Testing Matrix Rank, Optimally*

Maria-Florina Balcan[†] Yi Li[‡] David P. Woodruff[§] Hongyang Zhang[¶]

Abstract

We show that for the problem of testing if a matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ has rank at most d, or requires changing an ϵ-fraction of entries to have rank at most d, there is a *non-adaptive* query algorithm making $\widetilde{\mathcal{O}}(d^2/\epsilon)$ queries. Our algorithm works for any field $\mathbb{F}$. This improves upon the previous $\mathcal{O}(d^2/\epsilon^2)$ bound (Krauthgamer and Sasson, SODA '03), and bypasses an $\Omega(d^2/\epsilon^2)$ lower bound of (Li, Wang, and Woodruff, KDD '14) which holds if the algorithm is required to read a submatrix. Our algorithm is the first such algorithm which does not read a submatrix, and instead reads a carefully selected non-adaptive pattern of entries in rows and columns of $\mathbf{A}$. We complement our algorithm with a matching $\widetilde{\Omega}(d^2/\epsilon)$ query complexity lower bound for non-adaptive testers over any field. We also give tight bounds of $\widetilde{\Theta}(d^2)$ queries in the sensing model for which query access comes in the form of $\langle \mathbf{X}_i, \mathbf{A}\rangle := \mathrm{tr}(\mathbf{X}_i^\top \mathbf{A})$; perhaps surprisingly these bounds do not depend on ϵ.

Testing rank is only one of many tasks in determining if a matrix has low intrinsic dimensionality. We next develop a novel property testing framework for testing numerical properties of a real-valued matrix $\mathbf{A}$ more generally, which includes the stable rank, Schatten-p norms, and SVD entropy. Specifically, we propose a *bounded entry model*, where $\mathbf{A}$ is required to have entries bounded by 1 in absolute value. Such a model provides a meaningful framework for testing numerical quantities and avoids trivialities caused by single entries being arbitrarily large. It is also well-motivated by recommendation systems. We give upper and lower bounds for a wide range of problems in this model, and discuss connections to the sensing model above. We obtain several results for estimating the operator norm that may be of independent interest. For example, we show that if the stable rank is constant, $\|\mathbf{A}\|_F = \Omega(n)$, and the singular value gap $\sigma_1(\mathbf{A})/\sigma_2(\mathbf{A}) = (1/\epsilon)^\gamma$ for any constant $\gamma > 0$, then the operator norm can be estimated up to a $(1 \pm \epsilon)$-factor non-adaptively by querying $\mathcal{O}(1/\epsilon^2)$ entries. This should be contrasted to adaptive methods such as the power method, or previous non-adaptive sampling schemes based on matrix Bernstein inequalities which read a $1/\epsilon^2 \times 1/\epsilon^2$ submatrix and thus make $\Omega(1/\epsilon^4)$ queries. Similar to our non-adaptive algorithm for testing rank, our scheme instead reads a carefully selected pattern of entries.

1 Introduction

Data intrinsic dimensionality is a central object of study in compressed sensing, sketching, numerical linear algebra, machine learning, and many other domains [27, 20, 37, 36, 10, 41, 40]. In compressed sensing and sketching, the study of intrinsic dimensionality has led to significant advances in compressing the data to a size that is far smaller than the ambient dimension while still preserving useful properties of the signal [30, 3]. In numerical linear algebra and machine learning, understanding intrinsic dimensionality serves as a necessary condition for the success of various subspace recovery problems [15], e.g., matrix completion [6, 38, 13, 16, 33] and robust PCA [5, 39, 8]. The focus of this work is on the intrinsic dimensionality of matrices, such as the rank, stable rank, Schatten-p norms, and SVD entropy. The stable rank is defined to be the squared ratio of the Frobenius norm and the largest singular value, and the Schatten-p norm is the ℓ_p norm of the singular values (see Eqn. (2.3) for our definition of SVD entropy). We study these quantities in the framework of non-adaptive property testing [31, 9, 11]: given non-adaptive query access to the unknown matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ over a field $\mathbb{F}$, our goal is to determine whether $\mathbf{A}$ is of dimension d (where dimension depends on the specific problem), or is ϵ-far from having this property. The latter means that at least an ϵ-fraction of entries of $\mathbf{A}$ should be modified in order to have dimension d. Query access typically comes in the form of reading a single entry of the matrix, though we will also discuss sensing models where a query returns the value $\langle \mathbf{X}_i, \mathbf{A}\rangle := \mathrm{tr}(\mathbf{X}_i^\top \mathbf{A})$ for a given $\mathbf{X}_i$. Without making assumptions on $\mathbf{A}$, we would like to choose our sample pattern or set $\{\mathbf{X}_i\}$ of query matrices so that the query complexity

[*] A full version of the paper is available at `https://arxiv.org/abs/1810.08171`.

[†] Carnegie Mellon University. Email: ninamf@cs.cmu.edu.

[‡] Nanyang Technological University. Email: yili@ntu.edu.sg.

[§] Carnegie Mellon University. Email: dwoodruf@cs.cmu.edu.

[¶] Carnegie Mellon University. Email: hongyanz@cs.cmu.edu. Corresponding author.

is as small as possible.

Despite a large amount of work on testing matrix rank, many fundamental questions remain open. In the rank testing problem in the sampling model, one such question is to design an efficient algorithm that can distinguish rank-d vs. ϵ-far from rank-d with optimal sample complexity. The best-known sampling upper bound for non-adaptive rank testing for general d is $\mathcal{O}(d^2/\epsilon^2)$, which is achieved simply by sampling an $\mathcal{O}(d/\epsilon) \times \mathcal{O}(d/\epsilon)$ submatrix uniformly at random [20]. For arbitrary fields $\mathbb{F}$, only an $\Omega((1/\epsilon)\log(1/\epsilon))$ lower bound for constant d is known [23].

Besides the rank problem above, testing many numerical properties of real matrices has yet to be explored. For example, it is unknown what the query complexity is for the stable rank, which is a natural relaxation of rank in applications. Other examples for which previously we had no bounds are the Schatten-p norms and SVD entropy. We discuss these problems in a new property testing framework that we call the *bounded entry model*. This model has many realistic applications in the Netflix challenge [19], where each entry of the matrix corresponds to the rating from a customer to a movie, ranging from 1 to 5. Understanding the query complexity of testing numerical properties in the bounded entry model is an important problem in recommendation systems and applications of matrix completion, where often entries are bounded.

1.1 Problem Setup, Related Work, and Our Results

Our work has two parts: (1) we resolve the query complexity of non-adaptive matrix rank testing, a well-studied problem in this model, and (2) we develop a new framework for testing numerical properties of real matrices, including the stable rank, the Schatten-p norms and the SVD entropy. Our results are summarized in Table 1. We use $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ notation to hide polylogarithmic factors in the arguments inside. For the rank testing results, the hidden polylogarithmic factors depend only on d and $1/\epsilon$ and do not depend on n; for the other problems, they may depend on n.

Rank Testing. We first study the rank testing problem when we can only non-adaptively query entries. The goal is to design a sampling scheme on the entries of the unknown matrix $\mathbf{A}$ and an algorithm so that we can distinguish whether $\mathbf{A}$ is of rank d, or at least an ϵ-fraction of entries of $\mathbf{A}$ should be modified in order to reduce the rank to d. This problem was first proposed by Krauthgamer and Sasson in [20] with a sample complexity upper bound of $\mathcal{O}(d^2/\epsilon^2)$. In this work, we improve this to $\widetilde{\mathcal{O}}(d^2/\epsilon)$ for every d and ϵ, and complement this with a matching lower bound, showing that any algorithm with constant success probability requires at least $\widetilde{\Omega}(d^2/\epsilon)$ samples:

Theorems 3.1, 3.3, and 3.4 (Informal)**.** *For any matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ over any field, there is a randomized non-adaptive sampling algorithm which reads $\widetilde{\mathcal{O}}(d^2/\epsilon)$ entries and runs in $\mathsf{poly}(d/\epsilon)$ time, and with high probability correctly solves the rank testing problem. Further, any non-adaptive algorithm with constant success probability requires $\widetilde{\Omega}(d^2/\epsilon)$ samples over $\mathbb{R}$ or any finite field.*

Our non-adaptive sample complexity bound of $\widetilde{\mathcal{O}}(d^2/\epsilon)$ matches what is known with adaptive queries [23], and thus we show the best known upper bound might as well be non-adaptive.

New Framework for Testing Matrix Properties. Testing rank is only one of many tasks in determining if a matrix has low intrinsic dimensionality. In several applications, we require a less fragile measure of the collinearity of rows and columns, which is known as the stable rank [34]. We introduce what we call the *bounded entry model* as a new framework for studying such problems through the lens of property testing. In this model, we require all entries of a matrix to be bounded by 1 in absolute value. Boundedness has many natural applications in recommendation systems, e.g., the user-item matrix of preferences for products by customers has bounded entries in the Netflix challenge [19]. Indeed, there are many user rating matrices, etc., which naturally have a small number of discrete values, and therefore fit into a bounded entry model. The boundedness of entries also avoids trivialities in which one can modify a matrix to have a property by setting a single entry to be arbitrarily large, which, e.g., could make the stable rank arbitrarily close to 1.

Our model is a generalization of previous work in which stable rank testing was done in a model for which all rows had to have bounded norm [23], and the algorithm is only allowed to change entire rows at a time. As our non-adaptive rank testing algorithm will illustrate, one can sometimes do better by only reading certain carefully selected entries in rows and columns. Indeed, this is precisely the source of our improvement over prior work. Thus, the restriction of having to read an entire row is often unnatural, and further motivates our bounded entry model. We first informally state our main theorems on stable rank testing in this model.

Theorem 4.2 (Informal)**.** *There is a randomized algorithm for the stable rank testing problem to decide whether a matrix is of stable rank at most d or is ϵ-far*

Table 1: Query complexity results in this paper for non-adaptive testing of the rank, stable rank, Schatten-p norms, and SVD entropy. The testing of the stable rank, Schatten p-norm and SVD entropy are considered in the bounded entry model.

Testing Problems	Rank	Stable Rank	Schatten-p Norm	Entropy
Sampling	$\widetilde{\mathcal{O}}(d^2/\epsilon)$ (all fields) $\widetilde{\Omega}(d^2/\epsilon)$ (finite fields and $\mathbb{R}$)	$\widetilde{\mathcal{O}}(d^3/\epsilon^4)$ $\widetilde{\Omega}(d^2/\epsilon^2)^\dagger$	$\widetilde{\mathcal{O}}(1/\epsilon^{4p/(p-2)})$ $(p > 2)$ $\Omega(n)$ $(p \in [1, 2))$	$\Omega(n)^\dagger$
Sensing	$\mathcal{O}(d^2)$ (all fields) $\widetilde{\Omega}(d^2)$ (finite fields)	$\widetilde{\mathcal{O}}(d^{2.5}/\epsilon^2)$ $\widetilde{\Omega}(d^2/\epsilon^2)^\dagger$		

† The lower bound involves a reparameterization of the testing problem. See the respective theorem for details.

from stable rank at most d, with failure probability at most $1/3$, and which reads $\widetilde{\mathcal{O}}(d^3/\epsilon^4)$ entries.

Theorem 4.2 relies on a new $(1 \pm \tau)$-approximate non-adaptive estimator of the largest singular value of a matrix, which may be of independent interest.

Theorem 4.1 (Informal)**.** *Suppose that $\mathbf{A} \in \mathbb{R}^{n\times n}$ has stable rank $\mathcal{O}(d)$ and $\|\mathbf{A}\|_F^2 = \Omega(\tau n^2)$. Then in the bounded entry model, there is a randomized non-adaptive sampling algorithm which reads $\widetilde{\mathcal{O}}(d^2/\tau^4)$ entries and with probability at least 0.9, outputs a $(1 \pm \tau)$-approximation to the largest singular value of $\mathbf{A}$.*

We remark that when the stable rank is constant and the singular value gap $\sigma_1(\mathbf{A})/\sigma_2(\mathbf{A}) = (1/\tau)^\gamma$ for an arbitrary constant $\gamma > 0$, the operator norm can be estimated up to a $(1 \pm \tau)$-factor by querying $\mathcal{O}(1/\tau^2)$ entries non-adaptively. We refer the readers to the full version for the details.

