $\mathcal{O}(\#\text{nnz})$-time. This yields runtimes which scale with the sparsity of the problem instance, rather than the ambient dimension (which can be quadratically worse).

Regime of Dominance	Running time	Method
$\epsilon \geq 1/\text{poly}(d)$	$\frac{1}{\sqrt{\epsilon}} \times (\#\text{nnz})$	Lanczos, CG, AGD
$\kappa < \frac{1}{\epsilon}$	$\sqrt{\kappa} \log \frac{1}{\epsilon} \times (\#\text{nnz})$	Lanczos, CG, AGD
$\kappa, \epsilon^{-1} \geq d^2$	$d \times (\#\text{nnz})$	Lanczos & CG (not AGD)
$\#\text{nnz} = d^2$ (ties with above)	d^3	matrix inversion (naive)
$\#\text{nnz} \geq d^{\omega-1}$	$d^\omega + d^2 \log \frac{1}{\epsilon}$	matrix inversion (state of art)

Table 1: Methods for computing the largest eigenvector of $A \in R^{d \times d}$, and solving linear systems with d-data points, equivalent to computing $A^{-1}b$ for $b \in R^d$. Here $\#\text{nnz}$ denotes the number of nonzero entries of A, κ refers to an upper bound the condition number (in least squares) or eigengap (in PCA). For eigenvalue problems, above runtimes suppress $\log d$-dependence. Lanczos refers to the block Lanczos methods (Musco and Musco, 2015), CG to the conjugate gradient methods (e.g. Trefethen and Bau III (1997)), AGD to accelerated gradient descent. We use 'naive' matrix elimination refers to approaches such as those based on Gaussian elimination; 'state of art' matrix inversion denotes the theoretically state-of-art approach due to Williams (2012), which enjoys exponent $\omega \approx 2.3727$. Logarithmic factors associated with iteration complexity are included; logarithmic factors associated with numerical precision are suppressed.

The high accuracy regime. What is the computational complexity of obtaining high, inverse polynomial $\epsilon = \frac{1}{\text{poly}(d)}$ precision using a randomized algorithm, without a bound on the condition number or eigengap of the matrix? This is precise the gap in the literature that our results address.

In this regime, our understanding of even these most basic problems is poor by comparison. The best known algorithms for ϵ-approximation in this regime scale as $\mathcal{O}\left(d^\omega + d^2 \log \frac{1}{\epsilon}\right)$, where ω is the matrix inversion constant, currently around 2.37. These methods proceed by attempting to invert the matrix in question. Since the inverse of a sparse matrix is itself not necessarily sparse, these methods do not take advantage of data sparsity.

It is thus natural to ask if there is a randomized algorithm based on gradient-like queries that can exploit data sparsity, even for the simplest of linear algebra problems sketched above. We note that such faster algorithms would not necessarily require an inverse of the entire matrix, and therefore would not imply faster matrix inversion.

Our results. Our main result shows that first-order methods based on gradient-like queries cannot significantly surpass the performance of the existing conjugate gradient or Lanczos methods, even for the simplest of linear algebra problems sketched above, and crucially, *even in the high-accuracy regime.*

In a computation model where each iteration corresponds to one query of a matrix-vector product, we show that $\Omega(d)$ matrix-vector product oracle queries are necessary to obtain a $1/d^2$-accurate approximation to the largest eigenvector. This is tight, as d such queries are sufficient for the Lanczos method to obtain an exact solution (up to machine precision). Similarly, we show a lower bound of $\widetilde{\Omega}\left(d\right)$ queries for solving linear systems, which nearly matches the d-query upper bound of the conjugate gradient method.

Moreover, for instances with $\#\mathrm{nnz}(A) = \Theta(s^2)$ nonzero entries. we show a lower bound of $\Omega(s)$ queries necessary for high-precision eigenvalue approximation, and $\widetilde{\Omega}\left(s\right)$ for solving linear systems. This suggests an overall computational complexity of $\Omega(s^3)$ for first order methods. This in turn demonstrates that algebraic methods based on matrix inversion asymptotically outperform optimization-based approaches in the regime $s \geq d^{\omega/3}$.

Finally, our lower bounds are constructed so that the instance sparsity s encodes the eigengap (resp. condition number) parameters for eigenvector approximation (resp. least squares). In turn, these parameters can in turn be used to encode target accuracy ϵ in the low-accuracy regime. When translated in terms of these parameters, our guarantees are near-optimal up to logarithmic factors in terms of both eigengap/condition number and accuracy. In contrast to much existing work, our lower bounds are *information theoretic*, and apply to randomized algorithms, even those that do not satisfy Krylov restrictions. To our knowledge, this is the first work that provides lower bounds which apply to general randomized algorithms, and attain optimal dimension-dependence in the high accuracy regime when $\epsilon \ll 1/\mathrm{poly}(d)$. For a thorough discussion of the prior art, see our discussion of related work below.

Randomized algorithms Our work establishes lower bounds for *randomized* algorithms. These are more interesting than lower bounds for deterministic algorithms for several reasons. Of course, the former are stronger and more widely applicable than the latter. More importantly, there are problems for which randomized algorithms can outperform deterministic algorithms enormously, for instance, the only polynomial time algorithms for volume computation are randomized (Lovász and Vempala, 2006).

Lastly, the linear algebraic problems we consider are of great use in machine learning problems, which are frequently tackled using randomized approaches in order to avoid poor dimensional dependencies. As an example, randomized matrix sketching algorithms can substantially reduce the complexity of PCA or SVD for very large matrices. For instance, computing the top-k singular vectors of a $d \times d$ matrix requires kd^2 time for traditional (deterministic) iterative methods, but can be reduced to $d^2 + dk^2/\epsilon^4$ using a randomized sketching approach (Clarkson and Woodruff, 2017), which can be much better for moderate ϵ.

Related Work. There is an extensive literature on algorithms for linear algebraic tasks such as solving linear systems and computing eigenvalues and eigenvectors, see for example the survey of Sachdeva et al. (2014). In the interest of brevity, we focus on the relevant lower bounds literature.

The seminal work of Nemirovskii et al. (1983) establishes lower bounds which apply only to *deterministic* algorithms. These first order lower bounds enjoy essentially optimal dependence all relevant problem parameters, including dimension. However, these constructions are based on a so-called resisting oracle, and therefore *do not* extend to the randomized algorithms considered in this work.

For randomized algorithms, the lower bounds of Simchowitz et al. (2018) and Simchowitz (2018) yield optimal dependence on the eigengap and condition number parameters. However, these bounds require the dimension to be polynomial large in these parameters, which translates into a suboptimal dimension-dependent lower bound of $\widetilde{\Omega}\left(d^{1/3}\right)$.

A series of papers due to Woodworth and Srebro (2016, 2017) prove lower bounds for first order convex optimization algorithms which obtain optimal dependence on relevant parameters, but hold only in high dimensions. Furthermore, they are based on intricate, non-quadratic convex objectives which can effectively "hide" information in a way that linear algebraic instances cannot. Thus, they do not apply to the natural linear algebraic constructions that we consider. For high dimensional/low accuracy problems, there are also lower bounds for randomized algorithms that use higher order derivatives, see e.g. (Agarwal and Hazan, 2017). These, like the previously mentioned lower bounds, also only apply in high dimensions and imply dimension-dependent lower bounds like $\widetilde{\Omega}\left(d^{1/3}\right)$.

Finally, in concurrent, related work, Sun et al. (2019) study numerous other linear algebraic primitives in the same matrix-vector product oracle setting. They use a similar approach to proving lower bounds for other problems and randomized algorithms, but do not address the fundamental problems of maximum eigenvalue computation and linear regression as we do.

Proof Techniques. One of the greatest strengths of our results is the simplicity of their proofs. In general, establishing query lower bounds which apply to randomized algorithms requires great care to control the amount of information accumulated by arbitrary, randomized, adaptive queries. Currently, the two dominant approaches are either (a) to construct complex problem instances that obfuscate information from any sequence of queries made (Woodworth and Srebro, 2016), or (b) reduce the problem to estimating of some hidden component (Simchowitz et al., 2018; Simchowitz, 2018). The constructions for approach (a) are typically quite intricate, require high dimensions, and do not extend to linear algebraic problems. Approach (b) requires sophisticated information theoretic tools to control the rate at which information is accumulated.

In contrast, our work leverages simple problems of a classic random matrix ensemble known as the *Wishart* distribution Anderson et al. (2010). In particular, our lower bound for maximum eigenvalue computation is witnessed by a very natural instance $\mathbf{M} = \mathbf{W}\mathbf{W}^\top$ where the entries of $\mathbf{W}$ are i.i.d. Gaussian. This is plausibly a very benign instance as it is one of the simplest distributions over symmetric positive definite matrices that one might think of.

The simplicity of the problem instance, and existing understanding of the distribution of the spectrum of Wishart matrices allows for concise, straightforward proofs.

1.1. Notation

Let $\mathcal{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$, $\mathbb{S}^d := \{M \in \mathbb{R}^{d\times d} : M = M^\top\}$ and $\mathbb{S}^d_{++} := \{A \in \mathbb{S}^d : A \succ 0\}$. As a general rule, we use M for matrices which arise in eigenvector problems, and A for matrices which arise in least-squares problems. For $M \in \mathbb{S}^d$, we let $\mathrm{gap}(M) := \frac{\lambda_1(M)-\lambda_2(M)}{\lambda_1(M)}$, and for $A \in \mathbb{S}^d_{++}$, we set $\mathrm{cond}(A) := \frac{\lambda_1(A)}{\lambda_d(A)}$. We adopt the conventional notions $\mathcal{O}\left(\cdot\right), \Omega\left(\cdot\right), \Theta\left(\cdot\right)$ as suppressing universal constants independent of dimension and problem parameters, let $\widetilde{\mathcal{O}}\left(\cdot\right), \widetilde{\Omega}\left(\cdot\right)$ suppress logarithmic factors, and let $g(x) = \mathcal{O}^*\left(f(x)\right)$ denote a term which satisfies $g(x) \leq cf(x)$

for a particular, unspecified, but sufficiently small constant c. We say a matrix is s sparse if its number of nonzero entries is at most s.

2. Main Results

We begin the section by stating a lower bound for the problem of eigenvalue estimation and eigenvector approximation via matrix-vector multiply queries. Via a standard reduction, this bound will imply a lower bound for solving linear systems via gradient-queries.

We stress that, unlike prior lower bounds, our bounds for eigenvalue problems (resp. linear systems) both apply to arbitrary, randomized algorithms, *and* capture the correct dependence on the eigengap (resp. condition number), all the way up to a $\Omega(d)$ (resp. $\widetilde{\Omega}(d)$) worst-case lower bound in d dimensions. This worst-case lower bound is matched by the Lanczos (Musco and Musco, 2015) and Conjugate Gradient methods (see, e.g. Trefethen and Bau III (1997)), which, assuming infinite precision, efficiently recover the exact optimal solutions in at most d queries.

