Tight Query Complexity Lower Bounds for PCA via Finite Sample Deformed Wigner Law

Max Simchowitz
U.C Berkeley
Berkeley, C.A., U.S.A.
msimchow@berkeley.edu

Ahmed El Alaoui
U.C Berkeley
Berkeley, C.A., U.S.A.
elalaoui@berkeley.edu

Benjamin Recht
U.C Berkeley
Berkeley, C.A., U.S.A.
brecht@berkeley.edu

ABSTRACT

We prove a query complexity lower bound for approximating the top $r$ dimensional eigenspace of a matrix. We consider an oracle model where, given a symmetric matrix $M \in \mathbb{R}^{d \times d}$, an algorithm Alg is allowed to make $T$ exact queries of the form $w(i) = Mv(i)$ for $i$ in $\{1, \ldots, T\}$, where $v(i)$ is drawn from a distribution which depends arbitrarily on the past queries and measurements $\{v(j), w(i)\}_{1 \leq j \leq i}$. We show that for every gap $\eta \in (0, 1/2]$, there exists a distribution over matrices $M$ for which 1) gap$_r$($M$) = $\Omega$($\eta$) (where gap$_r$($M$) is the normalized gap between the $r$ and $r + 1$-st largest-magnitude eigenvector of $M$), and 2) any algorithm Alg which takes fewer than $c \cdot \log d / \sqrt{\eta}$ queries fails (with overwhelming probability) to identity a matrix $V \in \mathbb{R}^{d\times r}$ with orthonormal columns for which $\langle V, M V \rangle \geq \{1 - \text{const} \times \text{gap} \} \sum_{i=1}^{\eta} \lambda_i$. Our bound requires only that $d$ is a small polynomial in $1/\text{gap}$ and $r$, and matches the upper bounds of Musco and Musco '15. Moreover, it establishes a strict separation between convex optimization and randomized, "strict-saddle" non-convex optimization of which PCA is a canonical example: in the former, first-order methods can have dimension-free iteration complexity, whereas in PCA, the iteration complexity of gradient-based methods must necessarily grow with the dimension.

Our argument proceeds via a reduction to estimating a rank-$r$ spike in a deformed Wigner model $M = W + \lambda UU^\top$, where $W$ is from the Gaussian Orthogonal Ensemble, $U$ is uniform on the $d \times r$-Stiefel manifold and $\lambda > 1$ governs the size of the perturbation. Surprisingly, this ubiquitous random matrix model witnesses the worst-case rate for eigenspace approximation, and the "accelerated" inverse square-root dependence on the gap in the rate follows as a consequence of the correspondence between the asymptotic eigen-gap and the size of the perturbation $\lambda$, when $\lambda$ is near the "phase transition" $\lambda = 1$. To verify that $d$ need only be polynomial in $1/\text{gap}$ and $r$, we prove a finite sample convergence theorem for top eigenvalues of a deformed Wigner matrix, which may be of independent interest. We then lower bound the above estimation problem with a novel technique based on Fano-style data-processing inequalities with truncated likelihoods; the technique generalizes the Bayes-risk lower bound of Chen et al. '16, and we believe it is particularly suited to lower bounds in adaptive settings like the one considered in this paper.

CCS CONCEPTS

• Theory of computation → Nonconvex optimization;

KEYWORDS

Lower Bounds, Query Complexity, PCA, Optimization, Random Matrix Theory

ACM Reference Format:
Max Simchowitz, Ahmed El Alaoui, and Benjamin Recht. 2018. Tight Query Complexity Lower Bounds for PCA via Finite Sample Deformed Wigner Law. In Proceedings of 50th Annual ACM SIGACT Symposium on the Theory of Computing (STOC’18). ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3188745.3188796

1 INTRODUCTION

Eigenvector approximation is widely regarded as a fundamental problem in machine learning [21], numerical linear algebra [15], optimization, and numerous graph-related learning problems [26, 27, 32]. Interest in PCA has been driven further by the rush to understand non-convex optimization, as PCA has become the canonical example of a benign, but not-quite-convex objective. For one, there is a striking resemblance between eigenvector approximation algorithms and first-order convex optimization procedures [3, 17, 30]. Moreover, PCA is one of the simplest ‘strict saddle’ objectives: that is, a function whose first-order stationary points are either local minima, or saddle points at which the Hessian has a strictly negative eigenvalue. The strict saddle property extends to many popular nonconvex objectives, and enables efficient optimization by first order algorithms. Notably, Jin et al. [20] proposed a gradient algorithm which finds an approximate local minimum of a strict saddle objective in a number of iterations which matches first-order methods for comparable convex problems, up to poly-logarithmic factors in the dimension.

The aim of this paper is to understand the fundamental limits of randomized first-order methods for such benign non-convex problems by establishing sharp query-complexity lower bounds for approximating the top eigenspace of a symmetric matrix. Specifically, we consider randomized, adaptive algorithms Alg which access an unknown symmetric matrix $M \in \mathbb{R}^{d \times d}$ via $T$ queries of the form $\{v(i) = Mw(i)\}_{i \in \{1, \ldots, T\}}$. Letting gap$_r$($M$) denotes the (normalized) eigengap between the $r$- and $r + 1$-st singular value (or
eigenvalue-magnitude) of $M$, let $\text{abs}(M) := (M^2)^{1/2}$, we prove the following:

**Theorem 1 (Main Theorem).** There are universal constants $c_1, c_2 > 0$ such that for every $r \geq 1$ and gap $\in (0, 1/2]$, then there exist $d_0 = \text{poly}(\frac{1}{\text{gap}^2})$ such that, for all $d \geq d_0$, there exists a distribution over symmetric matrices $M \in \mathbb{R}^{d \times d}$ for which $\text{gap}(M) \geq \frac{1}{\text{gap}}$, and if Alg makes $T \leq r \log d \sqrt{\frac{\text{gap}}{\text{abs}(M)}}$ queries, then with probability at least $1 - \exp(-d^2)$, Alg cannot identify a matrix $\hat{V} \in \mathbb{R}^{d \times r}$ with orthonormal columns for which $(\hat{V}, \text{abs}(M)\hat{V}) \geq \left(1 - \frac{\text{abs}(M)}{2}\right) \sum_i \lambda_i(\text{abs}(M))$.\(^1\)

We emphasize that our lower bounds are information-theoretic, and do not place any computational or Krylov restrictions on how Alg generates its queries. Our bounds are tight, and are matched by the Block-Lanczos algorithm [23]. Finally, the presence of a logarithmic factor in the dimension establishes a strict separation between truly-convex and strict saddle objectives: whereas convex objectives admit first order algorithms whose query complexity is independent of the ambient dimension, strict saddle-objects necessarily incur dimension-dependent terms, even for randomized algorithms.

**New Techniques.** Our lower bound proceeds by a reduction from eigenspace computation to estimating a planted rank-$r$ component in a deformed Wigner model $M = W + \lambda UU^T$, where $W$ is from the Gaussian Orthogonal Ensemble (GOE), $U$ is uniform on the $d \times r$ Stiefel manifold, and $\lambda = \frac{1}{\sqrt{d}} \sqrt{\text{gap}^2 - \text{abs}(M)}$ is a parameter ensuring that $\text{gap}(M)$ concentrates around gap. Note that as $\text{gap} \to 0$, $\lambda \to 1$ placing us near the “phase transition” $\lambda = 1$ [16]. To ensure that we can take $d = \text{poly}(\frac{1}{\text{gap}}, r)$, we prove the first (to our knowledge) finite-sample convergence result for the top $r$ eigenvalues of a deformed Wigner matrix in the regime where $d$ is polynomial in $r$ and gap$^{-1}$. Along the way, we prove a variant of the Hanson-Wright inequality for the Stiefel manifold, and a pointwise convergence result for the Stieltjes transform; these results are outlined in Section 6.

After formalizing the reduction, our proof hinges on showing that when $r = 1$ and $U = u \in \mathbb{R}^d$, then our “information” about $u$, quantified by the squared-norm of the projection of $u$ onto the span of the first $k$ query vectors, can grow at a rate of at most $\lambda O(1) = 1 + O\left(\text{gap}^{-2}\right)$ per round. We generalize to the rank-$r$ case by leveraging the information-theoretic arguments from the rank-one case, but with a far more careful recursion to obtain the right dependence on $r$ (see Section 5 for details). For $r = 1$, our basic strategy mirrors Price and Woodruff’s [28] sparse recovery lower bound, which sequentially controls the mutual information between measurements of a sparse vector and a planted solution. However, since $\lambda = 1 + O\left(\text{gap}^{-1}\right)$, we require novel techniques in order to not overshoot the slow growth rate of $\lambda O(1)$ per round. Specifically, at each round $k \in [T]$, we apply a generalization of Fano’s inequality which replaces the KL-divergence with the expected $1 + \eta$-powers of appropriate likelihood ratios. The inequality is based on the Bayes risk lower bounds of Chen et al. [13], who generalize Fano’s inequality to arbitrary $f$-divergences, and show that their $\chi^2$-variant of Fano’s inequality (i.e. $\eta = 1$) yields sharper lower bounds in many non-adaptive problems. In our case, we tune $\eta$ as a function of $\lambda$ to get the correct rate.

Unfortunately, we cannot apply the bounds from Chen et al. [13] out of the box. This is because if there is even a small probability that Alg takes one highly informative measurement, then the expected likelihood ratios will overestimate the average information gain. This is an artifact of the fact that tails of likelihood ratios (unlike log-likelihoods) are ill-behaved. But this only becomes a problem in adaptive settings where measurements grow more informative over time. We circumvent this by proving a “truncated” variant of the bound in Chen et al. [13], which replaces the expected likelihood moments with an expectation restricted to the “good event” where Alg has yet to take an improbably-informative measurement. We prove this bound by generalizing $f$-divergences to arbitrary finite, non-normalized measures (e.g., measures obtained by restricting probability distributions to a given event), and establish that the data-processing inequality still holds in this general setting. Our information-theoretic tools are explained at length in Section 4.3.