Other measures of intrinsic dimensionality include matrix norms, such as the Schatten-p norm $\|\cdot\|_{\mathcal{S}_p}$, which measures the central tendency of the singular values. Familiar special cases are $p = 1$, 2 and ∞, which have applications in differential privacy [14] and non-convex optimization [5, 12] for $p = 1$, and in numerical linear algebra [29] for $p \in \{2, \infty\}$. Matrix norms have been studied extensively in the streaming literature [21, 24, 25, 26], though their study in property testing models is lacking.

We study non-adaptive algorithms for these problems in the bounded entry model. We consider distinguishing whether $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ is at least cn^p for $p > 2$ (at least $cn^{1+1/p}$ for $p < 2$), or at least an ϵ-fraction of entries of $\mathbf{A}$ should be modified in order to have this property, where c is a constant (depending only on p). We choose the threshold n^p for $p > 2$ and $n^{1+1/p}$ for $p < 2$ because they are the largest possible value of $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ for $\mathbf{A}$ under the bounded entry model. When $p > 2$, $\|\mathbf{A}\|_{\mathcal{S}_p}$ is maximized when $\mathbf{A}$ is of rank 1, and so this gives us an alternative "measure" of how close we are to a rank-1 matrix. Testing whether $\|\mathbf{A}\|_{\mathcal{S}_p}$ is large in sublinear time allows us to quickly determine whether $\mathbf{A}$ can be well approximated by a low-rank matrix, which could save us from running more expensive low-rank approximation algorithms. In contrast, when $p < 2$, $\|\mathbf{A}\|_{\mathcal{S}_p}$ is maximized when $\mathbf{A}$ has a flat spectrum, and so is a measure of how well-conditioned $\mathbf{A}$ is. A fast tester could save us from running expensive pre-conditioning algorithms. We state our main theorems informally below.

Theorem 4.4 (Informal)**.** *For constant $p > 2$, there is a randomized algorithm for the Schatten-p norm testing problem with failure probability at most $1/3$ which reads $\widetilde{\mathcal{O}}(1/\epsilon^{4p/(p-2)})$ entries.*

Results for Sensing Algorithms. We also consider a more powerful query oracle known as the *sensing model*, where query access comes in the form of $\langle \mathbf{X}_i, \mathbf{A}\rangle := \mathrm{tr}(\mathbf{X}_i^\top \mathbf{A})$ for some sensing matrices $\mathbf{X}_i$ of our choice. These matrices are chosen non-adaptively. We show differences in the complexity of the above problems in this and the above sampling model. For the testing and the estimation problems above, we have the following results in the sensing model:

Theorem 3.5 (Informal)**.** *Over an arbitrary finite field, any non-adaptive algorithm with constant success probability for the rank testing problem in the sensing model requires $\widetilde{\Omega}(d^2)$ queries.*

Theorems 4.2 and 4.3 (Informal)**.** *There is a randomized algorithm for the stable rank testing problem with failure probability at most $1/3$ in the sensing model with $\widetilde{\mathcal{O}}(d^{2.5}/\epsilon^2)$ queries. Further, any algorithm with constant success probability requires $\widetilde{\Omega}(d^2/\epsilon^2)$ queries.*

Theorem 4.5 (Informal)**.** *For $p \in [1, 2)$, any algorithm for the Schatten-p norm testing problem with failure probability at most $1/3$ requires $\Omega(n)$ queries.*

Theorem 4.1 (Informal)**.** *Suppose that $\mathbf{A} \in \mathbb{R}^{n\times n}$ has stable rank $\mathcal{O}(d)$ and $\|\mathbf{A}\|_F^2 = \Omega(\tau n^2)$. In the bounded entry model, there is a randomized sensing algorithm with sensing complexity $\widetilde{\mathcal{O}}(d^2/\tau^2)$ which* *outputs a* $(1 \pm \tau)$*-approximation to the largest singular value with probability at least* 0.9*. This sensing complexity is optimal up to polylogarithmic factors.*

We also provide an $\Omega(n)$ query lower bound for SVD entropy testing in the sensing model, see Section 4.3.

1.2 Our Techniques We now discuss the techniques in more detail, starting with the rank testing problem.

Prior to the work of [23], the only known algorithm for $d = 1$ was to sample an $\mathcal{O}(1/\epsilon) \times \mathcal{O}(1/\epsilon)$ submatrix. In contrast, for rank 1 an algorithm in [23] samples $\mathcal{O}(\log(1/\epsilon))$ blocks of varying shapes "within a random $\mathcal{O}(1/\epsilon) \times \mathcal{O}(1/\epsilon)$ submatrix" and argues that these shapes are sufficient to expose a rank-2 submatrix. For $d = 1$ the goal is to augment a 1×1 matrix to a full-rank 2×2 matrix. One can show that with good probability, one of the shapes "catches" an entry that enlarges the 1×1 matrix to a full-rank 2×2 matrix. For instance, in Figure 1, (r, c) is our 1×1 matrix and the leftmost vertical block catches an "augmentation element" (r', c') which makes $\begin{bmatrix} (r,c') & (r,c) \\ (r',c') & (r',c) \end{bmatrix}$ a full-rank 2×2 matrix. Hereby, the "augmentation element" means the entry by adding which we augment a $r \times r$ matrix to a $(r + 1) \times (r + 1)$ matrix. In [23], an argument was claimed for $d = 1$, though we note an omission in their analysis. Namely, the "augmentation entry" (r', c') can be the 1×1 matrix we begin with (meaning that $\mathbf{A}_{r',c'} \neq 0$, which might not be true), and since one can show that both (r, c) and (r', c') fall inside the same sampling block with good probability, the 2×2 matrix would be fully observed and the algorithm would thus be able to determine that it has rank 2. However, it is possible that $\mathbf{A}_{r',c'} = 0$ and (r', c') would not be a starting point (i.e., a 1×1 rank-1 matrix), and in this case, (r', c) may not be observed, as illustrated in Figure 1. In this case the algorithm will not be able to determine whether the augmented 2×2 matrix is of full rank. For $d > 1$, nothing was known. One issue is that the probability of fully observing a $d \times d$ submatrix within these shapes is very small. To overcome this, we propose what we call *rebasing* and *transformation to a canonical structure*. These arguments allow us to tolerate unobserved entries and conveniently obtain an algorithm for every d, completing the analysis of [23] for $d = 1$ in the process.

Rebasing Argument + Canonical Structure. The best previous result for the rank testing problem uniformly samples an $\mathcal{O}(d/\epsilon) \times \mathcal{O}(d/\epsilon)$ submatrix and argues that one can find a $(d + 1) \times (d + 1)$ full-rank submatrix within it when $\mathbf{A}$ is ϵ-far from rank-d [20]. In contrast, our algorithm follows from subsampling an $\mathcal{O}(\epsilon)$-fraction of entries in this $\mathcal{O}(d/\epsilon) \times \mathcal{O}(d/\epsilon)$ submatrix. Let $\mathcal{R}_1 \subseteq \cdots \subseteq \mathcal{R}_m$ and $\mathcal{C}_1 \supseteq \cdots \supseteq \mathcal{C}_m$ be the indices of subsampled rows and columns, respectively, with $m = \mathcal{O}(\log(1/\epsilon))$. We choose these indices uniformly at random such that $|\mathcal{R}_i| = \widetilde{\mathcal{O}}(d2^i)$ and $|\mathcal{C}_i| = \widetilde{\mathcal{O}}(d/(2^i\epsilon))$, and sample the entries in all m blocks determined by the $\{\mathcal{R}_i, \mathcal{C}_i\}$ (see Figure 1, where our sampled regions are enclosed by the dotted lines). Since there are $\widetilde{\mathcal{O}}(\log(1/\epsilon))$ blocks and in each block we sample $\widetilde{\mathcal{O}}(d^2/\epsilon)$ entries, the sample complexity of our algorithm is as small as $\widetilde{\mathcal{O}}(d^2/\epsilon)$.

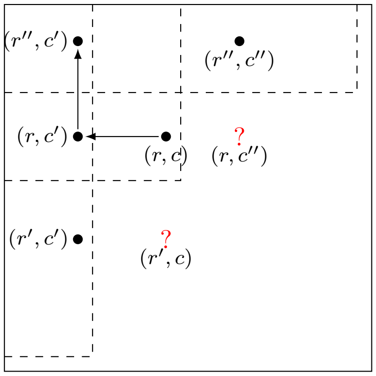

Figure 1: Our sampling scheme (the region enclosed by the dotted lines modulo permutation of rows and columns) and our path of augmenting a 1×1 submatrix. The whole region is the $\mathcal{O}(d/\epsilon) \times \mathcal{O}(d/\epsilon)$ submatrix sampled from the $n \times n$ matrix.

The correctness of our algorithm for $d = 1$ follows from what we call a rebasing argument. Starting from an empty matrix, our goal is to maintain and augment the matrix to a 2×2 full-rank matrix when $\mathbf{A}$ is ϵ-far from rank-d. By a level-set argument, we show an oracle lemma which states that *we can augment any* $r \times r$ *full-rank matrix to an* $(r + 1) \times (r + 1)$ *full-rank matrix by an augmentation entry in the sampled region*, as long as $r \le d$ and $\mathbf{A}$ is ϵ-far from rank-d. Therefore, as a first step we successfully find a 1×1 full-rank matrix, say with index (r, c), in the sampled region. We then argue that we can either (a) find a 2×2 fully-observed full-rank submatrix or a 2×2 submatrix which is not fully observed but we know must be of full rank, or (b) move our maintained 1×1 full-rank submatrix upwards or leftwards to a new 1×1 *full-rank* submatrix and repeat checking whether case (a) happens or not; if not, we implement case (b) again and repeat the procedure. To see case (a), by the oracle lemma, if the augmented entry is (r'', c') (see Figure 1), then we fully observe the submatrix determined by (r'', c') and (r, c) and so the algorithm is correct in this case. On the other hand, if the augmented entry is (r', c'), then we fail to see the entry at (r', c). In this case, when $\mathbf{A}_{r,c'} = 0$, then we must have $\mathbf{A}_{r',c'} \neq 0$; otherwise, (r', c') is not an augment of (r, c), which leads to a contradiction with the oracle lemma. Thus we find a 2×2 matrix with structure

$$(1.1) \qquad \begin{bmatrix} \mathbf{A}_{r,c'} & \mathbf{A}_{r,c} \\ \mathbf{A}_{r',c'} & \mathbf{A}_{r',c} \end{bmatrix} = \begin{bmatrix} 0 & \neq 0 \\ \neq 0 & ? \end{bmatrix},$$

which must be of rank 2 despite an unobserved entry, and the algorithm therefore is correct in this case. The remaining case of the analysis above is when $\mathbf{A}_{r,c'} \neq 0$. Instead of trying to augment $\mathbf{A}_{r,c}$, we augment $\mathbf{A}_{r,c'}$ in the next step. Note that the index (r, c') is to the left of (r, c). This leads to case (b). In the worst case, we move the 1×1 non-zero matrix to the uppermost left corner,[1] e.g., (r'', c'). Fortunately, since (r'', c') is in the uppermost left corner, we can, as guaranteed by the oracle lemma, augment it to a 2×2 *fully-observed* full-rank matrix. Again the algorithm outputs correctly in this case.