2.1. Eigenvalue Problems

Before introducing our results, we formalize the query model against which our lower bounds hold:

Definition 1 (Eigenvalue and Eigenvector Algorithms) *An eigenvalue approximation algorithm, or* $\mathrm{EigValueAlg}$*, is an algorithm* Alg *which interacts with an unknown matrix* $M \in \mathbb{S}^d_{++}$ *via* T *adaptive, randomized queries,* $\mathsf{w}^{(i)} = M\mathsf{v}^{(i)}$*, and returns an estimate* $\widehat{\lambda}$ *of* $\lambda_1(M)$*. An eigenvector approximation algorithm, or* $\mathrm{EigVecAlg}$*, operates in the same query model, but instead returns an estimate* $\widehat{v} \in \mathcal{S}^{d-1}$ *of* $v_1(M)$*. We call* $T := \mathrm{Query}(\mathsf{Alg})$ *the* query complexity *of* Alg.

We let $\mathbb{P}_{\mathbf{M}\sim\mathcal{D},\mathsf{Alg}}$ denote the probability induced by running Alg when the input is a random instance $\mathbf{M}$ drawn from a distribution $\mathcal{D}$. We now state our main query lower bound for $\mathrm{EigValueAlg}$'s, which we prove in Section 3. Our lower bound considers a distribution over symmetric matrices M which are also PSD, to show that our lower bounds hold even under the most benign, and restrictive conditions:[1]

Theorem 2 (Lower Bound for Eigenvalue Estimation) *There is a function* $d_0 : (0,1) \to \mathbb{N}$ *such that the following holds. For any* $\beta \in (0,1)$*, ambient dimension* $d \geq d_0(\beta)$*, and every sparsity level* $s \in [d_0(\beta), d]$*, there exists a distribution* $\mathcal{D} = \mathcal{D}(s,d,\beta)$ *supported on* s^2*-sparse matrices in* $\mathbb{S}^d_{++}$ *such that any* $\mathrm{EigValueAlg}$ Alg *with* $\mathrm{Query}(\mathsf{Alg}) \leq (1-\beta)s$ *satisfies*

$$\mathbb{P}_{\mathbf{M}\sim\mathcal{D},\mathsf{Alg}}\left[|\widehat{\lambda} - \lambda_1(\mathbf{M})| \geq \frac{1}{20s^2}\right] \geq \Omega(\sqrt{\beta})$$

Moreover, $\mathbf{M} \sim \mathcal{D}$ *satisfies* $\mathrm{gap}(\mathbf{M}) \geq \Omega_\beta(1)/s^2$*, and* $1 - \Omega_\beta(1)/s^2 \leq \lambda_1(\mathbf{M}) \leq 1$ *almost surely. Here,* $\Omega_\beta(1)$ *denotes a quantity lower bounded by a function of* β*, but not on* s *or* d.

In particular, any algorithm requires $\mathrm{Query}(\mathsf{Alg}) \geq \Omega(d)$ queries in ambient dimension d to estimate $\lambda_1(\mathbf{M})$ up to error $\mathcal{O}^*\left(d^{-2}\right)$ with constant probability, and in fact requires $(1 - o(1))d$ queries for a $1 - o(1)$ probability of error.

[1] Note that a lower bound on PSD matrices holds a fortiori for arbitrary symmetric and square matrices, since PSD are a subclass.

Remark 3 *The sparsity parameter in our construction can be used to encode the accuracy parameter ϵ. Specifically, by setting the parameter $s = 1/\sqrt{\epsilon}$, Theorem 2 implies $1/\sqrt{\epsilon}$ queries are necessary for ϵ-accuracy. Alternatively, choosing $s \geq \Omega(\sqrt{\mathrm{gap}(\mathbf{M})})$, we obtain a gap-dependent bound requiring $\Omega(1/\sqrt{\mathrm{gap}(\mathbf{M})})$ queries for $\mathcal{O}\left(\mathrm{gap}(\mathbf{M})\right)$ accuracy. Both bounds match the sharp lower bounds of Simchowitz et al. (2018) up to logarithmic factors, while also capturing the correct worst-case query complexity for ambient dimension d, namely $\Omega(d)$. Moreover, our proof is considerably simpler.*

Implications for sparsity. For $s \geq \sqrt{d}$, our lower bound says that first order methods require $\Omega(s)$ queries to approximate the top eigenvalue of matrices $\mathbf{M}$ with $\#nnz(\mathbf{M}) = \Theta(s^2)$. Therefore, implementations of first-order methods based on standard matrix-vector multiplies cannot have complexity better than $\Omega(s^3)$ in the worst case. On the other hand, matrix inversion has runtime d^{ω}, and is sufficient both for solving least squares, and for computing eigenvalues and eigenvectors (Garber and Hazan, 2015). Hence, for $s \in [d^{\omega/3}, d]$, we see that matrix inversion *outperforms* first-order based methods.

Approximating the top eigenvector. As a corollary, we obtain the analogous lower bound for algorithms approximating the top eigenvector of a symmetric matrix, and, in particular, an $\Omega(s)$ query complexity lower bound for $\epsilon = \mathcal{O}^*\left(s^{-2}\right)$-precision approximations:

Corollary 4 *In the setting of Theorem 2, any* $\mathrm{EigVecAlg}$ *with* $\mathrm{Query}(\mathsf{Alg}) \leq (1-\beta)s - 1$ *satisfies*

$$\mathbb{P}_{\mathbf{M}\sim\mathcal{D},\mathsf{Alg}}\left[\widehat{v}^{\top}\mathbf{M}\widehat{v} \leq \lambda_1(\mathbf{M})\left(1 - \frac{1}{20s^2}\right)\right] \geq \Omega(\sqrt{\beta})$$

Proof For ease, set $\epsilon := \frac{1}{20s^2}$. Let $\mathrm{Query}(\mathsf{Alg}) \leq (1-\beta)s - 1$, and let $\widehat{\lambda} = \widehat{v}^{\top}\mathbf{M}\widehat{v}$, which can be computed using at most 1 additional query. Since $\widehat{\lambda} \leq \max_{v \in \mathcal{S}^{d-1}} v^{\top}\mathbf{M}v = \lambda_1(\mathbf{M})$, and since $\lambda_1(\mathbf{M}) \leq 1$ we see that $\widehat{v}^{\top}\mathbf{M}\widehat{v} \geq \lambda_1(\mathbf{M}) - \epsilon$ only if $|\widehat{\lambda} - \lambda_1(\mathbf{M})| \leq \epsilon$. Thus, recalling $\lambda_1(\mathbf{M}) \leq 1$, we have $\mathbb{P}_{\mathcal{D},\mathsf{Alg}}\left[\widehat{v}^{\top}\mathbf{M}\widehat{v} \geq \lambda_1(\mathbf{M})(1-\epsilon)\right] \leq \mathbb{P}_{\mathcal{D},\mathsf{Alg}}\left[\widehat{v}^{\top}\mathbf{M}\widehat{v} \geq \lambda_1(\mathbf{M}) - \epsilon\right] = \mathbb{P}_{\mathcal{D},\mathsf{Alg}}\left[|\widehat{\lambda} - \lambda_1(\mathbf{M})| \leq \epsilon\right]$, which is at least $\Omega(\sqrt{\beta})$ by Theorem 2. ■

Again, the the sparsity s can be used to encode the accuracy parameter ϵ via $\epsilon := \frac{1}{20s^2}$.

2.2. Lower Bounds for Solving Linear Systems

We now present our lower bounds for minimizing quadratic functions. We consider the following query model:

Definition 5 (Gradient Query model for Linear System Solvers) *We say that* Alg *is an* $\mathrm{LinSysAlg}$ *if* Alg *is given initial point $x_0 \in \mathbb{R}^d$ and linear term $b \in \mathbb{R}^d$, and it interacts with an unknown symmetric matrix $A \in \mathbb{S}^d_{++}$ via T adaptive, randomized queries, $\mathsf{w}^{(i)} = A\mathsf{v}^{(i)}$, and returns an estimate $\widehat{x} \in \mathbb{R}^d$ of $A^{-1}b$. Again, we call T the* query complexity *of* Alg.

Defining the objective function $f_{A,b}(x) := \frac{1}{2}x^{\top}Ax - b^{\top}x$, we see that the query model of Definition 5 is equivalent to being given a gradient query at 0, $\nabla f_{A,b}(0) = b$, and making queries $\nabla f_{A,b}(\mathsf{v}^{(i)}) = A\mathsf{v}^{(i)} - b$. We shall use $\mathbb{P}_{\mathsf{Alg},(x_0,b,A)}$ do denote probability induced by running the a $\mathrm{LinSysAlg}$ Alg on the instance (x_0, b, A). Our lower bound in this model is as follows, stated in terms of the function suboptimality $\|\widehat{x} - A^{-1}b\|_A^2 = f_{A,b}(\widehat{x}) - \min_x f_{A,b}(x)$.

Theorem 6 (Lower Bound for Linear System Solvers) *Let $d_0 \in \mathbb{N}$ be a universal constant. Then for all ambient dimensions $d \geq d_0$, and all $s \in [d_0^2, d^2]$, any* LinSysAlg Alg *which satisfies the guarantee*

$$\mathbb{P}_{\mathsf{Alg},x_0,b,A}\left[\|\widehat{x} - A^{-1}b\|_A^2 \leq \frac{1}{e} \cdot \frac{\lambda_1(A)\|x_0\|^2}{s^2}\right] \geq 1 - \frac{1}{e} \tag{1}$$

for all $(d + s^2)$-sparse matrices $M \in \mathbb{S}_{++}^d$ with $\mathrm{cond}(M) \leq \mathcal{O}\left(s^2\right)$, and all $(x_0, b) \in \mathbb{R}^d \times \mathbb{R}^d$, must have query complexity at least

$$\mathrm{Query}(\mathsf{Alg}) \geq \Omega\left(s \cdot (\log^{2+\log} s)^{-1}\right),$$

where $\log^{p+\log}(x) := (\log^p x) \cdot (\log\log x)$.

In particular, any algorithm which ensures $\|\widehat{x} - A^{-1}b\|_A^2 \leq \mathcal{O}^*\left(\frac{\lambda_1(A)\|x_0\|^2}{d^2}\right)$ with probability $1 - \frac{1}{e}$ requires $\widetilde{\Omega}\left(d\right)$-queries.