**Related Work.** It is hard to do justice to the vast body of work on eigenvector computation, matrix approximation, and first order methods for convex and strict saddle objectives. We shall instead focus on situating our work in the lower bounds literature. As described above, our proof casts eigenvector computation as a sequential estimation problem. These have been studied at length in the context of sparse recovery and active adaptive compressed sensing [7, 11, 12, 28]. Due to the noiseless oracle model, our setting is most similar to that of Price and Woodruff [28], whereas other works [7, 11, 12] study measurements contaminated with noise. More broadly, query complexity has received much recent attention in the context of communication-complexity [6, 24], in which lower bounds on query complexity imply corresponding bounds against communication via lifting theorems. Similar ideas also arise the study of learning under memory constraints [29, 33, 34].

From an optimization perspective, our lower bound can be cast as a non-convex analogue of the seminal convex-optimization oracle lower bounds of Nemirovskii and Yudin [25]. But whereas the latter bounds match known upper bounds in terms of dependence on relevant parameters (accuracy, condition number, Lipschitz constant), Nemirovskii and Yudin consider worst-case initializations, and impose a strong Krylov space assumption. In the context of finite sums, Agarwal et al. [1] show that the Krylov space assumption can be removed, and Woodworth et al. [36] prove true information-theoretic lower bounds by considering randomized algorithms as we do in this work (albeit with different techniques). Lower bounds have also been established in the stochastic convex optimization [2, 19] where each gradient- or function-value oracle query is corrupted with i.i.d. noise, and Allen-Zhu et al. [3] prove analogues of these bounds for streaming PCA. While these lower bounds are information-theoretic, and thus unconditional, they are inapplicable to the setting considered in this work, where we are allowed to make exact, noiseless queries.
2 STATEMENT OF MAIN RESULTS

Let $\| \cdot \|$ denote the $\ell_2$ norm on $\mathbb{R}^d$, and let $S^{d-1} := \{ x \in \mathbb{R}^d : \|x\| = 1 \}$ denote the unit sphere and Stiefel($d,r$) denote the Stiefel manifold consisting of matrices $V \in \mathbb{R}^{d \times r}$ such that $V^T V = I$. Let $\mathbb{S}^{d \times d}$ denote the set of symmetric $d \times d$ matrices, and for $M \in \mathbb{S}^{d \times d}$ we let $\lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_d(M)$ denote its eigenvalues in decreasing order, $v_1(M), v_2(M), \ldots, v_d(M)$ denote the corresponding eigenvectors, let $\| M \|_{op}$ and $\| M \|_F$ denote the operator and Frobenious norms, and $abs(M) := (M^2)^{1/2}$. Finally, we define the eigengap of $M \in \mathbb{S}^{d \times d}$ as $\gamma_p(M) := \sigma_1(M) - \sigma_2(M)$, where $\sigma_1(M) = \lambda_1(M)$ is the $i$-th singular value of $M$. We will also use the notation $gap(M) := \gamma_p(M)$. We now introduce a definition of our query model:

**Definition 2.1** (Query Model). An randomized adaptive query algorithm $\text{Alg}$ with query complexity $T \in \mathbb{N}$ and accuracy $\epsilon \in (0, 1)$ is an algorithm which, for $r \in [T]$, queries an oracle with a vector $v^{(i)}$, and receives a noiseless response $w^{(i)} = Mv^{(i)}$. At the end $T$ rounds, the algorithm returns a matrix $V \in \text{Stiefel}(d,r)$. The queries $v^{(i)}$ and output $V$ are allowed to be randomized and adaptive, in that $v^{(i)}$ is a function of $(v^{(1)}, w^{(1)}), \ldots, (w^{(i-1)}, w^{(i-1)})$, as well as some random seed.

The goal of $\text{Alg}$ is to return a $\hat{V}$ satisfying

$$\langle \hat{V}, abs(M)\hat{V} \rangle \geq (1 - \epsilon \text{gap}(M)) \sum_{\ell = 1}^r \sigma_\ell(M)$$

for some small $\epsilon \in (0, 1)$. In the rank-one case, $\hat{V}$ is a vector $\vec{V} \in S^{d-1}$, and the above condition reduces to $\langle \hat{V}, abs(M)\hat{V} \rangle \geq (1 - \epsilon \text{gap}(M))\|M\|_{op}$.

**Example 2.1** (Examples of Randomized Query Algorithms). In the rank-one case, the Power Method and Lanczos algorithms [15] are both randomized, adaptive query methods. Even though the iterates $v^{(i)}$ of the Lanczos and power methods converge to the top eigenvector at different rates, they make identical queries: namely, they both identify $M$ on the Krylov space spanned by $v^{(i)}, Mv^{(i)}, \ldots, M^{i-1}v^{(i)}$. Lanczos differs from the Power Method by choosing $\hat{V}$ to be the optimal vector in this Krylov space, rather than the last iterate. Observe that even in the rank-$r$ case, our query model still permits each single vector-query to be chosen adaptively. Hence, our lower bound applies to subspace iterations (e.g., the block Krylov method of Musco and Musco [23]), and to algorithms which use deflation [4].

To state our results, we construct for every gap $\epsilon \in (0, 1)$ a distribution over matrices $M$ under which a gap($M$) $\gtrsim \epsilon$. To do so, we introduce the classical GOE or Wigner law [5]:

**Definition 2.2** (Gaussian Orthogonal Ensemble (GOE)). We say that $W \sim \text{GOE}(d)$ if the entries $[W]_{ij} \sim \mathcal{N}(0, 1/d)$, for $i \geq j$, $W_{ij} \sim \mathcal{N}(0, 2/d)$, and for $1 \leq j < i \leq d$, $W_{ji} = W_{ij}$.

In the rank-one case, we will then take our matrix to be $M := W + \lambda u u^T$, where $W \sim \text{GOE}(d)$, $u \sim S^{d-1}$, and $\lambda > 1$ is a parameter to be chosen. A critical result gives a finite-sample analogue of a classical result in random-matrix theory, which states that $\lambda_{\max}(M) = \lambda + \lambda^{-1}$ with high probability. On the other hand, $\| W\|_{op}$ concentrates around 2, and thus by eigenvalue interlacing $\lambda_{\max}(M) - \lambda_2(M) \gtrsim \lambda + \lambda^{-1} - 2$. Motivated by this, we define the asymptotic gap of $M$:

$$\text{gap} = \gamma_p(M) := \frac{\lambda + \lambda^{-1} - 2}{\lambda + \lambda^{-1}} = \left(1 - \frac{1}{\lambda + \lambda^{-1}}\right).$$

It is well known that, for a fixed $\lambda$, $\gamma_p(M) \xrightarrow{\text{prob}} \text{gap}$ as $d \to \infty$. We give a finite sample analogue:

**Proposition 2.1** (Finite Sample Eigengap of Deformed Wigner). Let $M = W + \lambda u u^T$, where $W \sim \text{GOE}(d)$ and $U \sim \text{Stiefel}(d, r)$ are independent. For $\gamma \in (0, 1)$, define the event

$$E_{\text{good}}(\gamma) := \{ \| W\|_{op} + (1 - \gamma)(\lambda + \lambda^{-1} - 2) \lambda_2(M) \}
\cap \{ \lambda_1(M) \leq (1 + \gamma)(\lambda + \lambda^{-1})\}.$$

There exists a polynomially bounded function $\text{gap}^{-1}(1 - \text{gap}^{-1}, 1 - \text{gap}^{-1})$ such that for $\gamma \in (0, 1/10)$, and $d \geq \text{gap}^{-1}(1 - \text{gap}^{-1}, 1 - \text{gap}^{-1})$, $E_{\text{good}}(\gamma) \geq 1 - \delta$. Moreover, on $E_{\text{good}}(\gamma)$, $\gamma_p(M) = \gamma_p(abs(M)) \geq 1 - \frac{1}{\text{gap}^2} \cdot \text{gap}$.

The explicit polynomial can be derived from a more precise statement, Theorem 6.1. We now state more precise version of Theorem 1:

**Theorem 2.2.** Fix a gap $\epsilon \in (0, 1)$ and any $d \geq \text{gap}^{-1}(1 - \text{gap}^{-1}, 1 - \text{gap}^{-1}, 1 - \text{gap}^{-1})$ such that $q$ is as in Proposition 2.1, and let $\lambda = \frac{\text{gap}^{-1}(1 - \text{gap}^{-1})}{1 - \text{gap}^{-1}}$ be the solution to Equation 1. Let $M = W + \lambda u u^T$ where $U \sim \text{Stiefel}(d, r)$ and $W \sim \text{GOE}(d)$. Then for any $\text{Alg}$ satisfying Definition 2.1, we have

$$\mathbb{E}_M \left[ \text{P}_{\text{Alg}} \left[ \langle \hat{V}, abs(M)\hat{V} \rangle \geq \left(1 - \frac{\text{gap}}{12}\right) \sum_{\ell = 1}^r \sigma_\ell(M) \right] \right] \geq \mathbb{E}_{E_{\text{good}}(1/2)} \left[ E_{\text{good}}(1/2) \right] \geq 2e \cdot \exp \left(-\frac{d}{78 \log(d) \text{gap}^3} \left(1 + \frac{\sqrt{\text{gap}(2 - \text{gap})}}{1 - \text{gap}}\right)^{-18(\epsilon + 2)}\right).$$

Where $\text{P}_{\text{Alg}}$ is the probability taken with respect to the randomness of the algorithm. Note that on $E_{\text{good}}(1/2)$, $\gamma_p(M) \geq \text{gap}/3$.

Observe $\frac{1 + \sqrt{\text{gap}(2 - \text{gap})}}{1 - \text{gap}} \leq 1 + O \left( \sqrt{\text{gap}} \right)$ as gap is bounded away from 1. Hence, if gap $\leq 1/2$, $d$ is a large enough polynomial in gap, then if $\left( 1 + O \left( \sqrt{\text{gap}} \right) \right)^{d/(1)} \leq d^{1/2}$, or equivalently, $T \leq 1 + \frac{\sqrt{\text{gap}}}{\text{gap}(2 - \text{gap})} \log d$, we see that the probability that $\langle \hat{V}, abs(M)\hat{V} \rangle \geq \left(1 - \frac{\text{gap}}{12}\right) \sum_{\ell = 1}^r \sigma_\ell(M)$ is at most $e^{-\gamma(l)}$, proving Theorem 1.