The analysis becomes more challenging for general d, since the number of unobserved/unimportant entries (i.e., those entries marked as "?") may propagate as we augment an $r \times r$ submatrix ($r = 1, 2, ..., d$) in each round. To resolve the issue, we maintain a structure (modulo elementary transformations) similar to structure (1.1) for the $r \times r$ submatrix, that is,

$$(1.2) \qquad \begin{bmatrix} 0 & 0 & \cdots & 0 & \cdots & 0 & \neq 0 \\ 0 & 0 & \cdots & 0 & \cdots & \neq 0 & ? \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & \neq 0 & \cdots & ? & \cdots & ? & ? \\ \neq 0 & ? & \cdots & ? & \cdots & ? & ? \end{bmatrix}.$$

Since the proposed structure has non-zero determinant, the submatrix is always of full rank. Similar to the case for $d = 1$, we show that we can either (a) augment the $r \times r$ submatrix to an $(r+1) \times (r+1)$ submatrix with the same structure (1.2) (modulo elementary transformations); or (b) find another $r \times r$ submatrix of structure (1.2) that is closer to the upper-left corner than the original $r \times r$ matrix. Hence the algorithm is correct for general d. More details are provided in the proof sketch of Theorem 3.1.

Pivot-Node Assignment. Our rank testing lower bound under the sampling model over a finite field $\mathbb{F}$ follows from distinguishing two hard instances $\mathbf{U}\mathbf{V}^\top$ vs. $\mathbf{W}$, where $\mathbf{U}, \mathbf{V} \in \mathbb{F}^{t \times d}$ and $\mathbf{W} \in \mathbb{F}^{t \times t}$ have i.i.d. entries that are uniform over $\mathbb{F}$. For an observed subset $\mathcal{S}$ of entries with $|\mathcal{S}| = \mathcal{O}(d^2)$, we bound the total variation distance between the distributions of the observed entries in the two cases by a small constant. In particular, we show that the probability $\Pr[(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}]$ is large for any observation $\mathbf{x} \in \mathbb{F}^{|\mathcal{S}|}$, by a *pivot-node assignment* argument, as follows. We reformulate our problem as a bipartite graph assignment problem $G = (L \cup R, E)$, where L corresponds to the rows of $\mathbf{U}$, R the rows of $\mathbf{V}$ and each edge of E one entry in $\mathcal{S}$. We want to assign each node a vector/affine subspace, meaning that the corresponding row in $\mathbf{U}$ or $\mathbf{V}$ will be that vector or in that affine subspace, such that they agree with our observation, i.e., $(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}$. Since $\mathbf{U}, \mathbf{V}$ are random matrices, we assign random vectors to nodes adaptively, one at a time, and try to maintain consistency with the fact that $(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}$. Note that the order of the assignment is important, as a bad choice for an earlier node may invalidate any assignment to a later node. To overcome this issue, we choose nodes of large degrees as *pivot nodes* and assign each non-pivot node adaptively in a careful manner so as to guarantee that the incident pivot nodes will always have valid assignments (which in fact form an affine subspace). In the end we assign the pivot node vectors from their respective affine subspaces. We employ a counting argument for each step in this assignment procedure to lower bound the number of valid assignments, and thus lower bound the probability $\Pr[(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}]$.

The above analysis gives us an $\Omega(d^2)$ lower bound for constant ϵ since $\mathbf{W}$ is constant-far from being of rank d. The desired $\Omega(d^2/\epsilon)$ lower bound follows from planting $\mathbf{U}\mathbf{V}^\top$ vs. $\mathbf{W}$ with $t = \sqrt{\epsilon}n$ into an $n \times n$ matrix at uniformly random positions, and padding zeros everywhere else.

New Analytical Framework for Stable Rank, Schatten-p Norm, and Entropy Testing. We propose a new analytical framework by reducing the testing problem to a sequence of estimation problems *without involving* $\mathsf{poly}(n)$ *in the sample complexity.* There is a two-stage estimation in our framework: (1) a constant-approximation to some statistic X of interest (e.g., stable rank) which enables us to distinguish $X \leq d$ vs. $X \geq 10d$ for the threshold parameter d of interest. If $X \geq 10d$, we can safely output "$\mathbf{A}$ is far from $X \leq d$"; otherwise, the statistic is at most $10d$, and (2) we show that X has a $(1 \pm \epsilon)$-factor difference between "$X \leq d$" and "far from $X \leq d$", and so we implement a more accurate $(1 \pm \epsilon)$-approximation to distinguish the two cases. The sample complexity does not depend on n polynomially because (1) the first estimator is "rough" and gives only a constant-factor approximation and (2) the

[1] The upper-left corner refers to the intersection of all sampled blocks, namely, $\mathcal{R}_1 \times \mathcal{C}_m$; it does not mean the top-left entry.

second estimator operates under the condition that $X \leq 10d$ and thus $\mathbf{A}$ has a low intrinsic dimension. We apply the proposed framework to the testing problems of the stable rank and the Schatten-p norm by plugging in our estimators in Theorem B.1 and Theorem B.3. This analytical framework may be of independent interest to other property testing problems more broadly.

In a number of these problems, a key difficulty is arguing about spectral properties of a matrix $\mathbf{A}$ when it is ϵ-far from having a property, such as having stable rank at most d. Because of the fact that the entries must always be bounded by 1 in absolute value, it becomes non-trivial to argue, for example, that if $\mathbf{A}$ is ϵ-far from having stable rank at most d, that its stable rank is even slightly larger than d. A natural approach is to argue that you could change an ϵ-fraction of rows of $\mathbf{A}$ to agree with a multiple of the top left or right singular vector of $\mathbf{A}$, and since we are still guaranteed to have stable rank at least d after changing such entries, it means that the operator norm of $\mathbf{A}$ must have been small to begin with (which says something about the original stable rank of $\mathbf{A}$, since its Frobenius norm can also be estimated). The trouble is, if the top singular vector has some entries that are very large, and others that are small, one cannot scale the singular vector by a large amount since then we would violate the boundedness criterion of our model. We get around this by arguing there either needs to exist a left or a right singular vector of large ℓ_1-norm (in some cases such vectors may only be right singular vectors, and in other cases only left singular vectors). The ℓ_1-norm is a natural norm to study in this context, since it is dual to the ℓ_∞-norm, which we use to capture the boundedness property of the matrix.

Our lower bounds for the above problems follow from the corresponding sketching lower bounds for the estimation problem in [25, 22], together with rigidity-type results [35] for the hard instances regarding the respective statistic of interest.

2 Preliminaries

We shall use bold capital letters $\mathbf{A}$, $\mathbf{B}$, ... to represent matrices, bold lower-case letters $\mathbf{u}$, $\mathbf{v}$, ... to represent vectors, and lower-case letters a, b, ... to represent scalars. We adopt the convention of abbreviating the set $\{1, 2, ..., n\}$ as $[n]$. We write $f \gtrsim g$ (resp. $f \lesssim g$) if there exists a constant $C > 0$ such that $f \geq Cg$ (resp. $f \leq Cg$).

For matrix functions, denote by $\mathsf{rank}(\mathbf{A})$ and $\mathsf{srank}(\mathbf{A})$ the rank and the stable rank of $\mathbf{A} \neq \mathbf{0}$, respectively. It always holds that $1 \leq \mathsf{srank}(\mathbf{A}) \leq \mathsf{rank}(\mathbf{A})$. For matrix norms, let $\|\mathbf{A}\|_{\mathcal{S}_p}$ denote the Schatten-p norm of $\mathbf{A}$, defined as $\|\mathbf{A}\|_{\mathcal{S}_p} = \left(\sum_{i=1}^n \sigma_i^p(\mathbf{A})\right)^{1/p}$. The Frobenius norm $\|\mathbf{A}\|_F$ is a special case of the Schatten-p norm when $p = 2$, the operator norm or the spectral norm (the largest singular value) of $\|\mathbf{A}\|$ equals to the limit as $p \to +\infty$. When $0 < p < 1$, $\|\mathbf{A}\|_{\mathcal{S}_p}$ is not a norm but is still a well-defined quantity, and it tends to $\mathsf{rank}(\mathbf{A})$ as $p \to 0^+$. Let $\|\mathbf{A}\|_0$ denote the number of non-zero entries in $\mathbf{A}$, and $\|\mathbf{A}\|_\infty$ denote the entrywise ℓ_∞ norm of $\mathbf{A}$, i.e., $\|\mathbf{A}\|_\infty = \max_{i,j} |\mathbf{A}_{i,j}|$. The *rigidity* of a matrix $\mathbf{A}$ over a field $\mathbb{F}$, denoted by $\mathcal{R}_{\mathbf{A}}^{\mathbb{F}}(r)$, is the least number of entries of $\mathbf{A}$ that must be changed in order to reduce the rank of $\mathbf{A}$ to a value at most r:

$$\mathcal{R}_{\mathbf{A}}^{\mathbb{F}}(r) := \min\{\|\mathbf{C}\|_0 : \mathsf{rank}_{\mathbb{F}}(\mathbf{A} + \mathbf{C}) \leq r\}.$$

Sometimes we may omit the subscript $\mathbf{A}$ in $\mathcal{R}_{\mathbf{A}}^{\mathbb{F}}(r)$ when the matrix of interest is clear from the context.

We define the entropy of an unnormalized distribution $(p_1, \ldots, p_n)$ $(0 < p_1 + \cdots + p_n \leq 1$ with $p_i \geq 0$ for all $i)$ to be

$$H(p_1, \ldots, p_n) = -\sum_i p_i \log p_i.$$

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, we define its (SVD) entropy as

$$(2.3) \quad H(\mathbf{A}) = H\left(\frac{\sigma_1^2(\mathbf{A})}{n^2}, \ldots, \frac{\sigma_n^2(\mathbf{A})}{n^2}\right) = \frac{-\sum_i \frac{\sigma_i^2(\mathbf{A})}{n^2} \log \frac{\sigma_i^2(\mathbf{A})}{n^2}}{\sum_i \frac{\sigma_i^2(\mathbf{A})}{n^2}}$$

with the convention that $0 \cdot \infty = 0$. For matrices $\mathbf{A}$ satisfying $\|\mathbf{A}\|_\infty \leq 1$, it holds that $\sigma_i(\mathbf{A}) \leq n$ for all i and the entropy above coincides with the usual Shannon entropy. Note that scaling only changes the entropy additively; that is, $H(\beta\mathbf{A}) = H(\mathbf{A}) - \log \beta^2$.

Let $\mathcal{G}(m, n)$ denote the distribution of $m \times n$ i.i.d. standard Gaussian matrix over $\mathbb{R}$ and $\mathcal{U}_{\mathbb{F}}(m, n)$ (or $\mathcal{U}(\mathcal{S})$) represent $m \times n$ i.i.d. uniform matrix over a finite field $\mathbb{F}$ (or a finite set $\mathcal{S}$). We use $d_{TV}(\mathcal{L}_1, \mathcal{L}_2)$ to denote the total variation distance between two distributions $\mathcal{L}_1$ and $\mathcal{L}_2$.

We shall also frequently use c, c', c_0, C, C', C_0, etc., to represent constants, which are understood to be absolute constants unless the dependency is otherwise specified.

3 Non-Adaptive Rank Testing

In this section, we study the problem of testing matrix rank, which is defined in Section 1.1.

3.1 Positive Results We provide the first non-adaptive algorithm for the rank testing problem in the sampling model with near-optimal $\widetilde{\mathcal{O}}(d^2/\epsilon)$ queries when $\epsilon \le 1/e$ and $d \ge 1$. Let $\eta \in (0, 1/2)$ be such that $\eta \log(1/\eta) = \epsilon$ and let $m = \lceil \log(1/\eta) \rceil$. We present our algorithm in Algorithm 1.