Remark 7 *As with the eigenvector lower bounds, we can use the sparsity parameters to encode accuracy. Specifically, by $s = \sqrt{\mathrm{cond}}$, obtaining a function suboptimality $f_{A,b}(\widehat{x}) - \min_x f_{A,b}(x) = \|\widehat{x} - A^{-1}b\|_A^2$ of $\mathcal{O}^*\left(1/\mathrm{cond}\right)$ requires $\widetilde{\Omega}\left(\sqrt{\mathrm{cond}}\right)$ queries, matching known upper bounds achieved by the conjugate gradient method up to logarithmic factors (Trefethen and Bau III, 1997), which in turn match information-theoretic lower bounds (Simchowitz, 2018). Moreover, these in turn can be converted into an* minimax *lower bound by in turn selecting the condition number parameter as* $\mathrm{cond} \propto 1/\epsilon$. *Indeed, this implies that to ensure $f_{A,b}(\widehat{x}) - \min_x f_{A,b}(x) \leq \epsilon$, one requires $\widetilde{\Omega}\left(1/\sqrt{\epsilon}\right)$ queries, entailing the minimax rate up to logarithmic factors.*

We prove Theorem 6 by leveraging a well-known reduction from eigenvector approximation to minimizing quadratic functions, known as "shift-and-invert" (Saad, 2011; Garber et al., 2016) . To state the result, we define a class of matrices to which the reduction applies:

$$\mathscr{M}_d(\mathtt{gap}, \alpha) := \left\{M \in \mathbb{S}_{++}^d : \mathrm{gap}(M) \geq \mathtt{gap},\ |\lambda_1(M) - 1| \leq \alpha\,\mathtt{gap},\ \lambda_1(M) \in \left[\tfrac{1}{2}, 2\right]\right\},$$

The term $\mathtt{gap}$ corresponds to $\mathrm{gap}(M)$, whereas α measures to how close $\lambda_1(M)$ is to 1. The rescaling to ensure $\lambda_1(M) \in \left[\frac{1}{2}, 2\right]$ is for simplicity, and more generally α corresponds to an approximate foreknowledge of $\lambda_1(M)$ (which is necessary to facilitate the reduction). We further note that the distribution $\mathcal{D}(s, d, \beta)$ from Theorem 2 satisfies, for some functions $c_{\mathrm{gap}}(\cdot)$ and $c_{\mathrm{eig}}(\cdot)$,

$$\mathbb{P}_{\mathbf{M}\sim\mathcal{D}(s,d,\beta)}\left[\mathbf{M} \in \mathscr{M}_d\left(\frac{c_{\mathrm{gap}}(\beta)}{s^2}, \frac{c_{\mathrm{eig}}(\beta)}{c_{\mathrm{gap}}(\beta)}\right)\right] = 1 \tag{2}$$

With this definition in hand, we provide a precise guarantee for the reduction, which we prove in Appendix A:

Proposition 8 (Eigenvector-to-Linear-System Reduction) *Let $d \geq d_{\min}$ for a universal $d_{\min}$. Fix a $\mathtt{gap} \in (0, 1)$, $\alpha > 0$, and suppose that* Alg *be a* LinSysAlg *which satisfies (1) with* $\mathtt{cond} :=$

$1 + \alpha + \frac{1}{\mathsf{gap}}$ *for all* $A \in \mathbb{S}^d_{++}$ *with* $\mathrm{cond}(A) \le \mathtt{cond}$. *Then, for any* $\delta \in (0, 1/e)$*, there exists an* EigVecAlg, $\mathsf{Alg}_{\mathrm{eig}}$, *which satisfies*

$$\mathbb{P}_{\mathsf{Alg}_{\mathrm{eig}},M}\left[\widehat{v}^\top M \widehat{v} \ge (1 - c\mathsf{gap})\lambda_1(M)\right] \ge 1 - \delta, \quad \forall M \in \mathscr{M}_d\left(\mathsf{gap}, \alpha\right)$$

with query complexity at most

$$\mathrm{Query}(\mathsf{Alg}_{\mathrm{eig}}) \le \mathrm{Query}(\mathsf{Alg}) \cdot \mathcal{O}_\alpha\left((\log \frac{1}{\delta}) \cdot \log^{2+\log} \frac{d}{\min\{c\mathsf{gap},\, 1\}}\right),$$

where $\mathcal{O}_\alpha\left(\cdot\right)$ *hides multiplicative and additive constants depending on* α.

We can now prove Theorem 6 by combining Proposition 8 and Theorem 2:

Proof We shall prove the theorem in the regime where $s = d$. The general s case is attained by embedding an instance of dimension s into dimension d, as in the proof of Theorem 2, and is deferred to Appendix B.

To begin, let $\beta = \frac{1}{2}$ (any constant in $(0, 1)$ suffices); throughout, β will be a universal constant, rather than a problem parameter. Next, fix an ambient dimension $d \ge d_0 := d_{\min} \vee d_0(\beta)$, where d_0 is from Theorem 2, and $d_{\min}$ from Proposition 8.

Lastly, let $\mathsf{gap} := \frac{c_{\mathrm{gap}}(\beta)}{d^2}$ and $\alpha := \frac{c_{\mathrm{eig}}(\beta)}{c_{\mathrm{gap}}(\beta)}$, where $c_{\mathrm{gap}}(\cdot)$ and $c_{\mathrm{eig}}(\cdot)$ are as in (2). Let $\mathbf{M} \sim \mathcal{D}(d, d, \beta)$. Then, (2) ensures $\mathbf{M} \in \mathscr{M}_d(\mathsf{gap}, \alpha)$ with probability 1. For the sake of contradiction, suppose that Alg is a LinSysAlg which satisfies the guarantee of (1) for all

$$A : \mathrm{cond}(A) \le 1 + \alpha + \frac{1}{\mathsf{gap}} = \mathcal{O}\left(d^2\right).$$

Then, for the universal constant $c := \frac{1}{20 c_{\mathrm{gap}}(\beta)}$, there exists an EigVecAlg $\mathsf{Alg}_{\mathrm{eig}}$ which satisfies, for all $M \in \mathscr{M}_d\left(\mathsf{gap}, \alpha\right)$,

$$\mathbb{P}_{\mathsf{Alg}_{\mathrm{eig}},M}\left[\widehat{v}^\top M \widehat{v} \ge \left(1 - \frac{1}{20d^2}\right)\lambda_1(M)\right] = \mathbb{P}_{\mathsf{Alg}_{\mathrm{eig}},M}\left[\widehat{v}^\top M \widehat{v} \ge (1 - c\,\mathsf{gap})\lambda_1(M)\right] \ge 1 - \Omega(\sqrt{\beta}),$$

whose query complexity $\mathrm{Query}(\mathsf{Alg}_{\mathrm{eig}})$ is bounded by

$$\mathrm{Query}(\mathsf{Alg}) \cdot \mathcal{O}_\alpha\left((\log \tfrac{1}{\Omega(\sqrt{\beta}) \wedge e}) \cdot \log^{2+\log} \tfrac{d}{1 \wedge c\,\mathsf{gap}, 1}\right) = \mathrm{Query}(\mathsf{Alg}) \cdot \mathcal{O}\left(\log^{2+\log} d\right),$$

where we use $\mathsf{gap} = \Theta(d^2)$, and that $\alpha, c, \beta, \Omega(\beta)$ depend on universal constants, and not on the choice of dimension d. By Theorem 2, we must have $\mathrm{Query}(\mathsf{Alg}_{\mathrm{eig}}) \ge d/2$, whence

$$\mathrm{Query}(\mathsf{Alg}) \ge \frac{c_1}{\sqrt{\mathsf{gap}}} \cdot \Omega\left((\log^{2+\log} d)^{-1}\right) = \Omega\left(d(\log^{2+\log} d)^{-1}\right).$$

■

3. Proof of Theorem 2

The proof of Theorem 2 follows by deriving a lower bound for the problem of estimating the least eigenvalue of a classical random matrix ensemble known as the (standard) *Wishart* matrices:

Definition 9 (Wishart Distribution) *We write* $\mathbf{W} \sim \mathrm{Wishart}(d)$ *to denote a random matrix with the distribution* $\mathbf{W} \overset{d}{=} \mathbf{X}\mathbf{X}^\top$, *where* $\mathbf{X} \in \mathbb{R}^{d \times d}$ *and* $\mathbf{X}_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{d}I)$.

We now state our main technical contribution, which lower bounds the number of queries required for estimation the smallest eigenvalue of a matrix $\mathbf{W} \sim \mathrm{Wishart}(d)$:

Theorem 10 (Lower Bound for Wishart Eigenvalue Estimation) *There exists a universal constant* p_0 *and function* $\mathsf{d} : (0,1) \to \mathbb{N}$ *such that the following holds: for all* $\beta \in (0,1)$, *and all* $d \geq \mathsf{d}(\beta)$, *we have that* $\mathbf{W} \sim \mathrm{Wishart}(d)$ *satisfies*

(a) Any algorithm Alg *which makes* $T \leq (1-\beta)d$ *adaptively chosen queries, and returns an estimate* $\widehat{\lambda}_{\min}$ *of* $\lambda_{\min}(\mathbf{W})$ *satisfies*

$$\mathbb{P}_{\mathbf{W},\mathsf{Alg}}\left[|\widehat{\lambda}_{\min} - \lambda_{\min}(\mathbf{W})| \geq \frac{1}{4d^2}\right] \geq \mathrm{c}_{\mathrm{wish}}\sqrt{\beta}.$$

(b) There exists constants $C_1(\beta)$ *and* $C_2(\beta)$ *such that*

$$\mathbb{P}_{\mathbf{W}}\left[\{\lambda_d(\mathbf{W}) \leq C_1(\beta)d^{-2}\} \cap \{\lambda_{d-1}(\mathbf{W}) - \lambda_d(\mathbf{W}) \geq C_2(\beta)d^{-2}\} \cap \{\|\mathbf{W}\| < 5\}\right] \geq 1 - \tfrac{\mathrm{c}_{\mathrm{wish}}\sqrt{\beta}}{2}.$$

Note that, by taking $\beta = o(1)$, Theorem 10 in fact demonstrates that $(1 - o(1))d$ queries are required for an $o(1)$ probability of failure, showing that no nontrivial improvements can be achieved.
Proof [Proof of Theorem 2] Fix $\beta \in (0,1)$, and let $d_0(\cdot) = \mathsf{d}(\cdot)$ denote the function from Theorem 10. For $s \geq d_0(\beta)$, let $\mathbf{W} \sim \mathrm{Wishart}(s)$, and define

$$\mathbf{M} = \begin{bmatrix} I_{s \times s} - \frac{1}{5}\mathbf{W} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{d \times d}.$$

By construction, $\mathbf{M}$ has sparsity s^2. Let us denote the event of part (b) of Theorem 10 $\mathcal{E}$. Then $\mathcal{E}$ occurs with with probability $1 - \frac{\mathrm{c}_{\mathrm{wish}}\sqrt{\beta}}{2}$.. On $\mathcal{E}$, we further have

- $0 \preceq \mathbf{M} \preceq 1$
- $\mathrm{gap}(\mathbf{M}) = \frac{\lambda_1(\mathbf{M}) - \lambda_2(\mathbf{M})}{\lambda_1(\mathbf{M})} \geq \frac{d^{-2}}{5}C_2(\beta) = \Omega_\beta(1) \cdot d^{-2}$
- $|\lambda_1(\mathbf{M}) - 1| = \frac{1}{5}\lambda_{\min}(\mathbf{W}) \leq \Omega_\beta(1)d^{-2}$