In Appendix A of the full paper [31] we present two additional results that follow as easy modifications of our proofs: Theorem A.1 presents an improved gap-dependence for $r = 1$, and generalizes to the setting where Alg is allowed T rounds of adaptivity, and makes a batch of B queries per round; Theorem A.3 presents a modification of Theorem 2.2 which establishes a sharp lower bound of $\Omega \left( \frac{\text{gap}^d - \gamma(l)}{\log(1 - \text{gap})} \right)$ in the “easy” regime where gap approaches one. Our techniques can be adapted to show sharp lower bounds for adaptively testing between $H_0 : M = W + \lambda u u^T$ against $H_0 : M = W$; we omit these arguments in the interest of brevity.
3 PROOF ROADMAP

3.1 Notation
In what follows, we shall use bold letters \( \mathbf{u}, \mathbf{U}, \mathbf{M} \) and \( \mathbf{W} \) to denote the random vectors and matrices which arise from the deformed Wigner law; blackboard font \( \mathbb{F} \) and \( \mathbb{E} \) will be used to denote laws governing these quantities. We will use standard typesetting (e.g. \( u, M \)) to denote fixed (non-random) quantities vectors, as well as problem dimension \( d \) and rank \( r \) of the plant \( U \).

Quantities relating to \( \text{Alg} \) will be in serif font; these include the queries \( v^{(i)} \), responses \( w^{(i)} \), and outputs \( \hat{v} \) and \( \hat{V} \). The law of these quantities under \( \text{Alg} \) will be denoted \( \mathbb{P} \) in bold serif.

Mathematical operators like \( \text{gap}(\mathbb{M}) = \lambda_1(\mathbb{M}) \) are denoted in Roman or standard font, and asymptotic quantities like gap in Courier.

3.2 Reduction from Eigenvector Computation to Estimating \( U \)
In this section, we show that an algorithm which adaptively finds a near-optimal \( \hat{V} \) implies the existence of a deterministic algorithm which plays a sequence of orthonormal queries \( v^{(1)}, \ldots, v^{(r)} \) for which \( \sum_{i=1}^r \|u^{(i)}\|^2 \) is large. Our first step is to show that if \( \hat{V} \) is near-optimal, then \( \hat{V} \) has a large overlap with \( U \), in the following sense:

**Lemma 3.1.** Given any \( \hat{V} \in \text{Stief}(d, r) \), any \( r' \in [r] \), and under the event \( \mathcal{E}_{\text{good}}(\lambda, 1/2) \), if \( \langle \hat{V}, \text{abs}(\mathbb{M}) \hat{V} \rangle \geq \left( 1 - \frac{(r+1-r')^2}{2r} \right) \), \( \sum_{\ell=1}^r \sigma_{\ell}(\mathbb{M}) \), then \( \lambda_{r'}(\hat{V}^\top U U^\top U \hat{V}) \gtrsim \frac{\mathbb{E}(\mathbb{P})}{4} \).

In the rank one case, with \( r = r' = 1 \), \( \hat{V} = \hat{V} \) and \( U = u \), the above lemma just implies that a near optimal \( \hat{V} \) satisfies \( \langle \hat{V}, u \rangle^2 \gtrsim \text{gap} \). In the more general case, we have that \( \lambda_{r'}(\hat{V}^\top U U^\top U \hat{V}) \gtrsim \text{gap} \) means that the image of \( \hat{V} \) needs to have "uniformly good" coverage of the planted matrix \( U^\top U \). The proof of Lemma 3.1 begins with the Lowner-order inequality

\[
\hat{V}^\top \text{abs}(\mathbb{M}) \hat{V} = \hat{V}^\top W \hat{V} + \hat{V} \leq \|W\|_r + \lambda \hat{V}^\top U U^\top U \hat{V}
\]

In the rank one case, this reduces to

\[
\hat{V}^\top \text{abs}(\mathbb{M}) \hat{V} \leq \|W\| + \lambda \langle \hat{V}, u \rangle^2.
\]

Hence, if we want \( \hat{V}^\top \text{abs}(\mathbb{M}) \hat{V} \gtrsim \lambda_{\text{max}}(\mathbb{M}) - \text{gap}/2, \) we must have that, since \( \lambda_{\text{max}}(\mathbb{M}) = \|W\| \) concentrates around gap, and

\[
\langle \hat{V}, u \rangle^2 \geq \frac{1}{\lambda} (\hat{V}^\top \text{abs}(\mathbb{M}) - \|W\| - \text{gap}/2) = \text{gap}/2 \lambda.
\]

which gives the lower bound. For \( r > 1 \), the proof becomes more technical, and is deferred to Appendix B.1.

Next, we argue that the performance of the optimal \( \hat{V} \) is bounded by a quantity depending only on the query vectors. As a first simplification, we argue that we may assume without loss of generality that \( v^{(1)}, v^{(2)}, \ldots \) are orthonormal.

**Observation 3.1.** We may assume that the queries are orthonormal, so that:

\[
V_k := [v^{(1)} \mid v^{(2)} | \cdots | v^{(k)}] \in \text{Stief}(d, k),
\]

and that, rather than returning responses \( w^{(k)} = \text{Mv}^{(k)} \), the oracle returns responses \( w^{(k)} = (I - V_k V_k^\top) \text{Mv}^{(k)} \), where we note that \( V_{k-1} V_k^\top \) is the projection onto \( \text{span}(v^{(1)}, v^{(2)}, \ldots, v^{(k-1)}) \).

The assumption that the queries are orthonormal are valid since we can always reconstruct \( k \)-queries \( v^{(1)}, \ldots, v^{(k)} \) from an associated orthonormal sequence obtained via the Gram-Schmidt procedure. The reason we can assume the responses are of the form \( w^{(k)} = (I - V_k V_k^\top) \text{Mv}^{(k)} \) is that Alg queries \( v^{(1)}, \ldots, v^{(k-1)} \), it knows \( Mv_{k-1} V_{k-1}^\top \), and thus, since \( M \) and \( V_{k-1} V_{k-1}^\top \) are symmetric, it also knows \( V_{k-1} V_{k-1}^\top M \), and thus \( Mv^{(k)} \) can be reconstructed from \( w^{(k)} = (I - V_{k-1} V_{k-1}^\top) \text{Mv}^{(k)} \). The next observation shows that it suffices to upper bound \( \lambda_r(\hat{V}^\top U U^\top U \hat{V}) \) with \( \lambda_r(\hat{V}^\top U U^\top U V^\top r) \).

**Observation 3.2.** We may assume without loss of generality that Alg makes \( r \) queries \( v^{(T_1)}, \ldots, v^{(r)} \) after outputting \( \hat{V} \), and that \( \lambda_r(\hat{V}^\top U U^\top U \hat{V}) \leq \lambda_r(\hat{V}^\top U U^\top U V^\top r) \).

This is valid because we can always modify the algorithm so that the queries \( v^{(T_1)}, \ldots, v^{(r)} \) ensures that \( \text{range}(\hat{V}) \subseteq \text{span}(v^{(1)}, v^{(2)}, \ldots, v^{(T_1)}, \ldots, v^{(r)}) = \text{range}(V^\top r) \).

In this case, we have that for all \( \ell \in [r] \) (in particular, \( \ell = r' \)),

\[
\hat{V}^\top \hat{V} \leq V^\top T_r V^\top r \quad \Rightarrow \quad U^\top U \hat{V} \leq U^\top V^\top T_r V^\top r U \quad \Rightarrow \quad \lambda_{\ell}(U^\top U \hat{V}) \leq \lambda_{\ell}(U^\top V^\top T_r V^\top r U) \quad \Rightarrow \quad \lambda_{\ell}(U^\top U \hat{V}) \leq \lambda_{\ell}(U^\top V^\top T_r V^\top r U).
\]

Lastly, suppose it is the case that \( \mathbb{E}_{U, M, \text{Alg}}[\lambda_r(V^\top T_r U^\top U V^\top T_r U)] \geq b \leq b \) for any deterministic algorithm \( \text{Alg}_{\text{det}} \), and some bounds \( B > 0 \) and \( b \in (0, 1) \). Then for any randomized algorithm \( \text{Alg} \), Fubini’s theorem implies

\[
\mathbb{P}_{U, M, \text{Alg}}[\lambda_r(V^\top T_r U^\top U V^\top T_r U) \geq b] \leq \mathbb{E}_{\text{Alg}} \mathbb{P}_{U, M}[\lambda_r(V^\top T_r U^\top U V^\top T_r U) \geq b] \leq \text{seed of Alg} \leq b.
\]

as well. Hence,

**Observation 3.3.** We may assume that, for all \( k \in [T + r] \), the query \( v^{(k)} \) is deterministic given the previous query-observation pairs \( (v^{(i)}, w^{(i)})_{1 \leq i \leq k} \).

3.3 Lower Bounding the Estimation Problem
As discussed above, we need to present lower bounds for the problem of sequentially selecting measurements \( v^{(1)}, v^{(2)}, \ldots, v^{(r)} \) for which the associated measurement matrix \( V^\top r \) has a large overlap with the planted matrix \( U \). Proving a lower bound for this sequential, statistical problem constitutes the main technical effort of this paper. We encode the entire history of \( \text{Alg} \) up to time \( i \) as \( Z_i := (v^{(1)}, w^{(1)}, \ldots, v^{(i)}, w^{(i)})_{1 \leq i \leq T+i} \); in particular, \( Z_{T+r} \) describes the entire history of the algorithm.