Algorithm 1 Non-adaptive testing of matrix rank in sampling model

```
Choose $\mathcal{R}_1, \dots, \mathcal{R}_m$ and $\mathcal{C}_1, \dots, \mathcal{C}_m$ from $\{1, 2, ..., n\}$ uniformly at random such that $\mathcal{R}_1 \subseteq \cdots \subseteq \mathcal{R}_m$, $\mathcal{C}_1 \supseteq \cdots \supseteq \mathcal{C}_m$, and $|\mathcal{R}_i| = c[\log d + \log\log(1/\eta)]d\log(1/\eta)2^i$, $|\mathcal{C}_i| = c[\log d + \log\log(1/\eta)]d\log(1/\eta)/(2^i\eta)$, where $c > 0$ is an absolute constant. To impose containment for $\mathcal{R}_i$'s, $\mathcal{R}_i$ can be formed by appending to $\mathcal{R}_{i-1}$ a uniformly random set of $|\mathcal{R}_i| - |\mathcal{R}_{i-1}|$ rows. The containment for $\mathcal{C}_i$'s can be imposed similarly.
Query the entries in $\mathcal{Q} = \bigcup_{i=1}^m (\mathcal{R}_i \times \mathcal{C}_i)$. Note that the entries in $(\mathcal{R}_m \times \mathcal{C}_1) \setminus \mathcal{Q}$ are unobserved. The algorithm solves the following minimization problem by filling in those entries of $\mathbf{A}_{(\mathcal{R}_m \times \mathcal{C}_1)\setminus\mathcal{Q}}$ given input $\mathbf{A}_\mathcal{Q}$
    (3.4)    $r := \min_{\mathbf{A}_{(\mathcal{R}_m \times \mathcal{C}_1)\setminus\mathcal{Q}}} \mathsf{rank}(\mathbf{A}_{\mathcal{R}_m, \mathcal{C}_1})$.
Output "$\mathbf{A}$ is $\epsilon$-far from having rank $d$" if $r > d$; otherwise, output "$\mathbf{A}$ is of rank at most $d$".
```

We note that the number of entries that Algorithm 1 queries is as small as $\widetilde{\mathcal{O}}(d^2/\epsilon)$. Furthermore, subproblem (3.4) in the algorithm can be solved in $\mathsf{poly}(d/\epsilon)$ time. The following theorem guarantees the correctness of Algorithm 1.

Theorem 3.1. (Sampling upper bound over all fields). *Let $\epsilon \le 1/e$ and $d \ge 1$. For any matrix $\mathbf{A}$ over any field $\mathbb{F}$, Algorithm 1 correctly solves the rank testing problem in the sampling model with probability at least $1 - 1/\mathsf{poly}(d\log(1/\epsilon))$. The algorithm queries $\widetilde{\mathcal{O}}(d^2/\epsilon)$ entries and can be implemented to run in $\mathsf{poly}(d/\epsilon)$ time.*

Proof. If $\mathbf{A}$ is of rank at most d, then the algorithm will never make mistake, so we assume that $\mathbf{A}$ is ϵ-far from being rank d in the proof below.

The idea is that, we start with the base case of an empty matrix, and augment it to a full-rank $r \times r$ matrix in r rounds, where in each round we increase the dimension of the matrix by exactly one. Each round may contain several steps in which we move the intermediate $j \times j$ matrix ($j \le r$) towards the upper-left corner without augmenting it; here, moving the matrix towards the upper-left corner means changing $\mathbf{A}_{\mathcal{R},\mathcal{C}}$ to $\mathbf{A}_{\mathcal{R}',\mathcal{C}'}$, of the same rank, with $|\mathcal{R}'| = |\mathcal{R}| = |\mathcal{C}'| = |\mathcal{C}| = j$ and $\mathcal{R}' \preceq \mathcal{R}$ and $\mathcal{C}' \preceq \mathcal{C}$, where $\mathcal{R}' \preceq \mathcal{R}$ means that, suppose that $r'_1 < r'_2 < \cdots < r'_j$ are the (sorted) elements in $\mathcal{R}'$ and $r_1 < r_2 < \cdots < r_j$ are the (sorted) elements in $\mathcal{R}$, it holds that $r'_i \le r_i$ for all $1 \le i \le j$, and $\mathcal{C}' \preceq \mathcal{C}$ has a similar meaning.

The challenge is that those unobserved entries ?'s may propagate as we augment the submatrix in each round. Our goal is to prove that starting from a *structural* $(r-1) \times (r-1)$ full-rank submatrix which might have ?'s as its entries, no matter what values of *all* ?'s are, with the augment operator we either (1) make progress for $(r-1) \times (r-1)$ submatrix, or (2) obtain an $r \times r$ full-rank submatrix *with the same structure*. Let us first condition on the event in the following lemma holds true.

Lemma 3.1. (Lemma 6, [23]) *For fixed $(\mathcal{R}, \mathcal{C})$, suppose that $(\mathcal{R}, \mathcal{C})$ has augment pattern i on $\mathbf{A}$. Let $\mathcal{R}', \mathcal{C}' \subseteq [n]$ be uniformly random such that $|\mathcal{R}'| = c2^i$, $|\mathcal{C}'| = c/(2^i\eta)$. Then the probability that $(\mathcal{R}', \mathcal{C}')$ contains at least one augment of $(\mathcal{R}, \mathcal{C})$ on $\mathbf{A}$ is at least $1 - 2e^{-c/2}$.*

Regarding the structure, we have the following claim.

Claim 1. *There exists a searching path for $r \times r$ full-rank submatrices with non-decreasing r which has the following lower triangular form modulo an elementary transformation*

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & \cdots & 0 & \neq 0 \\ 0 & 0 & \cdots & 0 & \cdots & \neq 0 & ? \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \neq 0 & \cdots & ? & ? \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & \neq 0 & \cdots & ? & \cdots & ? & ? \\ \neq 0 & ? & \cdots & ? & \cdots & ? & ? \end{bmatrix}, \tag{3.5}$$

where $\neq 0$ denotes the known entry which is non-zero, and ? denotes an entry which can be either observed or unobserved.

Proof. [Proof of Claim 1] Without loss of generality, we assume that all ?'s are unobserved, which is the most challenging case; otherwise, the proof degenerates to the discussion of *central submatrix* in Case (iii) which we shall specify later. We prove the claim by induction. The base case $r = 0$ is true. Suppose the claims holds for $r - 1$. We now argue the correctness for r.

Let (p, q) be the augment. Denote the augment row by

$$\begin{bmatrix} y_1 & \cdots & y_b & \mathbf{A}_{p,q} & y_{b+2} & \cdots & y_r \end{bmatrix},$$

and the augment column by

$$\begin{bmatrix} x_1 & \cdots & x_a & \mathbf{A}_{p,q} & x_{a+2} & \cdots & x_r \end{bmatrix}^\top.$$

We now discuss three cases based on the relation between $a+b$ and r.

Case (i). $a+b=r-1$ **($\mathbf{A}_{p,q}$ is on the antidiagonal of $r\times r$ submatrix).**

In this case, $y_{b+2},\dots,y_r$ and $x_{a+2},\dots,x_r$ are all ?'s. We argue that $x_1=x_2=\cdots=x_a=0$ and $y_1=y_2=\cdots=y_b=0$; otherwise, we can make progress. First consider y_i for $1\le i\le b$. If some $y_i\neq 0$, we can delete the $(r-i)$-th row in the $(r-1)\times(r-1)$ submatrix and insert the augment row (without the augment entry $\mathbf{A}_{p,q}$), which is above the deleted row. Thus we obtain a new $(r-1)\times(r-1)$ submatrix towards the upper-left corner, and furthermore, the new submatrix exhibits the structure in (3.5). The same argument applies for $x_1,x_2,...,x_a$. Therefore, if no progress is made, it must hold that $x_1=x_2=...=x_a=0$ and $y_1=y_2=...=y_b=0$. In this case, $\mathbf{A}_{p,q}\neq 0$; otherwise, (p,q) is not an augment. Therefore we obtain an $r\times r$ full-rank matrix of the form (3.5).

Case (ii). $a+b<r-1$ **($\mathbf{A}_{p,q}$ is above the antidiagonal of $r\times r$ submatrix).**

In this case, $y_{r-a+1},\dots,y_r$ and $x_{r-b+1},\dots,x_r$ are all ?'s. Similarly to Case (i), we shall argue that $x_1=\cdots=x_a=x_{a+2}=\cdots=x_{r-b}=0$ and $y_1=\cdots=y_b=y_{b+2}=\cdots=y_{r-a}=0$; otherwise, we can make progress. To see this, consider first y_i for $1\le i\le b$ and then for $b+2\le i\le r-a$. If $y_i\neq 0$ for some $i\le b$, we can delete the $(r-i)$-th row in the $(r-1)\times(r-1)$ submatrix and insert the augment row (without the augment entry $\mathbf{A}_{p,q}$), which is above the deleted row, and so we make progress. Now assume that $y_1=\cdots=y_b=0$. If $y_i\neq 0$ for some i such that $b+2\le i\le r-a$, we can delete the $(r-i+1)$-st row in the $(r-1)\times(r-1)$ submatrix of the last step and insert the augment row (without the augment entry $\mathbf{A}_{p,q}$), which is above the deleted row. So we make progress towards the most upper left corner. The same argument applies to $x_1,\dots,x_a,x_{a+2},\dots,x_{r-b}$. Therefore, $x_1=\cdots=x_a=x_{a+2}=...=x_{r-b}=0$ and $y_1=\cdots=y_b=y_{b+2}=\cdots=y_{r-a}=0$. In this case, $\mathbf{A}_{p,q}\neq 0$; otherwise, (p,q) is not an augment since all possible choices of ?'s cannot make the $r\times r$ submatrix non-singular. By exchanging the $(a+1)$-st row and the $(r-b)$-th row of the $r\times r$ submatrix or exchanging the $(b+1)$-st column and the $(r-a)$-th column, we obtain an $r\times r$ submatrix of the form (3.5).

Case (iii). $a+b>r-1$ **($\mathbf{A}_{p,q}$ is below the antidiagonal of $r\times r$ submatrix).**

In this case, we argue that $x_i=y_j=0$ for all $i\le r-b-1$ and $j\le r-a-1$; otherwise we can make progress as Cases (i) and (ii) for y_j. To see this, let us discuss from $j=1$ to $r-a-1$. If $y_j\neq 0$ $(j=1,2,\dots,r-a-1)$, we can delete the $(r-j)$-th row in the $(r-1)\times(r-1)$ submatrix and insert the augment row (without the augment entry $\mathbf{A}_{p,q}$), which is above the deleted row. So we make progress. The same argument applies to $x_1,\dots,x_{r-b-1}$. So $x_i=y_j=0$ for all $i\le r-b-1$ and $j\le r-a-1$.

Given that there is only one non-zero entry in the first $r-b-1$ rows and the first $r-a-1$ columns of the $r\times r$ submatrix (i.e., the Laplace expansion of the determinant), we only need to focus on a minor corresponding to a $\min\{a,b\}\times\min\{a,b\}$ *central submatrix*, which decides whether the determinant of the $r\times r$ submatrix is zero and is fully-observed because the augment (p,q) is at the lower right corner of the central submatrix (see the red part in Eqn. (3.6)).