Now consider an estimator $\widehat{\lambda}$ of $\lambda_{\max}(\mathbf{M})$. By considering the induced estimator $\widehat{\lambda}_{\min} := 5(1 - \widehat{\lambda})$ of $\lambda_{\min}(\mathbf{W})$, part (a) of Theorem 10 and a union bound implies that

$$\begin{aligned}\mathbb{P}_{\mathbf{M},\mathsf{Alg}}\left[\{|\widehat{\lambda}-\lambda_{\max}(\mathbf{M})|\le\frac{1}{20d^2}\}\mid\mathcal{E}\right] &= \mathbb{P}_{\mathbf{W},\mathsf{Alg}}\left[\{|\widehat{\lambda}_{\min}-\lambda_{\min}(\mathbf{W}))|\le\frac{1}{4d^2}\}\mid\mathcal{E}\right]\\ &= \mathbb{P}_{\mathbf{W},\mathsf{Alg}}\left[\{|\widehat{\lambda}_{\min}-\lambda_{\min}(\mathbf{W}))|\le\frac{1}{4d^2}\}\cap\mathcal{E}\right]\\ &\ge \mathbb{P}_{\mathbf{W},\mathsf{Alg}}\left[\{|\widehat{\lambda}_{\min}-\lambda_{\min}(\mathbf{W}))|\le\frac{1}{4d^2}\}\right]-\mathbb{P}_{\mathbf{W},\mathsf{Alg}}\left[\mathcal{E}\right]\\ &\overset{(i)}{\ge} \mathrm{c}_{\mathrm{wish}}\sqrt{\beta}-\frac{1}{2}\mathrm{c}_{\mathrm{wish}}\sqrt{\beta}=\frac{1}{2}\mathrm{c}_{\mathrm{wish}}\sqrt{\beta}\end{aligned}$$

where (i) uses Theorem 2. Hence, let $\mathcal{D}$ denote the distribution of $\mathbf{M}$ conditioned on $\mathcal{E}$, any $\mathrm{EigValueAlg}$ Alg with $\mathrm{Query}(\mathsf{Alg})\le(1-\beta)d$ queries satisfies

$$\mathbb{P}_{\mathbf{M}\sim\mathcal{D},\mathsf{Alg}}\left[\{|\widehat{\lambda}-\lambda_{\max}(\mathbf{M})|\le\frac{1}{20d^2}\}\right]\ge\frac{1}{2}\mathrm{c}_{\mathrm{wish}}\sqrt{\beta}=\Omega(\sqrt{\beta}).$$

■

3.1. Proof of Theorem 10

We begin the proof of Theorem 10 by collecting some useful facts from the literature regarding the asymptotic distribution of Wishart spectra.

Lemma 11 (Facts about Wishart Matrices) *Let $(\mathbf{z}_d^{(d)},\mathbf{z}_{d-1}^{(d)})\in R^2$ denote random variables with the (joint) law of $(d^2\lambda_d(\mathbf{W}^{(d)}),d^2\lambda_{d-1}(\mathbf{W}^{(d)}))$, where $\mathbf{W}^{(d)}\sim\mathrm{Wishart}(d)$. The following are true:*

1. *$(\mathbf{z}_d^{(d)},\mathbf{z}_{d-1}^{(d)})$ converge in distribution to $\mathcal{D}$ distribution with $\mathbb{P}_{(\mathbf{z}_d,\mathbf{z}_{d-1})\sim\mathcal{D}}[0<\mathbf{z}_d<\mathbf{z}_{d-1}]=1$* (Ramírez and Rider (2009, Theorem 1)).
2. *$\mathbf{z}_d$ has the density $f(x)=\mathbb{I}(x\ge 0)\cdot\frac{x^{-1/2}+1}{2}e^{-(x/2+\sqrt{x})}$* (e.g. Shen (2001, Page 3))

Moreover, for any $\epsilon>0$, $\lim_{d\to\infty}\mathbb{P}_{\mathbf{W}\sim\mathrm{Wishart}(d)}[\|\mathbf{W}\|_{\mathrm{op}}\ge 4+\epsilon]=0$ (e.g. Anderson et al. (2010, Exercise 2.1.18))

We note that we use $(\mathbf{z}_d^{(d)},\mathbf{z}_{d-1}^{(d)})$ for the normalized (by d^2) eigenvalues of $\mathbf{W}^{(d)}$. We convert these asymptotic guarantees into quantitative ones (proof in Section C.1).

Corollary 12 (Non-Asymptotic Properties) *There exists a maps $d_{\mathrm{reg}},d_{\mathrm{dens}}:(0,1)\to\mathbb{N}$, functions $C_1,C_2:(0,1)\to\mathbb{R}_{>0}$, and a universal constant p_0 such that the following holds: for any $\delta\in(0,1)$ and $d\ge d_{\mathrm{reg}}(\delta)$,*

$$\mathbb{P}_{\mathbf{W}\sim\mathrm{Wishart}(d)}\left[\left\{\mathbf{z}_{d-1}^{(d)}-\mathbf{z}_d^{(d)}\ge C_2(\delta)\right\}\cap\left\{\mathbf{z}_d^{(d)}\le C_1(\delta)\right\}\cap\{\|\mathbf{W}\|_{\mathrm{op}}\le 5\}\right]\ge 1-\delta \qquad (3)$$

Moreover, for any $\alpha\in(0,1)$ and $d\ge d_{\mathrm{dens}}(\alpha)$, $\mathbb{P}[\lambda_d(\mathbf{W}^{(d)})\ge d^{-2}]\ge p_0$, while and $\mathbb{P}[\lambda_d(\mathbf{W}^{(d)})\le\alpha^2 d^{-2}]\ge p_0\alpha$.

We now show establish, in an appropriate basis, $\mathbf{W} \sim \mathrm{Wishart}(d)$ admits a useful block decomposition:

Lemma 13 *Let* $\mathbf{W} \sim \mathrm{Wishart}(d)$*. Then, for any sequence of queries* $v^{(1)}, \dots, v^{(T)}$ *and responses* $w^{(1)}, \dots, w^{(T)}$*, there exists a rotation matrix* $\mathbf{V}$ *constructed solely as a function of* $v^{(1)}, \dots, v^{(T)}$ *such that the matrix* $\mathbf{V}\mathbf{W}\mathbf{V}^\top$ *can be written*

$$\mathbf{V}\mathbf{W}\mathbf{V}^\top = \begin{bmatrix} Y_1Y_1^\top & Y_1Y_2^\top \\ Y_2Y_1^\top & Y_2Y_2^\top + \widetilde{\mathbf{W}} \end{bmatrix}$$

where $\widetilde{\mathbf{W}}$ *conditioned on the event* $\mathsf{v}^{(1)} = v^{(1)}, \dots, \mathsf{v}^{(T)} = v^{(T)}$, $\mathsf{w}^{(1)} = w^{(1)}, \dots, \mathsf{w}^{(T)} = w^{(T)}$ *satisfies* $(\frac{d}{d-T}) \cdot \widetilde{\mathbf{W}} \sim \mathrm{Wishart}(d-T)$ *distribution.*

The above lemma is proven in in Appendix C.2. The upshot of the lemma is that after T queries, there is still a portion $\widetilde{\mathbf{W}}$ of $\mathbf{W}$ that remains unknown to the query algorithm. We now show that this unknown portion exerts significant influence on the smallest eigenvalue of $\mathbf{W}$. Specifically, the following technical lemma implies that $\lambda_{\min}(\mathbf{W}) = \lambda_{\min}(\mathbf{V}\mathbf{W}\mathbf{V}^\top) \le \lambda_{\min}(\widetilde{\mathbf{W}})$:

Lemma 14 *For* $A \in \mathbb{R}^{T \times T}$, $B \in \mathbb{R}^{(d-T)\times T}$*, and symmetric* $\widetilde{W} \in \mathbb{R}^{(d-T)\times(d-T)}$*, let*

$$M = \begin{bmatrix} AA^\top & AB^\top \\ BA^\top & BB^\top + W \end{bmatrix}$$

then $\lambda_{\min}(M) \le \lambda_{\min}(W)$.

Proof Let $v \in \mathbb{R}^{d-T}$ such that $\|v\| = 1$ and $v^\top W v = \lambda_{\min}(W)$. Define $z = \begin{bmatrix} -A^{-\top}B^\top v \\ v \end{bmatrix}$. Then

$$z^\top M z = z^\top \begin{bmatrix} -AB^\top v + AB^\top v \\ -BB^\top v + BB^\top v + Wv \end{bmatrix} = v^\top W v = \lambda_{\min}(W)$$

Therefore, $\lambda_{\min}(M) \le \lambda_{\min}(W)$. ■

With all the above ingredients in place, we are now ready to complete the proof of Theorem 10:

Proof [Proof of Theorem 10] Let $T \le (1-\beta)d$, and let $t = \frac{1}{2d^2}$, $\epsilon = \frac{1}{4d^2}$. Moreover, let $\mathsf{Z} := \{\mathsf{v}^{(1)}, \dots, \mathsf{v}^{(T)}, \mathsf{w}^{(1)}, \dots, \mathsf{w}^{(T)}\}$ encode the query-response information, and let $\widetilde{\mathbf{W}}$ denote the matrix from Lemma 13. Finally, define the error probability

$$p_{\mathrm{err}} := \mathbb{P}[|\widehat{\lambda}_{\min} - \lambda_{\min}(\mathbf{W})| \ge \epsilon].$$

We can now lower bound the probability of error by lower bounding the probability that the algorithm ouputs an estimate $\widehat{\lambda}_{\min}$ above a threshold t, while the corner matrix $\widetilde{\mathbf{W}}$ has smallest eignvalue below $t - \frac{1}{\epsilon}$. We can then decouple the probability of these events using independence of $\widetilde{\mathbf{W}}$ conditioned on the queries

$$\begin{aligned} p_{\mathrm{err}} \ge \mathbb{P}[\{\widehat{\lambda}_{\min} \ge t\} \cap \{t - \epsilon \ge \lambda_{\min}(\mathbf{W})\}] &\overset{(i)}{\ge} \mathbb{P}[\{\widehat{\lambda}_{\min} \ge t\} \cap \{t - \epsilon \ge \lambda_{\min}(\widetilde{\mathbf{W}})\}] \\ &= \mathbb{E}\left[\mathbb{P}[\{\widehat{\lambda}_{\min} \ge t\} \cap \{t - \epsilon \ge \lambda_{\min}(\widetilde{\mathbf{W}})\} \mid \mathsf{Z}]\right] \\ &\ge \mathbb{E}\left[\mathbb{P}[\widehat{\lambda}_{\min} \ge t \mid \mathsf{Z}] \cdot \mathbb{P}[\widetilde{\mathbf{W}} \le t - \epsilon]\right] \\ &= \mathbb{P}[\lambda_{\min}(\widetilde{\mathbf{W}}) \le t - \epsilon] \cdot \mathbb{P}[\widehat{\lambda}_{\min} \ge t], \end{aligned} \tag{4}$$

where (i) uses Lemma 14, and (ii) use the fact that $\widetilde{\mathbf{W}}$ has a Wishart distribution conditioned on Z, and thus $\lambda_{\min}(\widetilde{\mathbf{W}})$ is independent of Z. On the other hand, we have

$$p_{\text{err}} \ge \mathbb{P}[\{\widehat{\lambda}_{\min} < t\} \cap \{\lambda_{\min}(\mathbf{W}) \ge t + \epsilon\}] \ge \mathbb{P}[\lambda_{\min}(\mathbf{W}) \ge t + \epsilon] - \mathbb{P}[\widehat{\lambda}_{\min} \ge t],$$

so that $\mathbb{P}[\widehat{\lambda}_{\min} \ge t] \ge \mathbb{P}[\lambda_{\min}(\mathbf{W}) \ge t + \epsilon] - p_{\text{err}}$. Together with (4), this implies

$$\begin{aligned} p_{\text{err}} &\ge \mathbb{P}[\lambda_{\min}(\widetilde{\mathbf{W}}) \le t - \epsilon] \cdot \mathbb{P}[\widehat{\lambda}_{\min} > t] \\ &\ge \mathbb{P}[\lambda_{\min}(\widetilde{\mathbf{W}}) \le t - \epsilon] \cdot \left(\mathbb{P}[\lambda_{\min}(\mathbf{W}) \ge t + \epsilon] - p_{\text{err}}\right). \end{aligned}$$