Next, for \( U \in \text{Stief}(d, r) \), we let \( P_U \) denote the law of \( Z_{T+r} \), where \( M = W + \lambda U U^\top \) conditioned on \( U = U \). In the rank-one case, we \( P_U \) denotes the law of \( Z_{T+r} \), where \( M = W + \lambda u u^\top \) conditioned on \( u = u \). We will also abuse notation slightly by letting \( P_u \) denote the law obtained by running \( \text{Alg} \) on \( M = W \), i.e. with \( U = u = 0 \). In the rank-one case, we have the following theorem, whose proof is outlined in Section 4:
Theorem 3.2. Let $M = W + \lambda UU^T$, where $W \sim \text{GOE}(d)$, and $u \sim S^{d-1}$. Then for all $\delta \in (0, 1/e)$,
\[
\mathbb{E}_u P_u \left[ 3k \geq 1 : u^T V_k V_k^T u \geq 32\lambda^4k \text{ gap}^{-1/2} \right] \leq \delta.
\]

The above theorem essential states that the quantity $u^T V_k V_k^T u$ can grow at most geometrically at a rate of $\lambda^4k$, with an initial value sufficiently large in terms of the probability $\delta$ and gap. In Section 5, we prove an analogous bound, which gives a geometric control on $\lambda_r(U^T V_k V_k U)$:

Theorem 3.3. Let $M = W + \lambda UU^T$, where $U \sim \text{Stief}(d, r)$. Then for $d \geq \text{gap}^{-1/2}$ and $\delta \in (0, 1)$, and $r' \in [r]$
\[
\mathbb{E}_u P_u \left[ \forall k \in [d] : \lambda_r(U^T V_k V_k U) \leq \frac{26\lambda^2k/r' \log(2d^2) \log(e^\delta - 1)}{\text{gap}^2} \right] \geq 1 - \delta.
\]

In Section A.3, we combine Theorem 3.3, Lemma 3.1, and Observation 3.1 to prove Theorem 2.2. The final rate is a consequence of the fact that $\lambda = \frac{1+\sqrt{8d^2-2\text{gap}}}{d}$. As mentioned in the paragraph New Techniques, our main technical hammer for proving Theorems 3.2 and 3.3 is a novel data-processing lower bound (Proposition 4.4) which applies to “truncated” distributions; the techniques are explained in greater detail in Appendix F.

3.4 Conditional Likelihoods from Orthogonal Queries

We conclude with one further simplification which yields a closed form for the conditional distributions of our queries. Observe that it suffices to observe the queries $w(i) = (I - V_{i-1}V_{i-1}^T) M v(i) = M v(i) - V_{i-1}(M V_{i-1})^T v(i)$, our algorithm already “knows” the matrix $M V_{i-1}$ from the previous queries. Hence,

Observation 3.4. We may assume that we observe queries $w(i) = P_i M v(i)$, where $P_i := I - V_{i-1}V_{i-1}^T$.

We now show that, with our modified measurements $w(i) = P_i M v(i)$, then the query-observation pairs $(v(i), w(i))$ in the rank-one case have Gaussian likelihoods conditional on $Z_i$ and $u$.

Lemma 3.4 (Conditional Likelihoods). Let $P_i := I - V_iV_i^T$ denote the orthogonal projection onto the orthogonal complement of span($v(1), \ldots, v(i)$). Under $P_u$ (the joint law of $M$ and $Z_T$ on $\{u = u\}$), we have
\[
(P_{i-1} M v(i)|Z_{i-1}) \sim \mathcal{N} \left( \lambda(u^Tv(i)) P_{i-1} u, \frac{1}{d} \Sigma_i \right)
\]
where $\Sigma_i := P_{i-1} \left( I_d + \lambda(u^Tv(i)) P_{i-1} \right) P_{i-1}$.

In particular, $w(i)$ is conditionally independent of $w(1), \ldots, w(i-1)$ given $v(1), \ldots, v(i-1)$ and $u = u$.

Lemma 3.4 is proved in Appendix C.2. We remark that $\Sigma_i$ is rank-deficient, with its kernel being equal to the span of $(v(1), \ldots, v(i-1))$. Nevertheless, because the mean vector $\lambda(u^Tv(i)) P_{i-1} u$ lies in the orthogonal complement of $\ker \Sigma_i$, computing $\Sigma_i^{-1}(\lambda(u^Tv(i)) P_{i-1} u)$ can be understood as $\Sigma_i^{-1}(\lambda(u^Tv(i)) P_{i-1} u)\Sigma_i^{\dagger}$, where $\dagger$ denotes the Moore-Penrose pseudo-inverse [18].

4 PROOF OF THEOREM 3.2 ($r = 1$)

In this section, we prove a lower bound for the rank-one planted perturbation. The arguments in this section will also serve as the bedrock for the rank $r$ case, and exemplify our proof strategy.

Given any $V_k \in \text{Stief}(d, k)$, we introduce the notation $\Phi(V_k; u) := (V_k, uu^T V_k)$, which is just the square Euclidean norm of the projection of $u$ onto the span of $v(1), \ldots, v(k)$. $\Phi(V_k; u)$ will serve as a “potential function” which captures how much information the queries $v(1), \ldots, v(k)$ have collected about the planted solution $u$, in a sense made precise in Proposition 4.4 below. The core of our argument is the following proposition, whose proof is given in the following subsection:

Proposition 4.1. Let $(\tau_k)$ be a sequence such that $\tau_0 = 0$, and for $k \geq 1$, $\tau_k \geq 2k$. Then for all $\eta > 0$, one has
\[
\mathbb{E}_u P_u [\Phi(V_k; u) \leq \frac{\tau_k}{d} \cap \{ \Phi(V_{k+1}; u) > \frac{\tau_{k+1}}{d} \}] \leq \exp \left[ \frac{\eta}{2} \left( (1 + \eta) \frac{\tau_k}{d} - \sqrt{\frac{\tau_{k} + 2k}{2}} \right)^2 \right].
\]

The above proposition states that, given two thresholds $\tau_k, \tau_{k+1} > 0$, the probability that $d(V_{k+1}; u)$ exceeds the threshold $\tau_{k+1}$ on the event that $d(V_k; u)$ does not exceed the threshold $\tau_k$ is small. Hence, for a sequence of thresholds $0 = \tau_0 < \tau_1 < \ldots$, we have
\[
\mathbb{E}_u P_u [\exists k \geq 0 : \Phi(V_k; u) > \frac{\tau_{k+1}}{d} \cap \{ \Phi(V_{k+1}; u) > \frac{\tau_{k+1}}{d} \}] \leq \sum_{k=0}^{\infty} \mathbb{E}_u P_u [\Phi(V_k; u) \leq \frac{\tau_k}{d} \cap \{ \Phi(V_{k+1}; u) > \frac{\tau_{k+1}}{d} \}].
\]

Theorem 3.2 now follows by choosing the appropriate sequence $\tau_k(\delta)$, selecting $\eta$ appropriately, and verifying that the right hand side of the above display is at most $2\delta$. For intuition, setting $\eta = \lambda - 1$, we see that once $\tau_k$ gets large, it is enough to choose $\tau_{k+1} = \lambda^2 \tau_k$ to ensure that the exponent in Equation (3) is a negative number of sufficiently large magnitude. The details are worked out in Appendix D. We now turn to the proof of Proposition 4.1.

4.1 Proving Proposition 4.1

To prove Proposition 4.1, we argue that if $\tau_k$ is much smaller than $\tau_{k+1}$, then under the event $[\Phi(V_k; u) \leq \tau_k/d]$, the algorithm does not have enough information about $u$ to select a new query vector $v(k+1)$ for which $[\Phi(V_{k+1}; u) > \tau_{k+1}/d]$. The following proposition is proved in Section 4.3, and arises as a special case of a more general information theoretic tools introduced in that section.

Proposition 4.2. Let $D$ be any distribution supported on $S^{d-1}$, and let $\eta > 0$. Then,
We now state 4.4 which gives an upper bound on the information term as follows (see Appendix C.1 for proof):

**Lemma 4.3.** For any fixed $V \in \text{Stief}(d, k + 1)$ and $\tau_{k+1} \geq 2(k + 1)$, we have

\[
P_u\left[u^T V^+ V u \geq \tau_{k+1}/d\right] \leq \exp\left(-\frac{1}{2} \left(\sqrt{\tau_{k+1}} - \sqrt{2(k + 1)}\right)^2\right).
\]

We now state 4.4 which gives an upper bound on the information term. The proof is considerably more involved that of Lemma 4.3, and so we present a sketch in Section 4.2 below.

**Proposition 4.4.** For any $\tau_k \geq 0$ and any fixed $u \in S^{d-1}$, we have

\[
\mathbb{E}_u \mathbb{E}_{\theta_k} \left[\left(\frac{d\mu_{\theta_k}(Z_k)}{dP_0(Z_k)}\right)^{1+\eta} \left| (\Phi(V_k; u) \leq \tau_k)\right]\right] \leq e^{-\frac{(1+\eta)^2 \tau_k}{2}}.
\]

In particular, by taking an expectation over $u \sim S^{d-1}$, we have that

\[
\mathbb{E}_u \left[\left(\frac{d\mu_{\theta_k}(Z_k)}{dP_0(Z_k)}\right)^{1+\eta} \left| (\Phi(V_k; u) \leq \tau_k)\right]\right] \leq \exp\left(-\frac{(1+\eta)^2 \tau_k}{2}\right).
\]

This motivates the choice of $\Phi(V_k; u)$ as an information-potential, since it gives us direct control over bounds of the likelihood ratios. Proposition 4.1 now follows immediately from stringing together Proposition 4.2, Proposition 4.4 for the "information term", and Lemma 4.3 for the "entropy term".

### 4.2 Proof of Proposition 4.4 ("Information Term")

The difficulty in Proposition 4.4 is that truncating to the event $\{\Phi(V_k; u) \leq \tau_k\}$ introduces correlations between the conditional likelihoods that don’t arise in the conditionally independent likelihoods of Lemma 3.4. Nevertheless, we use a careful peeling argument (Appendix C.3) to upper bound the information term, an expected product of likelihoods, by a product of expected conditional likelihoods which we can compute. Formally, we have

**Proposition 4.5** (Generic upper bound on likelihood ratios). Fix an $u, s \in S^{d-1}$, and fix $r_u, r_s, \tau_0 \geq 0$. Define the likelihood functions

\[g_i(\tilde{V}_i) := \mathbb{E}_\rho \left[\frac{d\mu_{\eta_i}(Z_i|Z_{i-1})}{dP_0(Z_i|Z_{i-1})}\right]_{V_i = \tilde{V}_i}, \tag{5}\]

Then for any subset $V_k \subset \text{Stief}(d, k)$, we have

\[\mathbb{E}_\rho \left[\frac{d\mu_{\eta_i}(Z_i|Z_{i-1})}{dP_0(Z_i|Z_{i-1})}\right] \leq \sup_{V_k \in \text{Stief}(k)} \prod_{i=1}^k g_i(\tilde{V}_i), \tag{6}\]

where $\tilde{V}_{k,i}$ denotes the first $i$ columns of $V_k$.