(3.6)
$$\begin{bmatrix}
0 & \cdots & 0 & \cdots & 0 & \cdots & \neq 0\\
0 & \cdots & 0 & \cdots & 0 & \cdots & ?\\
\vdots & & \vdots & & \vdots & & \vdots\\
\vdots & & \vdots & & \vdots & & \vdots\\
0 & \cdots & \neq 0 & \cdots & \text{known} & \cdots & ?\\
\vdots & & \vdots & & \vdots & & \vdots\\
0 & \cdots & \text{known} & \cdots & \text{augment } \mathbf{A}_{p,q} & \cdots & ?\\
\vdots & & \vdots & & \vdots & & \vdots\\
0 & \cdots & ? & \cdots & ? & \cdots & ?\\
\neq 0 & \cdots & ? & \cdots & ? & \cdots & ?
\end{bmatrix}.$$

Since it is fully-observed, the minor must be non-zero; otherwise, (p,q) cannot be an augment for all choices of ?'s. Therefore, we can do an elementary transformation to make the central submatrix a lower triangular matrix with non-zero antidiagonal entries. More importantly, such an elementary transformation also transforms the $r\times r$ matrix to a lower triangular matrix with non-zero antidiagonal entries, because all the entries to the left and above of the central matrix are 0's, and all the entries to the right and below of the central matrix are ?'s. Hence any elementary transformation keeps 0's and ?'s unchanged, and we obtain therefore an $r\times r$ submatrix of the form (3.5).□

Now we are ready to prove Theorem 3.1. Note that Lemma 3.1 works only for *fixed* $(\mathcal{R},\mathcal{C})$. To make the lemma applicable "for all" $(\mathcal{R},\mathcal{C})$ throughout the augmentation process, we shall take a union bound by choosing $|\mathcal{R}|$ and $|\mathcal{C}|$ large enough. Specifically, for each i, we divide $\mathcal{R}_i=\bigcup_{k=1}^{\ell}\mathcal{R}_i^{(k)}$ uniformly at random into $\ell = d + d\log(\frac{1}{\eta})$[2] even parts $\mathcal{R}_i^{(1)}, \mathcal{R}_i^{(2)}, \ldots, \mathcal{R}_i^{(d)}$, where each $|\mathcal{R}_i^{(k)}| = c[\log(d) + \log\log(\frac{1}{\eta})]2^i$, and divide $\mathcal{C}_i = \bigcup_{k=1}^{d} \mathcal{C}_i^{(k)}$ uniformly at random into ℓ even parts $\mathcal{C}_i^{(1)}, \mathcal{C}_i^{(2)}, \ldots, \mathcal{C}_i^{(\ell)}$, where each $|\mathcal{C}_i^{(k)}| = c[\log(d) + \log\log(\frac{1}{\eta})]/(2^i\eta)$ for every k. We note that $\{\mathcal{R}_i^{(k)}\}_k$ (and $\{\mathcal{C}_i^{(k)}\}_k$) are independent of each other. It follows that the event in Lemma 3.1 holds with probability at least $1 - \frac{1}{\mathrm{poly}(d\log(1/\eta))}$. By a union bound over all $\ell^2 = \Theta(d^2\log^2(\frac{1}{\eta}))$ possible choices of $\{\mathcal{R}_i^{(k)}\} \times \{\mathcal{C}_i^{(k)}\}$ and Claim 1, with probability at least $1 - 1/\mathsf{poly}(d\log(\frac{1}{\epsilon}))$, Algorithm 1 answers correctly, when $\mathbf{A}$ is ϵ-far from having rank d. □

3.2 Hardness Results

Our positive result in Theorem 3.1 is supplemented by several negative results over various fields.

3.3 Sampling Lower Bound over Finite Field

We first provide a hardness result for the rank testing problem in the sampling model over any finite field, which shows that the sample complexity in Theorem 3.1 is tight in this case.

According to Yao's minimax principle, it suffices to provide a distribution on $n \times n$ input matrices $\mathbf{A}$ for which any deterministic testing algorithm fails with significant probability over the choice of $\mathbf{A}$. Before proceeding, we first state a hardness result that we want to reduce from.

Algorithm 2 Decomposing edges E

```
Input: A bipartite graph $G = (L \cup R, E)$.
Output: Partition of $E = E_1 \cup \cdots \cup E_t$ and the set
    of pivot nodes $\{w_t\}$.
$t \leftarrow 0$.
while $E \neq \emptyset$ do
    Find $v$ such that $1 \le \mathsf{deg}(v) \le \gamma d$.
    $t \rightarrow t + 1$.
    $E_t \leftarrow$ edges between $v$ and all its neighbours.
    $w_t \leftarrow v$.
    $E \leftarrow E \setminus E_t$.
return $E = E_1 \cup \cdots \cup E_t$ and $\{w_t\}$.
```

Lemma 3.2. *Let $G = (L \cup R, E)$ be a bipartite graph such that $|L| = |R| = n$ and $|E| < \gamma^2 d^2$ for $d \le n/\gamma$. Then Algorithm 2 returns a partition $E = E_1 \cup E_2 \cup \cdots \cup E_t$, where $t \le \gamma^2 d^2$ and $|E_i| \le \gamma d$ for all i.*

Proof. We first show that Algorithm 2 can be executed correctly, that is, whenever $E \neq \emptyset$ there always exists v such that $1 \le \mathsf{deg}(v) \le \gamma d$. We note that $1 \le \mathsf{deg}(v)$ is obvious because $E \neq \emptyset$. If all vertices with non-zero degree have degree at least γd, the total number of edges would be at least $\gamma dn \ge d^2\gamma^2$, contradicting our assumption on the size of E. When the algorithm terminates, it is clear that each E_i generates at most γd edges and the E_i's are disjoint and so $t \le \gamma^2 d^2$.□

Lemma 3.3. *Suppose that there are t groups of (fixed) vectors $\{\mathbf{v}_1^{(k)}, ..., \mathbf{v}_{s_k}^{(k)}\}_{k\in[t]} \subset \mathbb{F}^d$ such that the vectors in each group are linearly independent (denoted by $\perp$). Let $\mathbf{w}_1, ..., \mathbf{w}_r$ be random vectors in $\mathbb{F}^d$ such that each $\mathbf{w}_i$ is chosen uniformly at random from some set $\mathcal{S}_i \subseteq \mathbb{F}^d$ with $|\mathcal{S}_i| \ge |\mathbb{F}|^{(1-\gamma)d}$. Let $s = \max_k s_k$. When $s + r \le \gamma d$ for all k and $t \le \gamma^2 d^2$, it holds that*

$$\Pr_{\mathbf{w}_1,\ldots,\mathbf{w}_r}\left\{\mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \perp \cdots \perp \mathbf{w}_r, \forall k \in [t]\right\} \ge 1 - \frac{\gamma^3 d^3}{|\mathbb{F}|^{(1-2\gamma)d}}.$$

Proof. For fixed $\mathbf{w}_1, \ldots, \mathbf{w}_{i-1}$ such that $\mathbf{v}_1^{(k)}, \ldots, \mathbf{v}_{s_k}^{(k)}, \mathbf{w}_1, \ldots, \mathbf{w}_{i-1}$ are linearly independent for all $k \in [t]$, the probability that $\mathbf{w}_i \in \mathcal{S}_i$ is linearly independent of $\mathbf{v}_1^{(k)}, \ldots, \mathbf{v}_{s_k}^{(k)}, \mathbf{w}_1, \ldots, \mathbf{w}_{i-1}$ for all $k \in [t]$ is at least

$$1 - \frac{t|\mathbb{F}|^{s_k+i-1}}{|\mathcal{S}_i|} \ge 1 - \frac{t|\mathbb{F}|^{s_k+i-1}}{|\mathbb{F}|^{(1-\gamma)d}}$$
$$= 1 - \frac{t}{|\mathbb{F}|^{(1-\gamma)d-(s_k+i-1)}} \ge 1 - \frac{t}{|\mathbb{F}|^{(1-\gamma)d-(s+i-1)}}.$$

Therefore, for all $k \in [t]$ with $t \le \gamma^2 d^2$, we have

$$\Pr_{\mathbf{w}_1,\ldots,\mathbf{w}_r}\{\mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \perp \ldots \perp \mathbf{w}_r, \forall k\}$$
$$= \prod_{i=2}^{r} \Pr_{\mathbf{w}_i}\left\{\mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \perp \cdots \perp \mathbf{w}_i, \forall k \middle| \mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \perp \cdots \perp \mathbf{w}_{i-1}, \forall k\right\}$$
$$\times \Pr_{\mathbf{w}_1}\{\mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \text{ for all } k\}$$
$$\ge \prod_{i=1}^{r}\left(1 - \frac{t}{|\mathbb{F}|^{(1-\gamma)d-(s+i-1)}}\right)$$
$$\ge \prod_{i=1}^{r}\left(1 - \frac{\gamma^2 d^2}{|\mathbb{F}|^{(1-2\gamma)d}}\right) \quad (s+i-1 \le \gamma d \text{ and } t \le \gamma^2 d^2)$$
$$\ge 1 - \frac{r\gamma^2 d^2}{|\mathbb{F}|^{(1-2\gamma)d}} \quad ((1-x)^t \ge 1 - tx \text{ for } x \in (0,1))$$

[2] In the number of parts $d + d\log(\frac{1}{\eta})$, the first term follows from the operation of augmenting 1×1 submatrix to $d \times d$. The second term follows from moving the submatrix towards the upper left corner (from the lower-right corner in the worst case).

$$\geq 1 - \frac{\gamma^3 d^3}{|\mathbb{F}|^{(1-2\gamma)d}}. \quad (r \leq \gamma d).$$

□

When $|\mathcal{S}_i| = |\mathbb{F}|^{d-d_i}$ for $d_i \leq \gamma d$, it follows from Lemma 3.3 that the number of choices of the event

$$\begin{aligned}
&|\{(\mathbf{w}_1, \ldots, \mathbf{w}_r) \in \prod_{i=1}^{r} \mathcal{S}_i : \mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \\
&\qquad\qquad \perp \mathbf{w}_1 \perp \cdots \perp \mathbf{w}_r, \forall k\}| \\
&= \Pr_{\mathbf{w}_1,\ldots,\mathbf{w}_r} \left\{ \mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \perp \cdots \perp \mathbf{w}_r, \forall k \right\} \prod_{i=1}^{r} |\mathcal{S}_i| \\
&\geq \left(1 - \frac{\gamma^3 d^3}{|\mathbb{F}|^{(1-2\gamma)d}}\right) \cdot \prod_{i=1}^{r} |\mathcal{S}_i| \quad \text{(Lemma 3.3)} \\
&= \left(1 - \frac{\gamma^3 d^3}{|\mathbb{F}|^{(1-2\gamma)d}}\right) |\mathbb{F}|^{rd - \sum_{i=1}^{r} d_i}. \\
&\quad \text{(recall that } |\mathcal{S}_i| = |\mathbb{F}|^{d-d_i}\text{)}
\end{aligned} \tag{3.7}$$

Based on this result, we have the following lemma.

LEMMA 3.4. *Let $\mathbf{U}, \mathbf{V} \sim \mathcal{U}_{\mathbb{F}}(n, d)$, where $\mathcal{U}_{\mathbb{F}}(m, n)$ represents $m \times n$ i.i.d. uniform matrix over a finite field $\mathbb{F}$. Denote by $\mathcal{S}$ any subset of $[n] \times [n]$ such that $|\mathcal{S}| < \gamma^2 d^2$ for $\gamma \in (0, 1/4)$ and $d \leq n/\gamma$. It holds that for any $\mathbf{x} \in \mathbb{F}^{|\mathcal{S}|}$,*

$$\Pr[(\mathbf{U}\mathbf{V}^T)|_{\mathcal{S}} = \mathbf{x}] - \frac{1}{|\mathbb{F}|^{|\mathcal{S}|}} \geq -\frac{\gamma^5 d^5}{|\mathbb{F}|^{(1-2\gamma)d + |\mathcal{S}|}}.$$

Proof. Consider a bipartite graph $G = (L \cup R, E)$ where $|L| = |R| = n$ and $(i, j) \in E$ if and only if $(i, j) \in \mathcal{S}$. We run Algorithm 2 on graph G. By Lemma 3.2, we obtain a sequence of edge sets $E_1 \ldots, E_t$ with $w_1, \ldots, w_t$ (called *pivot nodes*), such that

1. $\{E_i, \ldots, E_t\}$ forms a partition of E;
2. $|E_i| \leq \gamma d$ for all i.

Since there is a one-by-one correspondence between the edges and the entries in $\mathcal{S}$, we will not distinguish edges and entries in the rest of the proof.

We associate each node v of G with an affine space $H_v \subseteq \mathbb{F}^d$ and a random vector $\mathbf{x}_v \in H_v$ as in Algorithm 3. Basically, Algorithm 3 first assigns the non-pivot nodes (to determine the affine subspace H_{w_i}) from the E_t down to the E_1), and in the end assigns all unassigned pivot nodes.

In the following argument, we number the for-loop iterations in Algorithm 3 backwards, i.e., the for-loop starts with the t-th iteration and goes down to the first iteration. In the i-th iteration, let r_i denote the number of nodes $v_j^{(i)}$ that are unassigned at the runtime of Line 6 and let $\#\mathcal{E}_i$ denote the number of good choices (which do not trigger abortion) of Step 8 over all r_i nodes to be assigned. Let $\#\mathcal{G}$ be the number of possible choices of Step 13 of Algorithm 3 and $s_0 = |\mathcal{S}_0|$ be the number of assigned pivot nodes by Step 13. Note that by the construction of Algorithm 2, the non-pivot nodes of E_i cannot be the pivot nodes of E_j for $j < i$. So Algorithm 3, if terminated successfully, can find an assignment such that $(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}$. We now lower bound the success probability.