Performing some algebra, and recalling $\epsilon = \frac{1}{4d^2}$, $t = 2\epsilon$,

$$p_{\text{err}} \ge \frac{\mathbb{P}[\lambda_{\min}(\widetilde{\mathbf{W}}) \le t - \epsilon] \cdot \mathbb{P}[\lambda_{\min}(\mathbf{W}) \ge t + \epsilon]}{1 + \mathbb{P}[\lambda_{\min}(\widetilde{\mathbf{W}}) \le t - \epsilon]} \ge \frac{\mathbb{P}[\lambda_{\min}(\widetilde{\mathbf{W}}) \le \frac{1}{2}] \cdot \mathbb{P}[\lambda_{\min}(\mathbf{W}) \ge 1]}{2}.$$

Finally, since $T \le (1 - \beta)d$, we have $(d - T)/d \ge \beta$. Thus,

$$\begin{aligned} \mathbb{P}\left[\lambda_{\min}(\widetilde{\mathbf{W}}) \le \frac{1}{2}\right] &= \mathbb{P}_{\widetilde{\mathbf{W}} \sim \text{Wishart}(T-d)}\left[\lambda_{\min}(\widetilde{\mathbf{W}}) \le \frac{d-T}{2d}\right] \\ &\ge \mathbb{P}_{\widetilde{\mathbf{W}} \sim \text{Wishart}(T-d)}\left[\lambda_{\min}(\widetilde{\mathbf{W}}) \le \frac{\beta}{2}\right]. \end{aligned}$$

Let $\mathsf{d}(\beta) = \beta^{-1} d_{\text{dens}}(\frac{\beta}{2})$, where d_{dens} d_{dens} is the function from Corollary 12. We then see that for all d for which $d \ge \mathsf{d}(\beta)$, then $\beta d \ge d_{\text{dens}}(\frac{\beta}{2})$, and thus Corollary 12 yields the existence of constant p_0 for which $\mathbb{P}[\lambda_{\min}(\mathbf{W}) \ge 1] \ge p_0$, and $\mathbb{P}_{\widetilde{\mathbf{W}} \sim \text{Wishart}(T-d)}[\lambda_{\min}(\widetilde{\mathbf{W}}) \le \frac{\beta}{2}] \ge \sqrt{\beta/2}p_0$. Hence, setting $\mathrm{c}_{\text{wish}} = \frac{p_0^2}{2\sqrt{2}}$, we conclude

$$\begin{aligned} p_{\text{err}} &\ge \frac{\mathbb{P}[\lambda_{\min}(\widetilde{\mathbf{W}}) \le \frac{1}{2}] \cdot \mathbb{P}[\lambda_{\min}(\mathbf{W}) \ge 1]}{2} \\ &\ge \frac{\mathbb{P}_{\widetilde{\mathbf{W}} \sim \text{Wishart}(T-d)}\left[\lambda_{\min}(\widetilde{\mathbf{W}}) \le \frac{1-\beta}{2}\right] \cdot \mathbb{P}[\lambda_{\min}(\mathbf{W}) \ge 1]}{2} \\ &\ge \frac{\sqrt{\beta/2}p_0 \cdot p_0}{2} = \frac{p_0^2}{2\sqrt{2}}\sqrt{\beta} = \mathrm{c}_{\text{wish}}\sqrt{\beta}. \end{aligned}$$

■

Acknowledgements MB's research is supported in part by the NSF Alan T. Waterman Award, Grant No. 1933331, a Packard Fellowship in Science and Engineering, and the Simons Collaboration on Algorithms and Geometry. EH is supported in part by NSF grant #1704860. MS is supported by an Open Philanthropy AI Fellowship, and this work was conducted while visiting Princeton University. BW is supported by the Google PhD fellowship program, and this work was conducted while an intern at Google AI Princeton. We thank Ramon Van Handel for his helpful discussions regarding random matrix theory.

References

Naman Agarwal and Elad Hazan. Lower bounds for higher-order convex optimization. *arXiv preprint arXiv:1710.10329*, 2017.

Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*, volume 118. Cambridge university press, 2010.

Kenneth L Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):1–45, 2017.

Dan Garber and Elad Hazan. Fast and simple pca via convex optimization. *arXiv preprint arXiv:1509.05647*, 2015.

Dan Garber, Elad Hazan, Chi Jin, Sham, Cameron Musco, Praneeth Netrapalli, and Aaron Sidford. Faster eigenvector computation via shift-and-invert preconditioning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2626–2634, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/garber16.html.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869, 2014.

Jacek Kuczyński and Henryk Woźniakowski. Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM journal on matrix analysis and applications*, 13 (4):1094–1122, 1992.

László Lovász and Santosh Vempala. Simulated annealing in convex bodies and an o*(n4) volume algorithm. *Journal of Computer and System Sciences*, 72(2):392–417, 2006.

Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems*, pages 1396–1404, 2015.

Arkadii Nemirovskii, David Borisovich Yudin, and Edgar Ronald Dawson. Problem complexity and method efficiency in optimization. 1983.

José A Ramírez and Brian Rider. Diffusion at the random matrix hard edge. *Communications in Mathematical Physics*, 288(3):887–906, 2009.

Yousef Saad. *Numerical methods for large eigenvalue problems: revised edition*, volume 66. Siam, 2011.

Sushant Sachdeva, Nisheeth K Vishnoi, et al. Faster algorithms via approximation theory. *Foundations and Trends® in Theoretical Computer Science*, 9(2):125–210, 2014.

Jianhong Shen. On the singular values of gaussian random matrices. *Linear Algebra and its Applications*, 326(1-3):1–14, 2001.

Max Simchowitz. On the randomized complexity of minimizing a convex quadratic function. *arXiv preprint arXiv:1807.09386*, 2018.

Max Simchowitz, Ahmed El Alaoui, and Benjamin Recht. Tight query complexity lower bounds for pca via finite sample deformed wigner law. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1249–1259. ACM, 2018.

Xiaoming Sun, David P Woodruff, Guang Yang, and Jialin Zhang. Querying a matrix through matrix-vector products. *arXiv preprint arXiv:1906.05736*, 2019.

Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

Virginia Vassilevska Williams. Multiplying matrices faster than coppersmith-winograd. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 887–898, 2012.

Blake Woodworth and Nathan Srebro. Lower bound for randomized first order convex optimization. *arXiv preprint arXiv:1709.03594*, 2017.

Blake E Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in neural information processing systems*, pages 3639–3647, 2016.

Appendix A. Proof of Proposition 8

The proof of Proposition 8 has two steps. First, we show that if Alg is LinSysAlg that can solve a linear system to *high precision* (in the Euclidean norm), then Alg implies the existance of an $\mathsf{Alg}_{\mathrm{eig}}$ which can recover the top eigenvector of a matrix up to roughly that precision:

Lemma 15 (Shift-and-Invert Reduction) *For parameters $\alpha > 0$ and $\mathtt{gap} \geq 1$, and recall the set*

$$\mathscr{M}_d(\mathtt{gap}, \alpha) := \left\{ M \in \mathbb{S}^d_{++} : \mathrm{gap}(M) \geq \mathtt{gap},\ |\lambda_1(M) - 1| \leq \alpha\, \mathtt{gap},\ \lambda_1(M) \in \left[\tfrac{1}{2}, 2\right] \right\},$$

and set $\mathtt{gap}_\alpha := \frac{1}{3+4\alpha}$ and $\mathtt{cond}_\alpha = \frac{1}{\mathtt{gap}} + (1+\alpha)$. Further, suppose Alg is a LinSysAlg *with query complexity T, and satisfies, for a given $\epsilon \in (0,1)$, and for all $A \in \mathbb{S}^d_{++}$ with $\mathrm{cond}(A) \leq \mathtt{cond}_\alpha$ and $b \in \mathbb{R}^d$,*

$$\mathbb{P}_{A,b,\mathsf{Alg}} \left[\|\widehat{x} - A^{-1}b\|_2^2 \leq \left(\tfrac{\epsilon \mathtt{gap}_\alpha}{5}\right)^2 \|A^{-1}b\|_2^2 \right] \geq 1 - \delta$$

Then, for any $\tau \geq 1$ and d for which $\epsilon \leq \frac{1}{\tau\sqrt{d}}$, and foran $R(\epsilon, \alpha) = \mathcal{O}\left(\frac{\log(1/\epsilon)}{\mathtt{gap}_\alpha}\right)$, there exists an an EigVecAlg, *$\mathsf{Alg}_{\mathrm{eig}}$, which has query complexity* $\mathrm{Query}(\mathsf{Alg}_{\mathrm{eig}}) \leq \mathrm{Query}(\mathsf{Alg}) \cdot R(\epsilon, \alpha)$, *and satisfies*

$$\mathbb{P}_{M,\mathsf{Alg}_{\mathrm{eig}}} \left[\widehat{v}^\top M \widehat{v} \geq \lambda_1(M)(1 - \epsilon^2) \right] \geq 1 - \delta R - \mathcal{O}\left(\tau^{-1}\right) - e^{-\Omega(d)}, \quad \forall M \in \mathscr{M}_d(\mathtt{gap}, \alpha)$$

This lemma is obtained by the so-called *shift-and-invert* procedure, which approximates $v_1(M)$ by running the power method on the A^{-1}, where $A = \gamma I - M$ is a "shifted" version of M for an appropriate shift parameter γ.