Here, we remark that the tilde notation $(\tilde{V}_k, \tilde{v}^{(1)}, \tilde{v}^{(2)}, \ldots)$ represents fixed vectors which the random quantities $V_k, \tilde{v}^{(1)}, \tilde{v}^{(2)}$, etc. For example, in the event $V_k = \tilde{V}_k$, $\tilde{V}_k$ is considered to be a deterministic matrix.

We can now invoke a computation of the $1 + \eta$-th moment of the likelihood ratios between two Gaussians, proved in Appendix C.4.

**Lemma 4.6.** Let $\mathbb{P}$ denote the distribution $\mathcal{N}(\mu_1, \Sigma)$ and $\mathbb{Q}$ denote $\mathcal{N}(\mu_2, \Sigma)$, where $\mu_1, \mu_2 \in (\ker \Sigma)^\perp$. Then

\[\mathbb{Q} \left[\mathbb{E}_{\Sigma} \left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)^{1+\eta} \right] = \exp\left(\frac{\eta(1+\eta)}{2} (\mu_1 - \mu_2)^T \Sigma^\dagger (\mu_1 - \mu_2)\right). \tag{7}\]

We are now in a position to prove Proposition 4.4:

**Proof of Proposition 4.4.** Fix a $u \in S^{d-1}$, and we shall and apply Proposition 4.5 in the language of Proposition 4.5, we have

\[g_i(\tilde{V}_i) = \mathbb{E}_{\rho} \left[\frac{d\mu_{\eta_i}(Z_i|Z_{i-1})}{dP_0(Z_i|Z_{i-1})}\right]_{V_i = \tilde{V}_i} = \mathbb{E}_{\rho} \left[\frac{d\mu_{\eta_i}(u_i|Z_{i-1})}{dP_0(Z_i|Z_{i-1})}\right]_{V_i = \tilde{V}_i} \tag{8}\]

Now, observe that, $P_{\mu}(u_i|Z_{i-1})$ is the density of $N(S_i(u_i, \Sigma_i) \cdot P_{i-1}(1) + \Sigma_i \cdot P_{i-1}(1)^T)$, and $P_{\rho}(u_i|Z_{i-1})$ is the density of $N(0, 1_{d} \Sigma_i)$. Since $\Sigma_i = P_{i-1}(1_d + 1_{d} \tilde{v}^{(i)} \tilde{v}^{(i)^T}) P_{i-1}(1)$, we have $P_{i-1}(1) \Sigma_i P_{i-1}(1) = P_{i-1}(1) \leq I$. Thus,

\[u^T \Sigma_i u \leq d \|u\|^2 \leq d \quad \forall u \in S^{d-1}. \tag{8}\]

Hence, by Lemma 4.6, we have

\[g_i(\tilde{V}_i) \leq \exp\left(\frac{\eta(1+\eta)^2}{2} u^T \Sigma_i u \right) \leq \exp\left(\frac{\eta(1+\eta)^2}{2} \right). \tag{9}\]

Hence, by Lemma 4.6, we have

\[g_i(\tilde{V}_i) \leq \exp\left(\frac{\eta(1+\eta)^2}{2} \right). \tag{9}\]
Hence, if \( \mathcal{V}_k := \{ \widetilde{v}_k \in \text{Stief}(d, k) : \Phi(\widetilde{v}_k; u) \leq \tau\} \), then Proposition 4.5 implies
\[
\mathbb{E}_u \left[ \left( \frac{d \mathbb{P}_u(\mathcal{Z}_k)}{d \mathbb{P}_0(\mathcal{Z}_k)} \right)^{1+\eta} I(\mathcal{V}_k \in \mathcal{V}_k) \right]
\leq \sup_{\mathcal{V}_k \in \mathcal{V}_k} \prod_{i=1}^{k} \mathbb{E}_u \left[ \frac{d(1+\eta)^{2} \cdot d(u, \mathcal{V}_k[i])^2}{2} \right]
= \sup_{\mathcal{V}_k \in \mathcal{V}_k} \mathbb{E}_u \left[ \frac{d(1+\eta)^{2} \Phi(\mathcal{V}_k; u)}{2} \right]
\leq \frac{\exp((1+\eta)^{2} \lambda_{2} \tau)}{2}.
\]

4.3 Proof of Proposition 4.2

We begin by introducing the general framework for Bayes risk lower bounds as presented in Chen et al. [13]. We begin with an estimand parameter \( \theta \) drawn from some prior \( \mathcal{P} \) over a measurable space \( (\Theta, \mathcal{G}) \). To each fixed \( \theta \in \Theta \) is associated a measure \( \mu_\theta \) over a measurable space \( (X, \mathcal{F}) \), governing a random variable \( X \). In our setting, consider the rank-one deformed Wigner \( M = W + \mu_\theta u \), a fixed Alg and round \( k \). Then the estimand is \( \theta = u \), the measures \( \mu_\theta \mid_{\mathcal{G}_0} \) is the measure correspond to the laws \( \mathbb{P}_u(.) \) over \( x = Z_k \).

We would like to use \( x \) to make an action which tells us something useful about \( \theta \). Formally, consider a space of action \( \mathcal{A} \) and a space \( \mathcal{H} \) of measurable action mappings \( \alpha : X \rightarrow \mathcal{A} \), and an indicator function \( I(\cdot, \cdot) : \alpha \times \Theta \rightarrow [0, 1] \) of a “good event” that we would want an algorithm to achieve. In our Wigner model, \( \mathcal{A} \) will denote the space \( \text{Stief}(d, k) \), each \( a \mid \mathcal{H} \) denote a mapping from the playout history \( Z_k \) to the measurements \( V_k \), and the good event will be
\[
I(\Phi(\mathcal{V}_k; \theta)) = I(\theta^T V_k V_k^T \theta > \tau)
\]
for some threshold \( \tau \).

As we want to show lower bounds, our goal will be to show that the quantity
\[
V_{\text{opt}} := \sup_{\alpha \in \mathcal{A}} \mathbb{E}_{\theta \sim \mathcal{P}} \mu_\theta[I(\alpha(x), \theta) = 1] \tag{9}
\]
cannot be too large. The key difference between this setup and typical information-theoretic lower bounds is that we will not require the measures \( \mu_\theta \) to be normalized (i.e., probability measures), only that they have finite mass \( \mu_\theta(X) \ll \infty \). Our motivation for this is that we will take \( \mu_\theta \) to be truncated probability measures, or measures \( \mu_\theta \) with \( \mu_\theta(X) \leq 1 \) for which there exists a probability distribution \( \overline{\mu}_\theta \) and an event \( B_\theta \subseteq \mathcal{F} \) such that
\[
\forall A \subseteq \mathcal{F} : \mu_\theta(A) = \overline{\mu}_\theta(A \cap B_\theta). \tag{10}
\]
To make this concrete, suppose in our above example that \( \mathbb{P}_u \) as our unnormalized measures \( \mathbb{P}_\theta \), and the sets \( B_u := \{ \Phi(V_{k-1}; u) \leq \tau_{k-1} \} \). Then, \( \mu_\theta \) correspond to the subdistribution
\[
A \mapsto \mathbb{P}_\theta(A \cap \{ \Phi(V_{k-1}; \theta) \leq \tau_{k-1} \}).
\]
Hence, we have
\[
V_{\text{opt}} = \sup_{\alpha \in \mathcal{A}} \mathbb{E}_u \mathcal{P}_\theta \left( \{ \Phi(\alpha(Z_k); u) \leq \tau \} \cap \{ \Phi(V_{k-1}; u) \leq \tau_{k-1} \} \right) \geq \mathbb{E}_{u \sim \mathcal{P}_\theta} \left( \{ \Phi(V_{k}; u) \leq \tau \} \cap \{ \Phi(V_{k-1}; u) \leq \tau_{k-1} \} \right). \tag{11}
\]

which is precisely the quantity we wish to control in Proposition 4.2. More generally, considering such truncated measures is desirable in adaptive settings when we may want to consider the probability than a sequential algorithm takes a certain action at stage \( k \), on the event that it has taken certain actions prior to stage \( k \). Our main theorem is as follows:

**Theorem 4.7** (Bayes risk lower bound for sub-distributions). Let \( \mathcal{P} \) be a prior distribution over \((\Theta, \mathcal{G})\), let \( v \) and \( (\mu_\theta) \) be a family of fine measures over \((X, \mathcal{F})\). Let \( \mathcal{A} \) denote an action space, let \( \mathcal{X} \) denote the space of decision rules from \( X \to \mathcal{A} \), and let \( I : \mathcal{A} \times \Theta \to [0, 1] \) be an indicator function. Let
\[
V_0 := \sup_{\alpha \in \mathcal{A}} \mathbb{E}_{\theta \sim \mathcal{P}}[I(\alpha, \theta) = 1] \tag{12}
\]
denote the optimal value of the best action taken without observing \( x \). If \( f \) is non-negative, convex, \( v(\mathcal{X}) \leq 1 \), \( \sup_{\theta \sim \mathcal{P}} \mu_\theta(X) \leq 1 \), and either \( i \) \( \rightarrow \) \( x f(1/x) \) is non-increasing or \( ii \) \( v(\mathcal{X}) \leq 1 \), then \( 2 \)
\[
\inf_{v \sim \mathcal{X}} \mathbb{E}_v[f(\frac{dv}{d\nu})] \geq V_0 f \left( \frac{V_{\text{opt}}}{V_0} \right). \tag{13}
\]

In essence, the above theorem relates two quantities on the right, a quantity comparing the optimal value \( V_{\text{opt}} \) to be the best “data-oblivious” value \( V_0 \), which depends only on the “spreadness” of the prior \( \mathcal{P} \) and not on the condition laws \( \mu_\theta \). The quantity \( \mathbb{E}_v[f(\frac{dv}{d\nu})] \) on the left hand side is known as a \( f \)-divergence [14] between \( \mu_\theta \) and \( v \), which measures the dissimilarity between the measures \( \mu_\theta \) and \( v \); we introduce them in full generality in Appendix F.1. If there exists a measure \( v \) for which \( \mathbb{E}_v[f(\frac{dv}{d\nu})] \) is small, it means that the measures \( \mu_\theta \) are in a sense similar on average, and hence the variable \( x \) doesn’t convey too much information about the estimand \( \theta \), and thus \( V_{\text{opt}} \) cannot be considerably larger than \( V_0 \).