Let $d_j^{(i)} = d - \mathsf{dim}(H_{v_j^{(i)}})$, which is either 0 or $|E_k|$ for some $k > i$. For any given realization $\mathbf{x}$, we have

$$\begin{aligned}
&\Pr\{(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}\} \\
&\geq \frac{\#\mathcal{E}_t \cdot \#\mathcal{E}_{t-1} \cdots \#\mathcal{E}_1}{|\mathbb{F}|^{d(r_t + \cdots + r_1)}} \frac{\#\mathcal{G}}{|\mathbb{F}|^{ds_0}} \text{ (by definition of } \#\mathcal{E}_i\text{)} \\
&\geq \prod_{i=1}^{t} \frac{1}{|\mathbb{F}|^{d_1^{(i)} + \cdots + d_{r_i}^{(i)}}} \left(1 - \frac{\gamma^3 d^3}{|\mathbb{F}|^{(1-2\gamma)d}}\right)^t \frac{\#\mathcal{G}}{|\mathbb{F}|^{ds_0}} \text{ (by (3.7))} \\
&\geq \frac{1}{|\mathbb{F}|^{\sum_{i=1}^{t} \sum_{j=1}^{r_i} d_j^{(i)}}} \left(1 - \frac{\gamma^5 d^5}{|\mathbb{F}|^{(1-2\gamma)d}}\right) \frac{\#\mathcal{G}}{|\mathbb{F}|^{d \times s_0}} \\
&= \frac{1}{|\mathbb{F}|^{\sum_{i=1}^{t} \sum_{j=1}^{r_i} d_j^{(i)}}} \left(1 - \frac{\gamma^5 d^5}{|\mathbb{F}|^{(1-2\gamma)d}}\right) \frac{1}{|\mathbb{F}|^{\sum_{s \in \mathcal{S}_0} |E_s|}} \\
&\geq \frac{1}{|\mathbb{F}|^{|E_1| + \cdots + |E_t|}} \left(1 - \frac{\gamma^5 d^5}{|\mathbb{F}|^{(1-2\gamma)d}}\right),
\end{aligned}$$

where the second to last inequality follows from the definition of $\#\mathcal{G}$ and the last inequality follows from that $|E_1| + \cdots + |E_t| = \sum_{j=1}^{t} \sum_{i=1}^{r_j} d_i^{(j)} + \sum_{s \in \mathcal{S}_0} |E_s|$ since every pivot and non-pivot node must be assigned exactly once by Algorithm 3 upon successful termination. (Recall that $d_i^{(j)}$ is either equal to 0 when $v_j^{(i)}$ is non-pivotal, or equal to $|E_k|$ when $v_j^{(i)} = w_k$.) □

Denote by $\mathcal{S} \subset [n] \times [n]$ a set of indices of an $n \times n$ matrix. For any distribution $\mathcal{L}$ over $\mathbb{F}^{n \times n}$, define $\mathcal{L}(\mathcal{S})$ on $\mathbb{F}^{|\mathcal{S}|}$ as the marginal distribution of $\mathcal{L}$ on the entries of $\mathcal{S}$, namely,

$$(\mathbf{X}_{p_1,q_1}, \mathbf{X}_{p_2,q_2}, \ldots, \mathbf{X}_{p_{|\mathcal{S}|},q_{|\mathcal{S}|}}) \sim \mathcal{L}(\mathcal{S}), \qquad \mathbf{X} \sim \mathcal{L}.$$

Now we are ready to show a lower bound of robust testing problem over any finite field.

THEOREM 3.2. *Suppose that $\mathbb{F}$ is a finite field and $\gamma \in (0, 1/4)$ is an absolute constant. Let $\mathbf{U}, \mathbf{V} \sim \mathcal{U}_{\mathbb{F}}(n, d)$ and $\mathbf{W} \sim \mathcal{U}_{\mathbb{F}}(n, n)$, where $\mathcal{U}_{\mathbb{F}}(m, n)$ represents $m \times n$ i.i.d. uniform matrix over a finite field $\mathbb{F}$.*

Algorithm 3 Path for assigning subspace H_v and random vector $\mathbf{x}_v$ to each node v

```
Input: Bipartite graph $G = (L \cup R, E)$, partition $E = E_1 \cup \cdots \cup E_t$ and pivot nodes $\{w_t\}$ by Algorithm 2,
    observed entries $\mathbf{x}|_E$.
Output: An affine space $H_v$ of vectors for every node $v$ and a vector $\mathbf{x}_v \in H_v$ for every node $v$.
$H_v \leftarrow \mathbb{F}^d$ for all $v$.
Set all nodes $v$ unassigned.
for $i \leftarrow t$ down to 1 do
    Let $v_1^{(i)}, ..., v_{|E_i|}^{(i)}$ be the non-pivot nodes in $E_i$ (i.e., the edges in $E_i$ are $(w_i, v_j^{(i)})$).
    for $j \leftarrow 1$ to $|E_i|$ do
        if $v_j^{(i)}$ is unassigned then
            $W_{v_j^{(i)}} \leftarrow H_{v_j^{(i)}} \setminus \bigcup_{k \le i : v_j^{(i)} \neq w_k} \mathrm{span}\{\mathbf{x}_{v_1}^{(i)}, \ldots, \mathbf{x}_{v_{j-1}}^{(i)},$ previously assigned non-pivot nodes in $E_k\}$.
            Choose $w_{v_j^{(i)}}$ uniformly at random from $H_{v_j^{(i)}}$.
            if $w_{v_j^{(i)}} \notin W_{v_j^{(i)}}$ then
                abort.
            Set $v_j^{(i)}$ to be assigned.
    Let $H_{w_i}$ be the solution set to the linear system (w.r.t. $\mathbf{x}_{w_i}$): $\mathbf{x}_{w_i}^\top [\mathbf{x}_{v_1}^{(i)}, \cdots, \mathbf{x}_{v_{|E_i|}}^{(i)}] = (\mathbf{x}|_{E_i})^\top$.
Choose $\mathbf{x}_{w_s}$ uniformly from $H_{w_s}$ of dimension $d - |E_s|$ for all $s \in \mathcal{S}_0 = \{p \in [t] \mid w_p$ is unassigned$\}$.
return $\{H_v\}$ and $\{\mathbf{x}_v\}$.
```

Consider two distributions $\mathcal{L}_1$ and $\mathcal{L}_2$ over $\mathbb{F}^{n \times n}$ defined by $\mathbf{U}\mathbf{V}^\top$ and $\mathbf{W}$, respectively. Let $\mathcal{S} \subset [n] \times [n]$. When $|\mathcal{S}| < \gamma^2 d^2$, it holds that

$$d_{TV}(\mathcal{L}_1(\mathcal{S}), \mathcal{L}_2(\mathcal{S})) \le C d^5 |\mathbb{F}|^{-cd},$$

where $C, c > 0$ are constants depending on γ, and $d_{TV}(\cdot, \cdot)$ represents the total variation distance between two distributions.

Proof. Let

$$\mathcal{X} = \left\{ \mathbf{x} \in \mathbb{F}^{|\mathcal{S}|} \;\middle|\; \Pr\left[(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}\right] < \frac{1}{|\mathbb{F}|^{|\mathcal{S}|}} \right\}.$$

It follows from the definition of total variation distance that

$$\begin{aligned} d_{TV}(\mathcal{L}_1(\mathcal{S}), \mathcal{L}_2(\mathcal{S})) &= \sum_{\mathbf{x} \in \mathcal{X}} \left[\frac{1}{|\mathbb{F}|^{|\mathcal{S}|}} - \Pr[(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}] \right] \\ &\le \sum_{\mathbf{x} \in \mathcal{X}} \frac{\gamma^5 d^5}{|\mathbb{F}|^{(1-2\gamma)d}} \frac{1}{|\mathbb{F}|^{|\mathcal{S}|}} \le \frac{\gamma^5 d^5}{|\mathbb{F}|^{(1-2\gamma)d}}, \end{aligned}$$

where the last inequality holds since $|\mathcal{X}| \le |\mathbb{F}|^{|\mathcal{S}|}$. □

Based on the above theorem, we have the following lower bound for the rank testing problem over finite field.

Theorem 3.3. (Sampling lower bound over finite field). *Let $d \le \sqrt{\epsilon} n$. Any non-adaptive algorithm for the rank testing problem over any finite field $\mathbb{F}$ requires $\Omega(d^2/\epsilon)$ queries.*

Proof. We first show that for constant ϵ, any non-adaptive algorithm for the rank testing problem over finite field $\mathbb{F}$ requires $\Omega(d^2)$ queries. Note that $\mathbf{W} \sim \mathcal{U}_{\mathbb{F}}(n, n)$ is ϵ-far from having rank less than d. It follows immediately from the preceding theorem that any algorithm which solves the matrix rank testing problem over a finite field must read $\Omega(d^2)$ entries; otherwise when d is large enough, it will hold that $d_{TV}(\mathcal{L}_1(\mathcal{S}), \mathcal{L}_2(\mathcal{S})) < 1/4$, contradicting the correctness of the algorithm on distinguishing $\mathcal{L}_1$ from $\mathcal{L}_2$.

We now prove the case for arbitrary ϵ. Denote by $\mathbf{A}$ and $\mathbf{B}$ the two hard instances in Theorem 3.2. We construct two hard instances $\mathbf{C}$ and $\mathbf{D}$ by uniformly at random planting the above-mentioned hard instances $\mathbf{A}$ and $\mathbf{B}$ of dimension $\sqrt{\epsilon} n \times \sqrt{\epsilon} n$, respectively, and padding zeros everywhere else. Note that $\mathbf{D}$ being ϵ-far from rank d is equivalent to $\mathbf{B}$ being constant-far from rank d. Suppose that we can request cd^2/ϵ queries with a small absolute constant c to distinguish the ranks of the hard instances $\mathbf{C}$ and $\mathbf{D}$, then in expectation (and with high probability by a Markov bound) we can request cd^2 queries of the hard instances $\mathbf{A}$ and $\mathbf{B}$ to distinguish their ranks, which leads to a contradiction. □

3.4 Other Lower Bounds for Rank Testing

Our second lower bound studies the hardness of the rank testing problem in the sampling model over $\mathbb{R}$.

Theorem 3.4. (Sampling lower bound over $\mathbb{R}$). *Let $d \le \sqrt{\epsilon} n$. Any non-adaptive algorithm for the rank* *testing problem in the sampling model over $\mathbb{R}$ requires $\Omega(d^2/\epsilon)$ queries.*

Proof Sketch. Our hard instances are $\mathbf{A}_1 = \mathbf{U}\mathbf{V}^\top$ vs. $\mathbf{A}_2 = \mathbf{U}\mathbf{V}^\top + n^{-14}\mathbf{G}$, where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n\times d}$ and $\mathbf{G} \in \mathbb{R}^{n\times n}$ have i.i.d. standard Gaussian entries. We observe that $\mathbf{A}_1$ is of rank d while $\mathbf{A}_2$ is constant-far from having rank d by the *rigidity* of the Gaussian random matrix. On the other hand, the total variation distance between the induced distributions on the samples of $\mathbf{A}_1$ and $\mathbf{A}_2$ is a small constant whenever we read fewer than $\mathcal{O}(d^2)$ entries.

The desired $\Omega(d^2/\epsilon)$ lower bound follows from uniformly at random planting the above hard instances $\mathbf{A}_1$ and $\mathbf{A}_2$ (replacing n above with $\sqrt{\epsilon}n$) into a $n \times n$ matrix, and padding zeros everywhere else. □

It can be more challenging to study the hardness of testing with a more powerful sensing oracle. We have the following result for the rank testing problem in the sensing model over any finite field.