Second, we show that if Alg can solve a linear system to moderate $\mathcal{O}\left(\frac{1}{\mathrm{gap}}\right)$-precision in the $\|\cdot\|_A$, it can be bootrsaped to obtain high precision solutions in $\|\cdot\|_2$:

Lemma 16 (Bootstrapping Moderate Precision Solves) *Fix $\mathtt{cond} \geq 1$, and suppose Alg satisfies, for all $A : \mathrm{cond}(A) \leq \mathtt{cond}$,*

$$\mathbb{P}_{x_0,b,\mathsf{Alg}} \left[\|\widehat{x} - A^{-1}b\|_A^2 \leq \frac{\|x_0 - A^{-1}b\|^2 \lambda_1(A)}{e\mathtt{cond}} \right] \geq 1 - \tfrac{1}{e} \tag{5}$$

Then, any for any $\epsilon, \delta \in (0, 1/e)$, there exist a LinSysAlg, *Alg' with* $\mathrm{Query}(\mathsf{Alg}') \leq \mathrm{Query}(\mathsf{Alg}) \cdot \mathcal{O}\left(Q(\epsilon, \delta)\right)$ *which satisfies*

$$\mathbb{P}_{A,x_0,b,\mathsf{Alg}} \left[\|\widehat{x} - A^{-1}b\|_2^2 \leq \epsilon \|x_0 - A^{-1}b\|_2^2 \right] \geq 1 - \delta,$$

where $Q(\epsilon, \delta) := (\log \frac{1}{\epsilon}) \log(\frac{1}{\delta} \log \frac{1}{\epsilon})$,

Proposition 8 now follows from combining these two lemmas above

Proof Let $\tau = \mathcal{O}(d)$ and $d \geq d_{\min}$, for $d_{\min}$ to be sufficiently large that the term $\mathcal{O}(\tau) + e^{-\Omega(d)} \leq 1/2e$. For a parameter c, we define the following constants.

$$\epsilon = \min\{c\,\mathtt{gap}, 1/\tau\sqrt{d}\} = \Omega(\min\{d^{-3/2}, c\,\mathtt{gap}\})$$
$$\epsilon' := \left(\frac{\epsilon \mathtt{gap}_\alpha}{5}\right)^2, \quad \delta := \frac{1}{2eR(\epsilon, \alpha)}.$$

Now, suppose that Alg_0 is a $\mathrm{LinSysAlg}$ which satisfies (5) with query compexity $\mathrm{Query}(\mathsf{Alg}_0) \le T$. Then, by Lemma 16 the exists a $\mathrm{LinSysAlg}$ Alg with query complexity $T \cdot Q(\epsilon', \delta)$ satisfying

$$\mathbb{P}_{A,b,\mathsf{Alg}}\left[\|\widehat{x} - A^{-1}b\|_2^2 \le \epsilon' \|A^{-1}b\|_2^2\right] \ge 1 - \delta,$$

Hence, by Lemma 15, there exists an $\mathrm{EigVecAlg}$ $\mathsf{Alg}_{\mathrm{eig}}$ with query complexity $T \cdot Q(\epsilon', \delta) \cdot R(\epsilon, \alpha)$ which satisfies, for all $M \in \mathscr{M}_d(\mathtt{gap}, \alpha)$

$$\mathbb{P}_{M,\mathsf{Alg}_{\mathrm{eig}}}\left[\widehat{v}^\top M \widehat{v} \ge \lambda_1(M) - c\mathtt{gap}\right] \ge 1 - \delta - \mathcal{O}\left(\tau\right) + e^{-\Omega(d)} \ge 1 - \frac{1}{e}.$$

We can increasing the success probability of $\mathsf{Alg}_{\mathrm{eig}}$ to $\ge 1 - \delta$ by restarting $\mathsf{Alg}_{\mathrm{eig}}$ $L = \mathcal{O}\left(\log \frac{1}{\delta}\right)$ times to obtain $\widehat{v}^{(1)}, \dots, \widehat{v}^{(L)}$, and returning

$$\overline{\widehat{v}} := \arg\max\{v \in \{\widehat{v}^{(j)}\}_{j \in [L]} : v^\top M v\},$$

In total, this requires at most $L + LT \cdot Q(\epsilon', \delta) \cdot R(\epsilon, \alpha) = T\,\mathcal{O}\left((\log \frac{1}{\delta}) \cdot Q(\epsilon', \delta) \cdot R(\epsilon, \alpha)\right)$ queries.

We conclude by boudning $Q(\epsilon', \delta) \cdot R(\epsilon, \alpha)$. We have that

$$\begin{aligned}
R(\epsilon, \alpha) &= \mathcal{O}\left(\frac{\log(1/\epsilon)}{\mathrm{gap}_\alpha}\right) \le \mathcal{O}\left(\frac{\log(1/\mathtt{gap}_\alpha \epsilon)}{\mathrm{gap}_\alpha}\right) \\
Q(\epsilon', \delta) &= \left(\log \frac{1}{\epsilon'}\right) \cdot \left(\log \frac{1}{\delta}(\log \frac{1}{\epsilon'})\right) \\
&= \mathcal{O}\left(\left(\log \frac{1}{\mathtt{gap}_\alpha \epsilon}\right) \cdot \left(\log \frac{1}{R(\epsilon.\alpha)} + \log\log \frac{1}{\mathtt{gap}_\alpha \epsilon})\right)\right) \\
&= \mathcal{O}\left(\left(\log \frac{1}{\mathtt{gap}_\alpha \epsilon}\right) \cdot \left(\log \frac{1}{\mathtt{gap}_\alpha} + \log\log \frac{1}{\mathtt{gap}_\alpha \epsilon})\right)\right)
\end{aligned}$$

Hence

$$\begin{aligned}
Q(\epsilon', \delta) \cdot R(\epsilon, \alpha) &= \mathcal{O}\left(\frac{1}{\mathtt{gap}_\alpha} \log^2 \frac{1}{\mathtt{gap}_\alpha \epsilon}\left(\log \frac{1}{\mathtt{gap}_\alpha} + \log\log \frac{1}{\mathtt{gap}_\alpha \epsilon}\right)\right) \\
&= \mathcal{O}\left(\frac{\log \frac{1}{\mathtt{gap}_\alpha}}{\mathtt{gap}_\alpha} \log^{2+\log} \frac{1}{\mathtt{gap}_\alpha \epsilon}\right) \\
&= \mathcal{O}_\alpha\left(\log^{2+\log} \frac{d}{\min\{1,\, c\mathtt{gap}\}}\right),
\end{aligned}$$

where we recall the notation $\log^{p+\log}(x) = (\log^p x) \log\log x$.

■

A.1. Proof of Lemma 16

Recall the function $f(x) = \frac{1}{2} x^\top A x - b^\top x$, and note that

$$f(x) - f(A^{-1}b) = \|x - A^{-1}b\|_A^2.$$

Let $q \geq 1$ be a parameter to be selected later, and let Alg_q denote the algorithm which (a) runs Alg q times to obtain estimates $\widehat{x}^{(1)}, \dots, \widehat{x}^{(q)}$, and (b) makes at most q additional queries to find

$$\widehat{x} \in \arg\min \left\{ f(x) : x \in \{\widehat{x}^{(1)}, \dots, \widehat{x}^{(q)}\} \right\}. \tag{6}$$

Then, by independence of the internal randomness of Alg, we can ensure

$$\mathbb{P}_{x_0,b,\mathsf{Alg}_q}\left[\|\widehat{x} - A^{-1}b\|_A^2 \leq \frac{\|x_0\|^2 \lambda_1(A)}{e\mathtt{cond}} \right] \geq 1 - e^{-q}$$

using at most $Tq + q$ queries.

Next, oberve that if $\|\widehat{x} - A^{-1}b\|_A^2 \leq \frac{\|x_0\|^2\lambda_1(A)}{e\mathtt{cond}} \leq \frac{\|x_0\|^2\lambda_1(A)}{e\mathrm{cond}(A)} = \frac{1}{e}\|x_0\|^2\lambda_d(A)$, then $\|\widehat{x} - A^{-1}b\|_2^2 \leq \frac{1}{e}\|x_0 - A^{-1}b\|_2^2$. Hence, by repeating Alg_q k-times, each time setting x_0 for the j-th repetition to coincide with $\widehat{x}$ from the $j-1$st, we find that

$$\mathbb{P}_{x_0,b,\mathsf{Alg}_q}\left[\|\widehat{x} - A^{-1}b\|_A^2 \leq e^{-k}\frac{\|x_0 - A^{-1}b\|^2 \lambda_1(A)}{e\mathtt{cond}} \right] \geq 1 - ke^{-q}$$

Hence, setting $k = \log(\frac{1}{\epsilon})$ and $q = \log\frac{1}{\delta}\log(\frac{1}{\epsilon})$, we obtain the lemma.

A.2. Proof of Lemma 15

Let $M \in \mathscr{M}_d(\alpha, \mathtt{gap})$, so that $\lambda_1(M) \in [1/2, 2]$, $\mathrm{gap}(M) \geq \mathtt{gap}$ and $|M - I| \leq \alpha\mathtt{gap}$. We define the associated "shifted" matrix $A := (1 + (1+\alpha)\mathtt{gap})I - M$. Crucially, A^{-1} and M have the same top eigenvector, and their eigenvalues are related by the correspondence

$$\lambda_j(A^{-1}) = \frac{1}{1 + 2\alpha\mathtt{gap} - \lambda_j(M)}$$

We can therefore compute the eigengap of A via

$$\begin{aligned}\frac{\lambda_2(A^{-1})}{\lambda_1(A^{-1})} = \frac{1 + (1+\alpha)\mathtt{gap} - \lambda_1(M)}{1 + (1+\alpha)\mathtt{gap} - \lambda_2(M)} &= \frac{1 + (1+\alpha)\mathtt{gap} - \lambda_1(M)}{1 + (1+\alpha)\mathtt{gap} - (1 - \mathtt{gap})\lambda_1(M)} \\ = \frac{1}{1 + \frac{\mathtt{gap}\lambda_1(M)}{1 + (1+\alpha)\mathtt{gap} - \lambda_1(M)}} &\overset{(i)}{\leq} \frac{1}{1 + \frac{\mathtt{gap}\lambda_1(M)}{(1+2\alpha)\mathtt{gap}}} \overset{(ii)}{\leq} \frac{1}{1 + \frac{1}{2(1+2\alpha)}}\end{aligned}$$

where (i) uses $|\lambda_1(M) - 1| \leq \alpha\mathtt{gap}$, and (ii) uses $\lambda_1(M) \geq 1/2$. Hence,

$$\mathrm{gap}(A^{-1}) \geq \frac{\frac{1}{2(1+2\alpha)}}{1 + \frac{1}{2(1+2\alpha)}} = \frac{1}{1 + 2(1+2\alpha)} = \frac{1}{3+4\alpha} := \mathtt{gap}_\alpha$$

In other words, the eigengap of A^{-1} depends on the parameter α, but *not* on the eigengap of M. Hence we can effectively run the power method on A^{-1} to compute the top eigenvector of M. Of course, we cannot query A^{-1}, but we can approximate a query $A^{-1}v$ by using a $\mathrm{LinSysAlg}$. To facillitate this reduction, we observe that $\mathrm{cond}(A) = \mathcal{O}\left(\mathtt{gap}^{-1}\right)$:

Claim 17 $\mathrm{cond}(A) \leq \mathtt{cond}_\alpha := \frac{1}{\mathtt{gap}} + (1+\alpha)$.