Theorem 4.7 is proven in Appendix F.1, along with a more general bound, Theorem F.3. We now conclude this subsection with the proof Proposition 4.2:

**Proposition 4.2.** We apply Theorem 4.7 with \( f(x) = x^{1+\eta} \) (which is non-negative on \((0, \infty)\), convex, and \( x f(1/x) = x^{-\eta} \) non-increasing). For clarity, we will index our truncated laws \( \mu_\theta(A) \) by \( u \in S^{d-1} \). Now, we take \( \mu_\theta(A) := \mathbb{P}_u(A \cap \{ \Phi(V_{k-1}; u) \leq \tau_{k-1} \}) \). We also take \( v \) to be the law of the law of \( Z_k \) under \( \mathbb{P}_0 \), the law of Alg under \( M = W \), without the rank-one spike. Since \( \mathbb{P}_\theta \ll \mathbb{P}_0 \) we see that \( \mu_\theta \ll v \). Moreover, we have that
\[
\frac{d \mu_\theta}{d \nu} = \frac{d \mathbb{P}_u}{d \mathbb{P}_0} \Phi(\mathcal{Z}_k; \theta) \leq \tau_{k-1}.
\]
Lastly we take \( \mathcal{P} \) to be the uniform distribution on the sphere. Hence, the right hand side of Theorem 4.7 reads
\[
\mathbb{E}_{u \sim S^{d-1}} \mathbb{E}_{\mathcal{Z}_k} \left[ \left( \frac{d \mathbb{P}_u(\mathcal{Z}_k)}{d \mathbb{P}_0(\mathcal{Z}_k)} \right)^{1+\eta} I(\Phi(V_{k-1}; u) \leq \tau_{k-1}) \right].
\]

On the other hand, we now choose the action space \( \mathcal{A} = \text{Stief}(d, k) \) and the indicator \( I(\mathcal{V}_k; u) := I(\Phi(V_{k}; u) > \tau) \). Using (9), we

\[\text{The notation } (\mu_\theta) \ll v \text{ means that for each } \theta \in \Theta, \mu_\theta \text{ is absolutely continuous with respect to } v \text{ (see e.g., Kalužna [22] for a review of absolute continuity).}\]
have
\[ V_0 f(V_{opt}/V_0) = V_0^{\eta \log(1+\eta)} \]
\[ \text{Eq. (12)} \]
\[ \sup_{V_{opt} \in \text{Stief}(d,k)} \mathbb{P}_{-d,\Delta \leq \lambda \leq d} \left[ \Phi(V_k; u) > \tau \right]^{\eta/\tau} \]
Solving for \( V_{opt} \), we have
\[ \left( \sup_{V_{opt} \in \text{Stief}(d,k)} \mathbb{P}_{-d,\Delta \leq \lambda \leq d} \left[ \Phi(V_k; u) > \tau \right] \right)^{\tau/\eta} \]
\[ \times \left( \mathbb{E}_{u \sim Z_k - \Phi_k} \left( \frac{d\mathbb{P}(Z_k)}{d\mathbb{P}(Z_k)} \right)^{1+\eta} I(\Phi(V_k; u) \leq \tau_k - 1) \right)^{\eta/\tau} \]
\[ \geq V_{opt} \]
\[ \text{Eq. (11)} \]
\[ \mathbb{E}_{u \sim Z_k - \Phi_k} \left( \Phi(V_k; u) > \tau \right) \cap \left( \Phi(V_k; u) \leq \tau_k - 1 \right) \]
This concludes the proof. \( \square \)

5 PROOF OF THEOREM 3.3 \((r \geq 1)\)

Here we present a proof outline of Theorem 3.3 which modifies the insights from the rank one case to get a recursion for the determinant \( \det(UV_k V_k U + \Delta I_r) \). We will use the rank-one potential from the last section \( \Phi(\cdot; \cdot) \), but will instead be interested in
\[ \Phi(V_k; U e) := e^T U^T V_k V_k^T U e, \quad e \in \mathbb{R}^r, \]
which measures the amount of information gathered about \( U \) in the direction of \( e \). We will want to show that, with high probability, the following event holds for an appropriate choice of parameters:
\[ \mathcal{E}(\tilde{\lambda}, \Delta, k_{max}) := \{ \forall e \in \mathbb{R}^r, k \in [1, \ldots, k_{max}] \}
\]
\[ \cap \{ \mathbb{E} \Phi(V_k; U e) + \Delta \leq \tilde{\lambda} (\mathbb{E} \Phi(V_{k-1}; U e) + \Delta) \}. \]

In other words, \( \mathcal{E}(\tilde{\lambda}, \Delta, k_{max}) \) corresponds to the event that, up to a translation by \( \Delta \), the potentials \( \Phi(V_k; U e) \) grows at most geometrically by a factor of \( \tilde{\lambda} \) in every direction. We should think of \( \tilde{\lambda} \) as being of order \( \Theta(1) \), which may be quite close to 1. Hence, the translation \( \Delta \) gives us additional slack which will be necessary for high-probability bounds.

Lemma 5.1. On \( \mathcal{E}(\tilde{\lambda}, \Delta, k_{max}) \), it holds that
\[ \det(UV_k V_k U + \Delta I_r) \leq \tilde{\lambda}^k \det(I_r) = \tilde{\lambda}^k \Delta^r \]

The proof of the above claim follows by first decomposing
\[ dU V_k V_k U + \Delta = dU V_{k-1} V_{k-1} U + \Delta I_r = U^T V_k V_k^T U + \Delta I_r \]
and applying the Sherman-Morrison rank-one update formula to control the growth of the \( \det(UV_k V_k U + \Delta I_r) \) in each stage. In Section E.1, we prove Theorem 3.3 by translating Lemma 5.1 into a growth bound on \( \lambda_r(U^T V_k V_k U) \), and control the probability of the event \( \mathcal{E}(\tilde{\lambda}, \Delta, k_{max}) \) with the following proposition:

Proposition 5.2. Let \( \rho \geq \lambda^3 c_{d,r}, \) fix \( k_{max} \geq 1 \), and set \( \Delta \geq \rho (2k_{max} + 1)/(\rho - 1) \).
\[ \mathbb{P}[\mathcal{E}(\rho^2, \Delta, k_{max})] \geq 1 - (20d/(\rho - 1))^r + \exp \left( -\frac{\lambda^3(\lambda - 1)\Delta}{2} \right) \]

5.1 Proof of Proposition 5.2

We will proceed by arguing that an analogue of \( \mathcal{E}(\tilde{\lambda}, \Delta, k_{max}) \) holds for a fixed \( e \in \mathbb{R}^{r-1} \), and then extending to all of \( S^{r-1} \) via a covering argument. Our first step is to prove an analogue of Proposition 4.1 for the potential \( \Phi(V_k; U e) \). This ends up between very similar to the rank-one case, with the modification that we end up conditioning on the matrix \( U(1 - e e^T) \), and consequently pay a slight penalty (see the factor \( c_{d,r} \) below) for reducing the effective problem dimension from estimating a random vector in \( \mathbb{R}^d \) to one in \( \mathbb{R}^{d-r-1} \). Precisely, we have the following:

Proposition 5.3. Define the constant \( c_{d,r} := \frac{d}{r - r - 1} \). Fix an \( \eta > 0 \), \( k \geq 0 \), and let \( \tau_k, \eta k + 1 \geq 0 \), with \( \tau_0 = 0 \). Then for any fixed \( e \in \mathbb{R}^r \)
\[ \mathbb{E} U^T P_k \left( \Phi(V_k; U e) \leq \tau_k/d \right) \cap (\Phi(V_{k+1}; U e) > \tau_{k+1}/d) \leq \exp \left( \frac{\eta^2}{2} \frac{(\Delta^2 - \Delta^2)}{1 + \eta} \right) \]

Proposition 5.3 is proved in Appendix E.4. We can now prove that a point-wise analogue of \( \mathcal{E}(\tilde{\lambda}, \Delta, k_{max}) \) for each \( e \in \mathbb{S}^{r-1} \) holds for \( \tilde{\lambda} = \lambda^3 c_{d,r} \).

Lemma 5.4. Let \( \rho \geq \lambda^3 c_{d,r}, \) fix \( k_{max} \geq 1 \), and set \( \Delta \geq \rho (2k_{max} + 1)/(\rho - 1) \).
\[ \mathbb{P}[\exists k \in [k_{max}] : \Phi(V_k; U e) + \Delta \geq \rho (\Phi(V_{k-1}; U e) + \Delta)] \leq \frac{d^2}{\rho - 1} \exp \left( \frac{-\lambda^3(\lambda - 1)\Delta}{2} \right) \]

The lemma is established by first fixing a \( k \in [k_{max}] \), “binning” \( d\Phi(V_{k-1}; U e) \) into at most \( d^2/(\rho - 1) \) intervals \( [\tau_{r-1} i, \tau_{r-1} i + 1] \), applying Proposition 5.3 and then using union bound over all \( k_{max} \) steps. To conclude the proof of Proposition 5.2, we invoke a simple covering argument to extend to all \( e \in \mathbb{S}^{r-1} \); details are given in Section E.2.

6 SPECTRUM OF DEFORMED WIGNER MODEL

In this section, we establish that for \( \lambda > 1 \), the top \( r \) eigenvalues of \( M = W + \lambda U U^T \) concentrate around \( \lambda + \lambda^{-1} \), while the magnitude of the remaining eigenvalues lie below \( 2 + o(1) \). While results of this flavor are standard in the asymptotic regime in which \( \lambda \) and \( r \) are held as fixed constants as \( d \to \infty \), [9, 10] our lower bounds require that \( d \) can be taken to be polynomial in \( r, \lambda \), and gap^{-1}.