Theorem 3.5. (Sensing lower bound over $\mathsf{GF}(p)$). *Any non-adaptive algorithm for the rank testing problem in the sensing model over $\mathsf{GF}(p)$ requires $\Omega(d^2/\log p)$ queries, where $\mathsf{GF}(p)$ denotes the Galois field of prime order p.*

Proof Sketch. Denote by $\mathsf{Matching}_{n,k,\epsilon}$ the *k-player simultaneous communication* problem of estimating the size of a maximum matching up to a factor of $(1 \pm \epsilon)$, where the edges of an n-vertex input graph are partitioned across the k players and the referee. For our purpose, we reduce from the problem of $\mathsf{Matching}_{n,k,\epsilon}$ to our problem of *rank testing*: we use the adjacency matrices of the hard instances of $\mathsf{Matching}_{n,k,\epsilon}$ [2] as our hard instances. □

4 Stable Rank, Schatten-p Norm, Entropy: A New Testing Framework

In this section, we study the problem of non-adaptively testing numerical properties of real-valued matrices. They can be studied under a unified framework in this section.

Roughly, our analytical framework reduces the testing problem to a sequence of estimation problems *without involving* $\mathsf{poly}(n)$ *in the sample complexity.* Our framework consists of two levels of estimation: (1) a constant-factor approximation to the statistic X of interest (e.g., stable rank), and (2) a more accurate $(1 \pm \tau)$-approximation to X.

4.1 Stable Rank Testing

We present our algorithm for stable rank testing in Algorithm 4. The sample complexity in Algorithm 4 depends on the bottleneck Step 10, which involves implementing a $(1\pm\epsilon/d^{1/4})$-approximation to the largest singular value. For the sampling and sensing models, we establish the following guarantees for the estimator.

Theorem 4.1. ($(1 \pm \tau)$-approximation to operator norm). *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a stable-rank-$\mathcal{O}(d)$ matrix such that $\|\mathbf{A}\|_F^2 = \Omega(\tau n^2)$ and $|\mathbf{A}_{i,j}| \le 1$ for all i, j. There exist (a) a non-adaptive sampling algorithm of query complexity $\mathcal{O}(d^2\log^2(n)/\tau^4)$ and (b) a non-adaptive sensing algorithm of query complexity $\mathcal{O}(\max\{\log^2(d\log(n)/\tau), d^2\log(n)\}/\tau^2)$, which both output a $(1 \pm \tau)$-approximation to the operator norm $\|\mathbf{A}\|$ with probability at least 0.9. The sensing complexity in (b) is optimal up to a polylogarithmic factor.*

Result (a) is built upon a non-uniform sampling of matrix rows and columns with probability proportional to their squared ℓ_2 norms, and the estimations of these squared ℓ_2 norms (see Algorithm 5). Result (b) is built upon a cycle estimator, akin to that in [23], for the Schatten-p norm with $p = \widetilde{\mathcal{O}}(1/\epsilon)$ (see Algorithm 7). We note that the sample complexity in Theorem 4.1 does not depend on n polynomially.

Now we are ready to prove the correctness of Algorithm 4.

Theorem 4.2. (Upper bounds). *Suppose that $d = \Omega((1/\epsilon)^{1/3})$. Then Algorithm 4 is a correct algorithm for the stable rank testing problem with failure probability at most 1/3 under both (a) the sampling model, with $\mathcal{O}(d^3/\epsilon^4 \cdot \log^2 n)$ sampled entries, and (b) the sensing model, with $\mathcal{O}(d^{2.5}/\epsilon^2 \cdot \log n)$ sensing queries.*

Proof Sketch. Denote by $\mathsf{srank}(\mathbf{A})$ the stable rank of $\mathbf{A}$. Our goal is to distinguish $\mathsf{H0}$: "$\mathsf{srank}(\mathbf{A}) \le d$" from $\mathsf{H1}$: "$\mathbf{A}$ is ϵ/d-far from $\mathsf{srank}(\mathbf{A}) \le d$". When $\mathbf{A} \in \mathsf{H1}$, we claim that $\|\mathbf{A}\|_F^2 \ge \epsilon n^2(1-\frac{1}{d})$. Otherwise, replacing any $\frac{\epsilon n}{d}$ rows of $\mathbf{A}$ each with an all-one row vector $\mathbf{1}^\top$ results in a new matrix $\mathbf{B}$ such that $\|\mathbf{B}\|^2 \ge \frac{\epsilon n^2}{d}$ and $\|\mathbf{B}\|_F^2 = \|\mathbf{A}\|_F^2 + \frac{\epsilon n^2}{d} \le \epsilon n^2(1 - \frac{1}{d}) + \frac{\epsilon n^2}{d} = \epsilon n^2$, leading to $\mathsf{srank}(\mathbf{B}) \le d$, a contradiction. Sample q_0 entries from $\mathbf{A}$ and stack them as vector $\mathbf{y}$. The resulting estimator $X = \frac{n^2}{q_0^2}\|\mathbf{y}\|_2^2$ is a $(1 \pm \tau)$-approximation to the squared Frobenius norm with $\tau = \epsilon/d^{1/4}$. So the algorithm is correct in Step 4.

Case (i). $\mathsf{srank}(\mathbf{A}) > c_1 d$ **when A is far from** $\mathsf{srank}(\mathbf{A}) \le d$**.** We first discuss the case when $\mathbf{A}$ is far from $\mathsf{srank}(\mathbf{A}) \le d$. Rudelson and Vershynin [32] show that uniformly sampling q rows of a matrix gives a subsampled matrix $\mathbf{A}_{\mathsf{row}}$ such that $\|\mathbf{A}_{\mathsf{row}}\| \lesssim \sqrt{\frac{q}{n}}\|\mathbf{A}\| + \sqrt{\log q}\|\mathbf{A}\|_{(n/q)}$, where $\|\mathbf{A}\|_{(n/q)}$ is the average of the n/q largest ℓ_2 norms of the columns of $\mathbf{A}$. Applying this result once for row sampling

Algorithm 4 Algorithm for stable rank testing under sampling/sensing model

```
Uniformly sample $q_0 = \mathcal{O}(\frac{\sqrt{d}}{\epsilon^{2.5}})$ entries $\mathbf{A}$, forming vector $\mathbf{y}$.
$X \leftarrow \frac{n^2}{q_0}\|\mathbf{y}\|_2^2$.                                  ▷ $X$ is an estimator of $\|\mathbf{A}\|_F^2$.
if $X \le \frac{9}{10}(1-\frac{1}{d})\epsilon n^2$ then
    Output "stable rank $\le d$".
else
    Uniformly sample a $q \times q$ submatrix $\widetilde{\mathbf{A}}'$ with $q = \mathcal{O}(\frac{d \log n}{\epsilon})$.
    if $\|\widetilde{\mathbf{A}}'\| \le C_0 \frac{\sqrt{X}}{\sqrt{c_1 d}} \frac{q}{n}$ then     ▷ $\|\widetilde{\mathbf{A}}'\|$ is a (1st-level) constant-approximation to the operator norm.
        Output "$\epsilon/d$-far from being stable rank $\le d$".
    else
        Implement an estimator $Z$ with $(1 \pm \epsilon/d^{1/4})$-approximation to the operator norm.
        if $Z^2 \ge \frac{X}{d}$ then                     ▷ $Z$ is the refined (2nd-level) estimator to the operator norm.
            Output "stable rank $\le d$".
        else
            Output "$\epsilon/d$-far from being stable rank $\le d$".
```

and once for column sampling, we obtain a submatrix $\mathbf{A}'$ such that $\|\widetilde{\mathbf{A}}'\| \lesssim \frac{q}{n}\frac{\|\mathbf{A}\|_F}{\sqrt{c_1 d}}$. On the other hand, Magdon-Ismail [28] shows that when $q = \mathcal{O}(\frac{d \log n}{\epsilon})$, uniformly sampling q rows of a stable-rank-$\mathcal{O}(d)$ matrix gives a submatrix $\mathbf{A}_{\mathsf{row}}$ such that $\|\mathbf{A}_{\mathsf{row}}\| \gtrsim \sqrt{\frac{q}{n}}\|\mathbf{A}\|$. So when $\mathsf{srank}(\mathbf{A}) \le d$, we have with high probability that $\|\widetilde{\mathbf{A}}'\| \gtrsim \frac{q}{n}\frac{\|\mathbf{A}\|_F}{\sqrt{d}}$. Therefore there is a constant-factor gap in stable rank between the two cases when c_1 is large enough, and we can thus distinguish H0 from H1 by checking $\|\widetilde{\mathbf{A}}'\|$ in case (i).

Case (ii). $\mathsf{srank}(\mathbf{A}) \le c_1 d$ when $\mathbf{A}$ is far from $\mathsf{srank}(\mathbf{A}) \le d$. We show that when $\mathbf{A} \in$ H1, $\mathsf{srank}(\mathbf{A}) > (1 + \Theta(\epsilon/d^{1/4}))d$. This can be demonstrated by replacing the $\frac{\epsilon n}{d}$ least correlated columns/rows (w.r.t. the leading left/right singular vector) with the signs of the top left/right singular vector. This forms a matrix $\mathbf{B}$. By the definition of "ϵ/d-far", we have that $\mathsf{srank}(\mathbf{B}) > d$. Expressing $\mathsf{srank}(\mathbf{B})$ in terms of $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|$, we can show that $\mathsf{srank}(\mathbf{A}) > (1 + \Theta(\epsilon/d^{1/4}))d$. Finally, we note that $\mathsf{srank}(\mathbf{A}) = \mathcal{O}(d)$ for both H0 and H1 in case (ii). Applying Theorem 4.1 with $\tau = \Theta(\epsilon/d^{1/4})$ gives the desired result. □

The positive results of our framework are complemented by a lower bound to formalize the hardness of the stable rank testing problem *in both sampling and sensing models.*

Theorem 4.3. (Lower bounds). *Let $\epsilon \in (0, 1/3)$ and $d \ge 4$. For any $\mathbf{A} \in \mathbb{R}^{(d/\epsilon^2)\times d}$ with all entries bounded by 1 in absolute value, any linear sketching algorithm that distinguishes "the stable rank of $\mathbf{A}$ is at most d_0" from "$\mathbf{A}$ is ϵ_0/d_0-far from having stable rank $\le d_0$" with error probability at most 1/6 requires $\Omega(d^2/(\epsilon^2 \log(d/\epsilon)))$ sketch length, where $d_0 = d/(1 + \Theta(\epsilon))$ and $\epsilon_0 = \Theta(\epsilon/\log^2(d/\epsilon))$.*

Proof Sketch. Our hard instances are $\mathbf{A}_1 = \frac{C}{\log(d/\epsilon)}\mathbf{G}$ vs. $\mathbf{A}_2 = \frac{C}{\log(d/\epsilon)}(\mathbf{G}_0 + 3\sqrt{\frac{\epsilon}{d}}\mathbf{u}\mathbf{v}^\top)$, where $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{n\times 1}$ and $\mathbf{G} \in \mathbb{R}^{n\times n}$ have i.i.d. standard Gaussian entries, and the scaling term $\frac{C}{\log(d/\epsilon)}$ guarantees that the entries of $\mathbf{A}_1$ and $\mathbf{A}_2$ are bounded by 1 with high probability. Observe that $\mathbf{A}_2$ is of stable rank d_0 while $\mathbf{A}_2$ is ϵ_0/d_0-far from having stable rank d_0 because of its rigidity. On the other hand, the total variation distance between distributions $\mathbf{A}_1$ and $\mathbf{A}_2$ is at most 1/10 if the sketch length is shorter than $\mathcal{O}(d^2/(\epsilon^2 \log(d/\epsilon)))$. □

4.2 Schatten-p Norm Testing

In this section, we develop our theory for Schatten-p norm testing based on our framework. We study the problem in the bounded entry model, where every entry is bounded by 1 in absolute value. Our goal is to distinguish "H0: $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ is at least cn^p for $p > 2$ or at least $cn^{1+1/p}$ for $p < 2$" from "H1: at least an ϵ-fraction of entries of $\mathbf{A}$ should be modified in order to have that property". We remark that n^p for $p > 2$ and $n^{1+1/p}$ for $p \in [1, 2)$ are the largest possible values of $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ conditioned on $\|\mathbf{A}\|_F^2 \le n^2$, and are thus the largest possible values of $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ for $\mathbf{A}$ in the bounded entry model.