Proof Since $M \preceq 0$, $\lambda_1(A) \le 1 + (1+\alpha)\mathtt{gap}$. Moreover, since $|\lambda_1(M) - 1| \le \alpha\mathtt{gap}$, $\lambda_{\min}(A) = 1 + 2\alpha\mathtt{gap} - \lambda_1(M) \ge \mathtt{gap}$. ■

We our now ready to present the reduction. Let Alg be LinSysAlg satisfying the condition $\mathbb{P}_{A,b,\mathsf{Alg}}\left[\|\widehat{x} - A^{-1}b\|_2^2 \le \left(\frac{\epsilon\mathtt{gap}_\alpha}{5}\right)^2 \|A^{-1}b\|_2^2\right] \ge 1-\delta$ for all $A : \mathrm{cond}(A) \ge \mathtt{cond}_\alpha$ and $b \in \mathbb{R}^d$,

For a round number $R \ge 1$ to be selected later, precision ϵ, and failure probability δ, we define a procedure $\mathsf{Alg}_{\mathrm{eig}}$ in Algorithm 1, which uses Alg as a primitive to run an approximate power method on A^{-1}, up to the errors:

$$\Delta_r := \|\widehat{x}^{(r)} - A^{-1}u_r\|_2.$$

```
Algorithm 1: $\mathsf{Alg}_{\mathrm{eig}}$ (Approximate Power Method via $\mathsf{Alg}$)
Input: Confidence Parameter $\epsilon$, accuracy parameter $\delta$
Draw $u_0 \overset{\mathrm{unif}}{\sim} \mathcal{S}^{d-1}$
for rounds $r = 1, 2, \dots, R$ do
    call $\mathsf{Alg}$ to obtain $\widehat{x}^{(r)}$ such that
        $\mathbb{P}\left[\|\widehat{x}^{(r)} - A^{-1}u_{r-1}\|_2^2 \le \left(\frac{\epsilon\mathtt{gap}_\alpha}{5}\right)^2 \|A^{-1}u_{r-1}\|_2^2\right] \ge 1-\delta$    (7)
    Set $u_r := \widehat{x}^{(r)}/\|\widehat{x}^{(r)}\|_2$
Return $\widehat{v} = u_r$.
```

This "noisy" power method admits a black-box analysis due to Hardt and Price (2014):

Lemma 18 (Corollary 1.1 in Hardt and Price (2014), $k = p = 1$, specialized to Algorithm 1) *Fix a parameter $\tau > 1$, and an $\epsilon \le \frac{1}{\tau\sqrt{d}}$. Then, if*

$$\Delta_r \le \left(\frac{\lambda_1(A^{-1}) - \lambda_2(A^{-1})}{5}\right)\epsilon,$$

then for an $R = \mathcal{O}\left(\frac{\log(d\tau/\epsilon)}{\mathrm{gap}(A^{-1})}\right)$, $\sqrt{1 - \langle u_R, v_1(M)\rangle^2} \le \epsilon$ with probability $1 - \mathcal{O}\left(\tau^{-1}\right) - e^{-\Omega(d)}$ over the draw of x_0, where c is a universal constant.

We first interpret the bound $\sqrt{1 - \langle u_R, v_1(M)\rangle^2}$ in terms of the subotimality $\lambda_1(M) - u_R^\top M u_R$. Since $0 \preceq M \preceq 2I$, we have that if the conclusion of Lemma 18 is satisfied,

$$\begin{aligned} u_R^\top M u_R &= \lambda_1(M)(u_R^\top v_1(M))^2 + \sum_{i=2}^d \lambda_i(M) \cdot (u_R^\top v_i(M))^2 \\ &\ge \lambda_1(M)(u_R^\top v_1(M))^2 = \lambda_1(M)(1 + (1 - (u_R^\top v_1(M))^2)) = \lambda_1(M)(1-\epsilon^2). \end{aligned}$$

We can now conclude the proof by verifying

$$\begin{aligned}
&\mathbb{P}[u_R^\top M u_R \geq \lambda_1(M) - 2\epsilon^2] + \mathcal{O}\left(\tau\right) + e^{-\,\Omega(d)} \\
&\geq \mathbb{P}\left[\forall r \in [R] \Delta_r \leq \left(\tfrac{\lambda_1(A^{-1}) - \lambda_2(A^{-1})}{5}\right)\epsilon\right] \\
&= \mathbb{P}\left[\forall r \in [R] \Delta_r \leq \lambda_1(A^{-1})\left(\tfrac{\mathrm{gap}(A^{-1})}{5}\right)\epsilon\right] \\
&\overset{(i)}{\geq} \mathbb{P}\left[\forall r \in [R] \Delta_r \leq \|A^{-1}u_{r-1}\|_2 \left(\frac{\mathtt{gap}_\alpha}{5}\right)\epsilon\right] \overset{(ii)}{\geq} R\delta,
\end{aligned}$$

where (i) follows from the bound $\mathrm{gap}(A^{-1}) \leq \mathtt{gap}_\alpha$ and $\|A^{-1}u_{r-1}\|_2 \leq \|u_{r-1}\|_2 \lambda_1(A^{-1}) = \lambda_1(A^{-1})$, and (ii) by a union bound over the event in (7), with $R = \mathcal{O}\left(\frac{\log(d\tau/\epsilon)}{\mathrm{gap}(A^{-1})}\right) = \mathcal{O}\left(\frac{\log(1/\epsilon)}{\mathtt{gap}_\alpha}\right)$, where we recall $\epsilon \leq \frac{1}{\tau\sqrt{d}}$.

Appendix B. Proof of Theorem 6 for Arbitrary Condition Number

In this section, we given proof of Theorem 6 for general condition number. Using Proposition 8 directly for matrices with larger gap incurs a dimension on $\log$ of the ambient dimension.

To sharpen this, we state a slightly refined reduction. For this to go through, define, for a subspace $\mathcal{V} \subset \mathbb{R}^d$, let

$$\mathscr{M}_d(\mathtt{gap}, \alpha, \mathcal{V}) := \{M \in \mathscr{M}_d(\mathtt{gap}, \alpha) : v_1(M) \in \mathcal{V}\}$$

to denote the restriction of $\mathscr{M}_d(\mathtt{gap}, \alpha)$ to matrices whose top eigenvector is *know* to lie in a subspace $\mathcal{V}$. For this class, we can improve Proposition 8 as follows:

Proposition 19 (Eigenvector-to-Linear-System Reduction, Known Subspace) *Fix a* $\mathtt{gap} \in (0, 1)$*,* $\alpha > 0$*, and suppose that* Alg *be a* LinSysAlg *which which satisfies* (1) *with* $\mathtt{cond} := 1 + \alpha + \frac{1}{\mathtt{gap}}$ *for all* $A \in \mathbb{S}^d_{++}$ *with* $\mathrm{cond}(A) \leq \mathtt{cond}$*. Then, for any* $\delta \in (0, 1/e)$*, there exists an* EigVecAlg*,* $\mathsf{Alg}_{\mathrm{eig}}$*, which satisfies*

$$\mathbb{P}_{\mathsf{Alg}_{\mathrm{eig}}, M}\left[\widehat{v}^\top M \widehat{v} \geq (1 - c\mathtt{gap})\lambda_1(M)\right] \geq 1 - \delta, \quad \forall M \in \mathscr{M}_d\left(\mathtt{gap}, \alpha, \mathcal{V}\right)$$

with query complexity at most

$$\mathrm{Query}(\mathsf{Alg}_{\mathrm{eig}}) \leq \mathrm{Query}(\mathsf{Alg}) \cdot \mathcal{O}_\alpha\left((\log\frac{1}{\delta}) \cdot \log^{2+\log}\frac{\dim(\mathcal{V})}{\min\{c\mathtt{gap},\, 1\}}\right),$$

where $\mathcal{O}_\alpha\left(\cdot\right)$ *hides multiplicative and additive constants depending on* α*.*

Proof The proof is nearly identical to the proof of Theorem 8. The only difference is that, by initializing u_0 to be uniform on $\mathcal{S}^{d-1} \cap \mathcal{V}$, the guarantee of the noisy power method (Lemma 18) can be improved to depend on $\dim(\mathcal{V})$ instead of the ambient dimension d. ■

The proof of Theorem 6 for general $\mathtt{cond}$ is as now as follows: Fix $s \in [d_0(1/2) \vee d_{\min}, d]$. Let $\beta = \frac{1}{2}$, $\mathbf{M} \sim \mathcal{D}(s, s, \beta)$, and define the embedded matrix $\overline{\mathbf{M}} = \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{d \times d}$. Then, $v_1(\overline{\mathbf{M}})$

lies in the s-dimensional subspace $\mathcal{V}$ corresponding to the first s entries, it is easy to check that $\overline{\mathbf{M}} \in \mathscr{M}_d(\mathtt{gap}, \alpha, \mathcal{V})$, where $\mathtt{gap} := \frac{c_{\mathrm{gap}}(\beta)}{s^2}$ and $\alpha := \frac{c_{\mathrm{eig}}(\beta)}{c_{\mathrm{gap}}(\beta)}$ analogously to the $s = d$ case of Theorem 6.

Retracing the steps, and replacing the dependence of d with s, we find if Alg satisfies the guarantee of (1) for all $\mathrm{cond}(A) \le \mathtt{cond} := 1 + \alpha + \frac{1}{\mathtt{gap}} = \Theta(s^2)$,

$$\mathrm{Query}(\mathsf{Alg}) \ge \frac{c_1}{\sqrt{\mathtt{gap}}} \cdot \Omega\left((\log^{2+\log} s)^{-1}\right) = \Omega\left(s(\log^{2+\log} s)^{-1}\right).$$

Appendix C. Omitted Proofs from Section 3

C.1. Proof of Corollary 12

For the first point, fix a $\delta \in (0,1)$. Then by Lemma 11, the limiting normalized distributions of the eigenvalues $(\mathbf{z}_d, \mathbf{z}_{d-1})$ satisfy $\mathbb{P}[\{\mathbf{z}_d \ge C_1(\delta)\} \cap \{\mathbf{z}_{d-1} - \mathbf{z}_d \le C_2(\delta)\}] \le \frac{\delta}{3}$ for appropriate constants $C_1(\delta), C_2(\delta)$. By convergence in distribution, and the fact that $\{z_1 \ge C_1(\delta)\} \cap \{z_2 - z_1 \le C_2(\delta)\}$ corresponds to the event that (z_1, z_2) lie in a closed set, $\lim_{d\to\infty} \mathbb{P}[\{\mathbf{z}_d^{(d)} \ge C_1(\delta)\} \cap \{\mathbf{z}_{d-1}^{(d)} - \mathbf{z}_d^{(d)} \le C_2(\delta)\}] \le \frac{\delta}{3}$, so that for all d sufficently large as a function of δ, $\mathbb{P}[\{\mathbf{z}_d^{(d)} \ge C_1(\delta)\} \cap \{\mathbf{z}_{d-1}^{(d)} - \mathbf{z}_d^{(d)} \le C_2(\delta)\}] \le \frac{2\delta}{3}$. Finally, for all d sufficiently large as a function of δ, we have $\mathbb{P}[\lambda_{\max}(\mathbf{W}) \ge 5] \le \frac{\delta}{3}$. The result now follows from a union bound.