Theorem 6.1. There exists a universal constant \( C \geq 0 \) such that the following holds. Let \( M = W + \lambda U U^T \) in \( \mathbb{S}^d \), and let gap be as in (1). Let \( \kappa \leq 1/2, e \leq \text{gap} \cdot \min \{1/2, 1/\lambda + 1/\lambda^{-1} \} \), and \( \delta > 0 \). Then for
\[ d \geq C \left( \frac{r + \log(1/\delta)}{\text{gap} e^2} + (\text{gap})^{-3} \log(1/\text{gap}) \right), \]
the event \( \mathcal{E}_M \) defined below holds with probability at least \( 1 - 9\delta \):
\[ \mathcal{E}_M := \left\{ \|W\|_{\text{op}} \leq 2 + \kappa (\lambda + \lambda^{-1} - 2) \right\} \]
\[ \cap \left\{ [\lambda_r(M), \lambda_1(M)] \subset (\lambda + \lambda^{-1})[1 - e, 1 + e] \right\} \]
Moreover, on \( \mathcal{E}_M \), \( \lambda_r(M) - \|W\|_{\text{op}} \geq (1 - \text{gap}^{-1})/4 \) \( \lambda + \lambda^{-1} \) gap \( \geq (\lambda + \lambda^{-1}) \text{gap}/4 \).
The proof begins with the statement that the eigenvalues of $M$ are precisely the zeros of the function $z \mapsto \det(zI - M) = \det(zI - W + \lambda UU^\top)$. In particular, if $z > \lambda_{\max}(W)$, then $zI - W$ is invertible, and by standard determinant identities, we have

$$
\det(zI - M) = \det(zI - W + \lambda UU^\top) = \det(zI - W)\det((I - \lambda U^\top(zI - W)^{-1}U)).
$$

In other words, $z > \lambda_{\max}(W)$ is in $\spec(M)$ if and only if $\det(I - \lambda U^\top(zI - W)^{-1}U) = 0$. Given $\epsilon, \kappa$ as in Theorem 6.1, we show that for $z' \leq 2 + \kappa(\lambda + \lambda^{-1} - 2) = 2 + o_d(1)$, and for the values

$$
a_{\text{low}} = (\lambda + \lambda^{-1})(1 - \epsilon), \quad \text{and} \quad a_{\text{up}} = (\lambda + \lambda^{-1})(1 - \epsilon),
$$

it simultaneously holds with high probability that $\|W\|_\op \leq z'$ and $z \mapsto \det(I - \lambda U^\top(zI - W)^{-1}U)$ vanishes for $r$ distinct values of $z \in [a_{\text{low}}, a_{\text{up}}]$. This will imply that at least $r$ of the eigenvalues of $M$ lie in $[a_{\text{low}}, a_{\text{up}}]$. Note that, by eigenvalue interlacing, it also follows that the remaining eigenvalues of $M$ lie in $[\lambda_{\min}(W), \lambda_{\max}(W)] \subseteq [-\|W\|_\op, \|W\|_\op]$. We will proceed by showing that the eigenvalues of the matrix $I_r - \lambda U^\top(zI - W)^{-1}U$ are all negative when $z' < z < a_{\text{up}}$ and are all positive when $z > a_{\text{up}}$. This motivates the definition of the events $\mathcal{A}(z') = \{\|W\|_\op \leq z'\}$,

$$
\mathcal{E}_{\text{low}}(z) := \{I_r - \lambda U^\top(zI - W)^{-1}U \leq 0\}, \quad \text{and} \quad \mathcal{E}_{\text{up}}(z) := \{I_r - \lambda U^\top(zI - W)^{-1}U \geq 0\}.
$$

Then, using a continuity argument, we derive a useful condition for $\det(I_r - \lambda U^\top(zI_r - W)^{-1}U)$ to vanish at $r$ distinct points.

**Proposition 6.2.** There exists a zero-measure event $N$ such that, on $N^c \cap \mathcal{A}(z') \cap \mathcal{E}_{\text{up}}(a_{\text{up}}) \cap \mathcal{E}_{\text{low}}(a_{\text{low}})$, the function $z \mapsto \det(I - \lambda U^\top(zI - W)^{-1}U)$ vanishes at $r$ distinct points in $[a_{\text{low}}, a_{\text{up}}]$.

We prove Proposition 6.2 in Section I. We are now left with controlling the probabilities of $\mathcal{E}_{\text{low}}$, $\mathcal{E}_{\text{up}}(a_{\text{up}})$ and $\mathcal{E}_{\text{low}}(a_{\text{low}})$. To control $\mathcal{A}(z')$, we combine a non-asymptotic bound on the spectral norm of a Wigner matrix $W$ by Bandeira and van Handel [8] with a standard concentration inequality. Note that asymptotic results of the above statement can be found in references as [5]. Vershynin [35] gives bounds that are sharp up to constant factors.

**Proposition 6.3 (Bound on $\|W\|_\op$).** Let $d \geq 250$, and fix a $p \in (0, 1)$. Then,

$$
P[\|W\|_\op > z'] \leq p, \quad (17)
$$

where $z' = z'(p) := 2 + 21d^{-1/3}\log^2(3/d) + 2\sqrt{\log(1/p)/d}.

The above proposition is proved in Section J. We must now control the probabilities of $\mathcal{E}_{\text{low}}$ and $\mathcal{E}_{\text{up}}$. Since $U$ is uniform on Stiefel$(d, r)$ and independent of $W$, we expect by concentration that

$$
U^\top(zI_d - W)^{-1}U \approx \mathbb{E}_U[U^\top(zI_d - W)^{-1}U] = \frac{1}{d}\text{tr}(zI_d - W)^{-1}I_r = Sw(z)I_r,
$$

where $Sw(z) := \frac{1}{d}\text{tr}(zI_d - W)^{-1}$ the Stieltjes transform of the empirical spectral measure of the Wigner matrix $W$. As $d \to \infty$, it is well known [5] that for all $z > 2$,

$$
Sw(z) := \frac{1}{d}\text{tr}(zI_d - W)^{-1} \xrightarrow{\text{prob}} s(z) \quad (18)
$$

where $s(z) := \frac{z - \sqrt{z^2 - 4}}{2}$.

Therefore we see that $\|U^\top(zI_d - W)^{-1}U - \lambda s(z)I_r\|_\op = o_d(1)$. Finally, we see that the equation $\det(zI - W) = 0$ is solved by $z = \lambda + \lambda^{-1}$, and that $s(a_{\text{low}}) > \lambda^{-1}$ and $s(a_{\text{up}}) < \lambda^{-1}$. Thus, our goal will be to verify that, on $\mathcal{A}(z')$, the following holds for $z \in [a_{\text{low}}, a_{\text{up}}]$ with high probability:

$$
|Sw(z) - s(z)| + \|Sw(z)I_r - U^\top(zI_d - W)^{-1}U\|_\op \leq |s(z) - \lambda^{-1}|. \quad (19)
$$

Indeed, we see that for $z = a_{\text{low}} < \lambda + \lambda^{-1}$, the above equation implies $\mathcal{E}_{\text{low}}$ by the triangle inequality, and similarly for $z = a_{\text{up}} > \lambda + \lambda^{-1}$. To handle the error $|Sw(z)I_r - U^\top(zI_d - W)^{-1}U\|_\op$, we fix $W$ and reason about the above quadratic form in $U$ using the following Hanson-Wright style inequality proved in Section K.

**Proposition 6.4 (Stiefel Hanson-Wright).** Let $A, B \in \mathbb{R}^{d \times d}$ be any fixed symmetric matrix, and let $U \overset{\text{unif}}{\sim} \text{Stiefel}(d, r)$ be uniform on the Stiefel manifold. Then for all $t \leq d/4$,

$$
P\left[\left\|U^\top AU - \text{tr}(A)I_r\right\|_\op > \frac{8}{d^2} (t|AA|_F + t|A\|_\op)/(d - 2\sqrt{t}/d)\right] \leq 3e^{-t + 2r}. \quad (20)
$$

In particular, if we choose $A = (zI_d - W)^{-1}$ and condition on $\mathcal{A}(z')$, we can bound $\|A\|_\op \leq (z - z')^{-1}$, $\|A\|_F \leq \sqrt{d}(z - z')^{-1}$, and hence conclude that $U^\top(zI_d - W)^{-1}U = Sw(z)I_r + o_d(1)$. Next, the term $|Sw(z) - s(z)|$ is upper bounded by Theorem 6.5 (proved below), which establishes a finite sample version of Equation (19).

**Theorem 6.5 (Stieltjes transform).** Fix $p, \delta \in (0, 1)$ and let $z'$ given in Proposition 6.3. Fix an $a \in (2 + (z - z')^{-1}/2, d/2)$, and assume that $\tau := (d(a - z')^2)^{-1/2} \leq \min\left\{\frac{1}{16\sqrt{2}}, \frac{1}{32}, \frac{1}{a - z'}\right\}$, and $p/3 < \tau/8$. Then there exists an event $\mathcal{E}_S(a) := \{\mathbb{E}[\text{Zar}(a)^2] \leq 1 - \delta\}$ and on $\mathcal{E}_S(a) \cap \mathcal{A}(z')$, $|Sw(a) - s(a)| \leq c_5 z'^2 + d_0 a^{3/2}p^{1/6}$

where $c_5 := 4\sqrt{2} + 2\sqrt{\log(2)/8}$.

Finally, Lemma G.1 in the Appendix establishes a lower bound on $|s(a) - \lambda^{-1}|$ (note that this is deterministic). In Section G.1, we put the pieces together to show for our choice $\epsilon, \kappa$, and an appropriate $z'$, Theorem 6.5 and Proposition 6.4 imply that Equation (19) holds with high probability.
respect to the Lebesque measure, and so integrating the summand
d(a−λ_{max}(W)) in the neighborhood of a will cause the expectation
to diverge. Luckily, the probability that λ_{max}(W) is close to a is vanishly small in d, and so we will still be able to establish concentration by estimating S_W(z), where z = a + bi and b > 0 is very samill relative to a. This ensures that E[S_W(z)] will converge, and in fact we will be able to both compute the latter expectation and show that S_W(z) concentrates around it. Before establishing with these acts, we show that if b is sufficiently small and a is not close to z^∗, then S_W(z) ≈ S_W(a).

Lemma 6.6. On A(z^∗), |Re(S_W(a + ib)) − S_W(a)| ≤ \frac{b^2}{(a−z^∗)^2} for any a > z^∗.

The proof of the above lemma is deferred to Appendix H.6. Our first step to control S_W(z) is estimating its expectation:

Proposition 6.7. Define the deterministic quantity

Err(z) := E[S_W(z)]^2 + \frac{1}{d} E[(tr(zI − X)^2 − E[S_W(z)])^2].

Then, as long as a^2 − 4 > b^2 + 4|Re(Err)| and b > |3m(Err(z))|, one has

\begin{align*}
|\text{Re}(E[S_W(z)]) − s(a)| & \leq \sqrt{|b^2 + 4\text{Re}(Err)| + |2ab + 3\text{Re}(Err)|}.
\end{align*}

The proposition, proved in Appendix H.5, follows the standard arguments given in Section 2.4 of [5]. At a high level, we show that E[S_W(z)] satisfies a quadratic equation whose roots are approximately \frac{a + √a^2−4}{2}. We need to take care that we choose the correct root, which imposes the conditions a^2 − 4 > b^2 + 4|Re(Err)| and |b > |3m(Err(z))|. In particular, the requirement b > |3m(Err(z))| will force us to take special care to show that 3m(Err(z)) is dominated by b.

The remaining part of the proof requires us to establish two results: first, that Re(Err(z)) and 3m(Err(z)) are sufficiently small, and second, that S_W(z) concentrates around its expectation. Since W has Gaussian entries, one may be tempted to argue both result by using the Lipschitz property of the map W → S_W(z). Unfortunately, the Lipschitz constant of this map scales with 1/b, which will become quite large if we take b to be too small.

Instead, we define a modified matrix \tilde{W} such that, \tilde{W} = W on A(z^∗), and the composition of maps W → \tilde{W} → S_W(z) has a suitably large Lipschitz constant, even when b is vanishly small function in d (e.g. d^{−10}). Specifically, given the eigendecomposition W = OΛO^T, define the matrix \tilde{W} := O^T diag(min(Λ_{ii}, z^∗))O to be the matrix obtained by truncating the eigenvalues of W to lie in (−∞, z^∗]. Observe that under A(z^∗), one has \tilde{W} = W. Moreover, λ_1(\tilde{W}) ≤ z^∗ almost surely.

This latter observation is crucial in establishing the following lemma, which allows us to approximate the real and imaginary parts of Err(z) - a quantity defined in terms of the raw Wigner matrix W - by variance-like quantities involving the Stieltjes transform of the modified matrix \tilde{W}:

Lemma 6.8. Suppose that 0 < b < min(1, a − z^∗), and define ρ := E[|A(z^∗)|]. Then,

\begin{align*}
&\text{Re(Err(z))} − \text{Re}(E[S_W(z)] − E[S_W(z)])^2 \leq \frac{1}{d(a−z^∗)^2} + \frac{4p^2}{b^2}.
\end{align*}

Moreover, on A(z^∗),

\begin{align*}
&\text{Im(Err(z))} − \text{Im}(E[S_W(z)] − E[S_W(z)])^2 \leq \frac{2b}{d(a−z^∗)^2} + \frac{4p^2}{b^2}.
\end{align*}

This leaves us with the two tasks before we can finally optimize Proposition 6.7 to get a high probability bound on S_W(z). First, we need to control Err(z) by bounding the size of the variance-like terms \text{Re}(E[S_W(z)] − E[S_W(z)])^2 and \text{Im}(E[S_W(z)] − E[S_W(z)])^2. Secondly, we shall need to argue that S_W(z) concentrates around E[S_W(z)]. Both tasks amount to controlling the deviations of S_W(z), which we can achieve by leveraging the fact that S_W(z) is a Lipschitz function of the underlying standard gaussian matrix X (recall that X_{ij} i.i.d. N(0,1), and that W = \frac{1}{√d}(X + X^T).)

Lemma 6.9. Let z = a + bi, and define the map Ψ : X → S_W(z). Then if a − z^∗ > |b|,

\begin{align*}
\text{Lip}(\text{Re}(\Psi)) & \leq \frac{\sqrt{2b}}{d^2(a−z^∗)^2} \quad \text{and} \quad \text{Lip}(\text{Im}(\Psi)) \leq \frac{4\sqrt{2b}}{d^2(a−z^∗)^2}.
\end{align*}

To control the variance terms, we use the Gaussian Poincare inequality, which states that if f : R^d → R is an L Lipschitz function, and let x ∈ R^d is a standard gaussian vector, then Var[f(x)] ≤ L^2. This lets us control \text{Re}(Err(z)) and \text{Im}(Err(z))

Lemma 6.10. Suppose that b ≤ (a − z^∗)/2 and d ≥ (a − z^∗)^2. Then,

\begin{align*}
&\text{Im}(\text{Err(z)}) \leq \frac{8\sqrt{2b}}{d} \cdot \max \left\{ \frac{1}{(a−z^∗)^2}, \frac{1}{d(a−z^∗)^3} \right\} + \frac{4p^2}{b^2}.
\end{align*}

\text{Re}(\text{Err(z)}) \leq \frac{1}{4} \left( d(a−z^∗)^2 + \rho^2/b^2 \right).

Appendix H.1, finally put together all these pieces to prove Theorem 6.5. Proofs of all the supporting claims can be found in the following subsections of Appendix H.

REFERENCES

[1] Alekh Agarwal and Leon Bottou. 2014. A lower bound for the optimization of finite sums. arXiv preprint arXiv:1410.0723 (2014).
[2] Alekh Agarwal, Martin J Wainwright, Peter L Bartlett, and Pradeep K Ravikumar. 2009. Information-theoretic lower bounds on the oracle complexity of convex optimization. In Advances in Neural Information Processing Systems. 1–9.
[4] Zeyuan Allen-Zhu and Yuanzhi Li. 2016. First Efficient Convex Convergence for Streaming k-PCA: a Global, Gap-Free, and Near-Optimal Rate. arXiv preprint arXiv:1607.07837 (2016).
[5] Zeyuan Allen-Zhu and Yuanzhi Li. 2016. LazySVD: Even faster SVD decomposition yet without agonizing precision. In Advances in Neural Information Processing Systems. 974–982.
[6] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. 2010. An introduction to random matrices. Vol. 118. Cambridge university press.
[7] Anurag Anshu, Naresh B Goud, Rahul Jain, Srijita Kundu, and Priyanka Mukhopadhyay. 2017. Lifted randomized query complexity to randomized communication complexity. arXiv preprint arXiv:1703.07521 (2017).
[8] Ery Arias-Castro, Emmanuel J Candes, and Mark A Davenport. 2013. On the fundamental limits of adaptive sensing. IEEE Transactions on Information Theory 59, 1 (2013), 472–481.
Tight Query Complexity Lower Bounds for PCA

[8] Afonso S Bandeira, Ramon van Handel, et al. 2016. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability* 44, 4 (2016), 2479–2506.

[9] Florent Benaych-Georges and Raj Rao Nadakuditi. 2011. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics* 227, 1 (2011), 494–521.

[10] Mireille Capitaine, Catherine Donati-Martin, and Delphine Féral. 2009. The largest eigenvalues of finite rank deformation of large Wigner matrices: convergence and nonuniversality of the fluctuations. *Annals of Probability* (2009), 1–47.

[11] Rui M Castro et al. 2014. Adaptive sensing performance lower bounds for sparse signal detection and support estimation. *Bernoulli* 20, 4 (2014), 2217–2246.

[12] Rui M Castro and Ervin Tánczos. 2017. Adaptive Compressed Sensing for Support Recovery of Structured Sparse Sets. *IEEE Transactions on Information Theory* 63, 3 (2017), 1535–1554.

[13] Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. 2016. On Bayes Risk Lower Bounds. *Journal of Machine Learning Research* 17 (2016), 1–58.

[14] Imre Csiszár. 1972. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica* 2, 1-4 (1972), 191–213.

[15] James W Demmel. 1997. *Applied numerical linear algebra*. SIAM.

[16] Delphine Féral and Sandrine Péché. 2007. The largest eigenvalue of rank one deformation of large Wigner matrices. *Communications in mathematical physics* 272, 1 (2007), 185–228.

[17] Dan Garber, Elad Hazan, Chi Jin, Sham M Kakade, Cameron Musco, Praneeth Netrapalli, and Aaron Sidford. 2016. Faster eigenvector computation via shift-and-invert preconditioning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML*. 1396–1404.

[18] Roger A Horn and Charles R Johnson. 2012. *Matrix analysis*. Cambridge university press.

[19] Kevin G Jamieson, Robert Nowak, and Ben Recht. 2012. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*. 2672–2680.

[20] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. 2017. How to Escape Saddle Points Efficiently. *arXiv preprint arXiv:1703.00887* (2017).

[21] Ian Jolliffe. 2002. *Principal component analysis*. Wiley Online Library.

[22] Olav Kallenberg. 2006. *Foundations of modern probability*. Springer Science & Business Media.

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

[24] Jelani Nelson, Jakub Pachocki, and Zhengyu Wang. 2017. Optimal lower bounds for universal relation, samplers, and finding duplicates. *arXiv preprint arXiv:1703.08139* (2017).

[25] Arkadi Nemirovskii, David Borinovich Yudin, and Edgar Ronald Dawson. 1983. Problem complexity and method efficiency in optimization. (1983).

[26] Andrew Y Ng, Michael I Jordan, et al. [n. d.]. On spectral clustering: Analysis and an algorithm.

[27] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web*. Technical Report. Stanford InfoLab.

[28] Eric Price and David P Woodruff. 2013. Lower bounds for adaptive sparse recovery. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 652–663.

[29] Ohad Shamir. 2014. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Advances in Neural Information Processing Systems*. 163–171.

[30] Ohad Shamir. 2015. Fast stochastic algorithms for SVD and PCA: Convergence properties and convexity. *arXiv preprint arXiv:1507.08788* (2015).

[31] Max Simchowitz, Ahmed El Alaoui, and Benjamin Recht. 2018. Tight Query Complexity Lower Bounds for PCA via Finite Sample Deformed Wigner Law. *arXiv preprint (to appear)* (2018).

[32] Daniel A Spielman. 2007. Spectral graph theory and its applications. In *Foundations of Computer Science, 2007. FOCS’07. 48th Annual IEEE Symposium on*. IEEE, 29–38.

[33] Jacob Steinhardt and John C Duchi. 2015. Minimax rates for memory-bounded sparse linear regression. In *COLT*. 1564–1587.

[34] Jacob Steinhardt, Gregory Valiant, and Stefan Wagner. 2015. Memory, Communication, and Statistical Queries. In *Electronic Colloquium on Computational Complexity (ECCC)*, Vol. 22. 1–2.

[35] Roman Vershynin. [n. d.]. High Dimensional Probability. ([n. d.]).

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