For the second point, we use that the limiting distribution of $d^2\lambda_1(\mathbf{W})$ has density $f(x) = \mathbb{I}(x \ge 0) \cdot \frac{x^{-1/2}+1}{2} e^{-(x/2+\sqrt{x})}$. Recalling the notation $\mathbf{z}_d$ for a random variable with said limiting distribution, integrating the density shows that there exists a constant p such that, for all for all $\alpha \in (0,1)$, $\mathbb{P}[\mathbf{z}_d \ge 1] \ge p$ and $\mathbb{P}[\mathbf{z}_d \le \alpha^2] \ge p\alpha$. The bound now follows by invoking convergence in distribution and setting, say $p_0 = p/2$.

C.2. Proof of Lemma 13

Without loss of generality, the query vectors $v^{(1)}, \ldots, v^{(T)}$ are orthonormal because the response to any sequence of queries can be calculated from the responses to a sequence of orthonormal queries.

We define a sequence of matrices $\mathbf{V}_1, \ldots, \mathbf{V}_T$ such that for each $t \le T$, the product $\mathbf{V}_{1:t} = \mathbf{V}_t\mathbf{V}_{t-1}\ldots\mathbf{V}_1$ is an orthonormal matrix whose first t rows are $v^{(1)}, \ldots, v^{(t)}$. This can be accomplished by choosing $\mathbf{V}_t$ to be an arbitrary orthonormal matrix whose first $t-1$ rows are $e_1, \ldots, e_{t-1}$, and whose t^{th} row is chosen so that $\mathbf{V}_t[t,:]\mathbf{V}_{1:t-1} = v^{(t)\top}$. This is always possible because each $\mathbf{V}_t$ is a rotation matrix and thus $\mathbf{V}_{1:t-1}$ is full rank and orthonormal. Importantly, the matrix $\mathbf{V}_t$ can be constructed as a function of $v^{(1)}, \ldots, v^{(t)}$ only, and does not depend in any way on the later queries $v^{(t+1)}, \ldots, v^{(T)}$.

Similarly, we define another sequence of matrices $\mathbf{R}_1, \ldots, \mathbf{R}_T$ such that for each $t \le T$, the product $\mathbf{R}_{1:t} = \mathbf{R}_1\mathbf{R}_2\ldots\mathbf{R}_t$ is an orthonormal matrix whose first t columns form an orthonormal basis for the first t rows of $\mathbf{V}_{1:t}\mathbf{X}$. This can be accomplished by choosing $\mathbf{R}_t$ to be an arbitrary orthonormal matrix whose first $t-1$ columns are $e_1, \ldots, e_{t-1}$, and whose t^{th} column is the (normalized) component of the t^{th} row of $\mathbf{V}_{1:t}\mathbf{X}$ that lies outside the span of the first $t-1$ rows of $\mathbf{V}_{1:t}\mathbf{X}$. Importantly, the matrix $\mathbf{R}_t$ can be constructed as a function of the first t rows of $\mathbf{V}_{1:t}\mathbf{X}$ only.

By the construction of $\mathbf{V}_{1:t}$ and $\mathbf{R}_{1:t}$, for each $t \leq T$, querying $v^{(t)}$ and observing the response $w^{(t)} = \mathbf{X}\mathbf{X}^\top v^{(t)}$ is equivalent to querying e_t and observing the response $\widetilde{w}^{(t)} = \mathbf{V}_{1:t} w^{(t)} = \mathbf{V}_{1:t}\mathbf{X}\mathbf{R}_{1:t}\mathbf{R}_{1:t}^\top \mathbf{X}^\top \mathbf{V}_{1:t}^\top e_t$.

Let $\mathbf{V}^{\|}_{1:t}$ denote the first t rows of $\mathbf{V}_{1:t}$, and let $\mathbf{V}^{\perp}_{1:t}$ denote the remaining $d-t$ rows. Similarly, let $\mathbf{R}^{\|}_{1:t}$ denote the first t columns of $\mathbf{R}_{1:t}$, and let $\mathbf{R}^{\perp}_{1:t}$ denote the remaining $d-t$ columns. Then, for any $t \leq T$ we can decompose

$$\left(\mathbf{V}_{1:t}\mathbf{X}\mathbf{R}_{1:t}\right)\left(\mathbf{V}_{1:t}\mathbf{X}\mathbf{R}_{1:t}\right)^\top = \begin{bmatrix} \left(\mathbf{V}^{\|}_{1:t}\mathbf{X}\mathbf{R}^{\|}_{1:t}\right)\left(\mathbf{V}^{\|}_{1:t}\mathbf{X}\mathbf{R}^{\|}_{1:t}\right)^\top & \left(\mathbf{V}^{\|}_{1:t}\mathbf{X}\mathbf{R}^{\|}_{1:t}\right)\left(\mathbf{V}^{\perp}_{1:t}\mathbf{X}\mathbf{R}^{\|}_{1:t}\right)^\top \\ \left(\mathbf{V}^{\perp}_{1:t}\mathbf{X}\mathbf{R}^{\|}_{1:t}\right)\left(\mathbf{V}^{\|}_{1:t}\mathbf{X}\mathbf{R}^{\|}_{1:t}\right)^\top & \begin{matrix}\left(\mathbf{V}^{\perp}_{1:t}\mathbf{X}\mathbf{R}^{\|}_{1:t}\right)\left(\mathbf{V}^{\perp}_{1:t}\mathbf{X}\mathbf{R}^{\|}_{1:t}\right)^\top \\ +\left(\mathbf{V}^{\perp}_{1:t}\mathbf{X}\mathbf{R}^{\perp}_{1:t}\right)\left(\mathbf{V}^{\perp}_{1:t}\mathbf{X}\mathbf{R}^{\perp}_{1:t}\right)^\top\end{matrix} \end{bmatrix} \tag{8}$$

We will now prove the lemma by induction.

Base case: Recall that $\mathbf{V}_1$ is constructed solely as a function of $v^{(1)}$, which is independent of $\mathbf{X}$. Therefore, $\mathbf{V}^{\|}_1\mathbf{X}$ is independent of $\mathbf{V}^{\perp}_1\mathbf{X}$. The matrix $\mathbf{R}_1$ is constructed as a function of $\mathbf{V}^{\|}_1\mathbf{X}$, and is thus independent of $\mathbf{V}^{\perp}_1\mathbf{X}$. Consequently, $\mathbf{V}^{\perp}_1\mathbf{X}\mathbf{R}^{\|}_1$ is independent of $\mathbf{V}^{\perp}_1\mathbf{X}\mathbf{R}^{\perp}_1$.

Finally, we observe from (8) that $\widetilde{w}^{(1)} = \mathbf{V}_1\mathbf{X}\mathbf{R}_1\mathbf{R}_1^\top\mathbf{X}^\top\mathbf{V}_1^\top e_1$ is measurable with respect to the matrices $\mathbf{V}^{\|}_1\mathbf{X}\mathbf{R}^{\|}_1$ and $\mathbf{V}^{\perp}_1\mathbf{X}\mathbf{R}^{\|}_1$, both of which are independent of $\mathbf{V}^{\perp}_1\mathbf{X}\mathbf{R}^{\perp}_1$.

We conclude that $\mathbf{V}^{\perp}_1\mathbf{X}\mathbf{R}^{\perp}_1$ is independent of both the first query $v^{(1)}$ and the first observation $w^{(1)}$, and thus has a entries $(\mathbf{V}^{\perp}_1\mathbf{X}\mathbf{R}^{\perp}_1)_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{d})$ conditioned on $\mathsf{v}^{(1)} = v^{(1)}$ and $\mathsf{w}^{(1)} = w^{(1)}$.

Inductive step: Suppose that for all $t < T$ the matrix $\mathbf{V}^{\perp}_{1:t}\mathbf{X}\mathbf{R}^{\perp}_{1:t}$ has entries entries which are $(\mathbf{V}^{\perp}_{1:t}\mathbf{X}\mathbf{R}^{\perp}_{1:t})_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{d})$ conditioned on the first t queries and responses.

The algorithm's T^{th} query is a function of the first $T-1$ queries and observations, thus $\mathbf{V}_T$ and $\mathbf{V}_{1:T}$ are independent of $\mathbf{V}^{\perp}_{1:T-1}\mathbf{X}\mathbf{R}^{\perp}_{1:T-1}$ by the inductive hypothesis. Therefore, $\mathbf{V}^{\|}_{1:T}\mathbf{X}\mathbf{R}^{\perp}_{1:T-1}$ is independent of $\mathbf{V}^{\perp}_{1:T}\mathbf{X}\mathbf{R}^{\perp}_{1:T-1}$. The matrices $\mathbf{R}_T$ and $\mathbf{R}_{1:T}$ are constructed as a function of $\mathbf{V}^{\|}_{1:T}\mathbf{X}$ and are thus independent of $\mathbf{V}^{\perp}_{1:T}\mathbf{X}\mathbf{R}^{\perp}_{1:T-1}$. We conclude that $\mathbf{V}^{\perp}_{1:T}\mathbf{X}\mathbf{R}^{\|}_{1:T}$ is independent of $\mathbf{V}^{\perp}_{1:T}\mathbf{X}\mathbf{R}^{\perp}_{1:T}$.

Finally, we observe from (8) that $\widetilde{w}^{(T)} = \mathbf{V}_{1:T}\mathbf{X}\mathbf{R}_{1:T}\mathbf{R}_{1:T}^\top\mathbf{X}^\top\mathbf{V}_{1:T}^\top e_T$ is measurable with respect to $\mathbf{V}^{\|}_{1:T}\mathbf{X}\mathbf{R}^{\|}_{1:T}$ and $\mathbf{V}^{\perp}_{1:T}\mathbf{X}\mathbf{R}^{\|}_{1:T}$, both of which are independent of $\mathbf{V}^{\perp}_{1:T}\mathbf{X}\mathbf{R}^{\perp}_{1:T}$.

Therefore, by induction $\mathbf{V}^{\perp}_{1:T}\mathbf{X}\mathbf{R}^{\perp}_{1:T}$ is independent of the algorithm's T queries and thus has entries $(\mathbf{V}^{\perp}_{1:T}\mathbf{X}\mathbf{R}^{\perp}_{1:T})_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{d})$ for $1 \leq i, j \leq d - T$ conditioned on $\mathsf{v}^{(1)} = v^{(1)}, \ldots, \mathsf{v}^{(T)} = v^{(T)}$ and $\mathsf{w}^{(1)} = w^{(1)}, \ldots, \mathsf{w}^{(T)} = w^{(T)}$. Hence, $\widetilde{\mathbf{W}} := (\mathbf{V}^{\perp}_{1:T}\mathbf{X}\mathbf{R}^{\perp}_{1:T})(\mathbf{V}^{\perp}_{1:T}\mathbf{X}\mathbf{R}^{\perp}_{1:T})^\top$ satisfies $(\frac{d}{d-T}) \cdot \widetilde{\mathbf{W}} \sim \operatorname{Wishart}(d-T)$.