Our positive result shows that for $p > 2$, there is an algorithm in the sampling model which correctly solves the Schatten-p norm testing problem with high probability, and the sampling complexity depends on n only logarithmically.

Theorem 4.4. (Upper bounds for $p > 2$). *Let $p > 2$ be a constant. There exist constants $c =$*

$c(p)$, $C = C(p)$ and $\epsilon_0 = \epsilon(p)$ such that for any $\epsilon \in [C/n, \epsilon_0]$, there is a randomized algorithm which reads $\mathcal{O}\left(\epsilon^{-4p/(p-2)} \log^2 n\right)$ entries and with probability ≥ 0.99 solves the Schatten-p norm testing problem.

The proof framework for Theorem 4.4 is similar to that of Theorem 4.2.

In contrast, for Schatten-p norm testing for $p < 2$, we have a negative result. We show that any algorithm which distinguishes $\mathsf{H0}$ from $\mathsf{H1}$ correctly must have query complexity linear in n, based on the hard instance for Schatten-p norm estimation in [22].

Theorem 4.5. (Lower bounds for $p \in [1, 2)$). *Let $p \in [1, 2)$ be a constant. There exist constants $c = c(p)$ and $\epsilon_0 = \epsilon_0(p)$ such that for any $\epsilon \leq \epsilon_0$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$, any non-adaptive algorithm that correctly tests $\mathsf{H0}$ against $\mathsf{H1}$ with probability at least 0.99 must make $\Omega(n)$ sensing queries.*

4.3 Entropy Testing

We consider the problem of testing the matrix entropy $H(\mathbf{A})$, defined in (2.3), for matrices $\mathbf{A}$ in the bounded entry model. Our goal is to distinguish "$\mathsf{H0}. : H(\mathbf{A}) \leq \log n + \log\log\log n - c$" and "$\mathbf{A}$ is $\frac{\epsilon}{\log n \log\log n}$-far from having entropy at most $\log n + \log\log\log n - c$" for some absolute constant c. We show a lower bound of $\Omega(n)$ queries.

Theorem 4.6. *There exist absolute constants $c > 0$ and $\epsilon_0 > 0$ such that for any $\epsilon \leq \epsilon_0$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$, any non-adaptive algorithm that correctly tests $\mathsf{H0}$ against $\mathsf{H1}$ with probability at least 0.99 must make $\Omega(n)$ queries (i.e., the sketch size is $\Omega(n)$).*

Acknowledgements. Part of this work was done while D. W. and H. Z. were visiting the Simons Institute for the Theory of Computing. We thank Sepehr Assadi, Yury Polyanskiy, Ran Raz, Avishay Tal, Zhengyu Wang, and Yihong Wu for useful discussions. This work was supported in part by grants NSF CCF-1422910, NSF CCF-1535967, NSF IIS-1618714, Office of Naval Research (ONR) grant N00014-18-1-2562, and a Microsoft Research Fellowship.

References

[1] Noga Alon, Troy Lee, Adi Shraibman, and Santosh Vempala. The approximate rank of a matrix and its algorithmic applications: approximate rank. In *ACM Symposium on Theory of Computing*, pages 675–684, 2013.

[2] Sepehr Assadi, Sanjeev Khanna, and Yang Li. On estimating maximum matching size in graph streams. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1723–1742, 2017.

[3] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Annual Conference on Learning Theory*, pages 152–192, 2016.

[4] Maria-Florina Balcan and Nicholas JA Harvey. Learning submodular functions. In *ACM Symposium on Theory of Computing*, pages 793–802, 2011.

[5] Maria-Florina Balcan, Yingyu Liang, David P. Woodruff, and Hongyang Zhang. Matrix completion and related problems via strong duality. In *Innovations in Theoretical Computer Science*, volume 94, 2018.

[6] Maria-Florina Balcan and Hongyang Zhang. Noise-tolerant life-long matrix completion via adaptive sampling. In *Advances in Neural Information Processing Systems*, pages 2955–2963, 2016.

[7] Siddharth Barman, Arnab Bhattacharyya, and Suprovat Ghoshal. Testing sparsity over known and unknown bases. *arXiv preprint arXiv:1608.01275*, 2016.

[8] Thierry Bouwmans, Sajid Javed, Hongyang Zhang, Zhouchen Lin, and Ricardo Otazo. On the applications of robust PCA in image and video processing. *Proceedings of the IEEE*, 106(8):1427–1457, 2018.

[9] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203, 2014.

[10] Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *ACM Symposium on Theory of Computing*, pages 81–90, 2013.

[11] Constantinos Daskalakis, Ilias Diakonikolas, Rocco A Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1833–1852, 2013.

[12] Amit Deshpande, Madhur Tulsiani, and Nisheeth K Vishnoi. Algorithms and hardness for subspace approximation. In *ACM-SIAM symposium on Discrete Algorithms*, pages 482–496, 2011.

[13] Moritz Hardt. Understanding alternating minimization for matrix completion. In *IEEE Symposium on Foundations of Computer Science*, pages 651–660, 2014.

[14] Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems*, pages 2339–2347, 2012.

[15] Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Annual Conference on Learning Theory*, pages 354–375, 2013.

[16] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *ACM Symposium on Theory of Computing*, pages 665–674, 2013.

[17] Ashish Khetan and Sewoong Oh. Matrix norm

estimation from a few entries. In *Advances in Neural Information Processing Systems*, pages 6424–6433. 2017.

[18] Weihao Kong and Gregory Valiant. Spectrum estimation from samples. *The Annals of Statistics*, 45(5):2218–2247, 2017.

[19] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

[20] Robert Krauthgamer and Ori Sasson. Property testing of data dimensionality. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 18–27, 2003.

[21] Yi Li, Huy L. Nguyen, and David P. Woodruff. On sketching matrix norms and the top singular vector. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1562–1581, 2014.

[22] Yi Li, Huy L. Nguyên, and David P. Woodruff. On approximating matrix norms in a stream. 2017. Submitted.

[23] Yi Li, Zhengyu Wang, and David P. Woodruff. Improved testing of low rank matrices. In *International Conference on Knowledge Discovery and Data Mining*, pages 691–700, 2014.

[24] Yi Li and David P. Woodruff. On approximating functions of the singular values in a stream. In *ACM Symposium on Theory of Computing*, pages 726–739, 2016.

[25] Yi Li and David P. Woodruff. Tight bounds for sketching the operator norm, Schatten norms, and subspace embeddings. In *International Conference on Approximation Algorithms for Combinatorial Optimization Problems, and International Conference on Randomization and Computation*, pages 39:1–39:11, 2016.

[26] Yi Li and David P. Woodruff. Embeddings of Schatten norms with applications to data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 60:1–60:14, 2017.

[27] Zhouchen Lin and Hongyang Zhang. *Low-rank Models in Visual Analysis: Theories, Algorithms, and Applications*. Academic Press, 2017.

[28] Malik Magdon-Ismail. Row sampling for matrix algorithms via a non-commutative Bernstein bound. *arXiv preprint arXiv:1008.0587*, 2010.

[29] Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[30] Vasileios Nakos, Xiaofei Shi, David P Woodruff, and Hongyang Zhang. Improved algorithms for adaptive compressed sensing. In *International Colloquium on Automata, Languages, and Programming*, pages 90:1–90:14, 2018.

[31] Michal Parnas and Dana Ron. Testing metric properties. *Information and Computation*, 187(2):155–195, 2003.

[32] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, 54(4):21, 2007.

[33] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via nonconvex factorization. In *IEEE Symposium on Foundations of Computer Science*, pages 270–289, 2015.

[34] Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[35] Leslie Valiant. Graph-theoretic arguments in low-level complexity. *Mathematical Foundations of Computer Science*, pages 162–176, 1977.

[36] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[37] Hongyang Zhang, Zhouchen Lin, and Chao Zhang. A counterexample for the validity of using nuclear norm as a convex surrogate of rank. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 226–241, 2013.

[38] Hongyang Zhang, Zhouchen Lin, and Chao Zhang. Completing low-rank matrices with corrupted samples from few coefficients in general basis. *IEEE Transactions on Information Theory*, 62(8):4748–4768, 2016.

[39] Hongyang Zhang, Zhouchen Lin, Chao Zhang, and Edward Y Chang. Exact recoverability of robust PCA via outlier pursuit with tight recovery bounds. In *AAAI Conference on Artificial Intelligence*, pages 3143–3149, 2015.

[40] Hongyang Zhang, Zhouchen Lin, Chao Zhang, and Junbin Gao. Robust latent low rank representation for subspace clustering. *Neurocomputing*, 145:369–373, 2014.

[41] Hongyang Zhang, Zhouchen Lin, Chao Zhang, and Junbin Gao. Relations among some low-rank subspace recovery models. *Neural Computation*, 27(9):1915–1950, 2015.

A Other Related Works

Property Testing of Low-Rank Matrices. Krauthgamer and Sasson [20] studied the problem of property testing of data dimensionality, building upon earlier work of Parnas and Ron [31]. They presented algorithms for testing low dimensionality of a set of vectors and for testing whether a matrix is of low rank. Their algorithm achieves $\mathcal{O}(d^2/\epsilon^2)$ non-adaptive samples by uniformly sampling an $\mathcal{O}(d/\epsilon) \times \mathcal{O}(d/\epsilon)$ submatrix. Later Li et al. [23] studied the *adaptive* testing of matrix rank with a sample complexity upper bound $\widetilde{\mathcal{O}}(d^2/\epsilon)$. Despite a large amount of work on the positive results of rank testing, non-trivial negative results in this direction remain absent. Barman et al. [7] studied a slightly different setting of the rank problem by testing whether $\mathsf{H0}$: $\mathsf{rank}(\mathbf{A}) \le d$ or $\mathsf{H1}$: ϵ-far from $\mathsf{rank}(\mathbf{A}) \le 20d/\epsilon^2$ with a different definition of "ϵ-far" in terms of ϵ-approximate rank [1]. The ϵ-approximate rank is defined as the minimum rank over matrices that approximate every entry of $\mathbf{A}$ to within an additive ϵ. In contrast to these works, we provide the first $\widetilde{\mathcal{O}}(d^2/\epsilon)$ sample complexity upper bound for the more traditional rank testing problem without any rank gap between H0 and H1. We complement this positive result with various matching negative results, showing that any algorithm requires at least $\widetilde{\Omega}(d^2/\epsilon)$ samples in order to succeed with constant probability over various fields. We also extend the results to sensing oracles and obtain an $\mathcal{O}(d^2)$ upper bound and an $\widetilde{\Omega}(d^2)$ matching lower bound.

Property Testing of Stable Rank. To the best of our knowledge, this is the first work that studies the stable rank (and the Schatten-p norm) testing problem in the *bounded entry model*. Perhaps the most related work to ours is [23], which studied non-adaptive testing of stable rank in the *bounded row model*. In this model, the rows of $\mathbf{A}$ and of the matrix after change have Euclidean norm at most 1. The algorithm determines if $\mathbf{A}$ has stable rank at most d, or requires changing an ϵ/d-fraction of *rows* to have stable rank at most d. For this problem, Li et al. [23] provided a tight $\Theta(d/\epsilon^2)$ bound. We argue that the bounded entry model is more challenging than the bounded row model, as our restrictions are in fact weaker, which allows for more flexible options of changing the matrix $\mathbf{A}$.

Estimation of Rank. Estimating the matrix rank is a learning version of the rank testing problem. Balcan and Harvey [4] showed that the rank of a subsampled submatrix is highly concentrated around its expectation. Balcan and Zhang [6] proved that uniformly sampling an $\mathcal{O}(\mu d \log d) \times \mathcal{O}(\mu d \log d)$ submatrix suffices to preserve the rank of the original matrix $\mathbf{A}$, under the standard incoherence assumption that the underlying rank-d matrix $\mathbf{A}$ admits a skinny SVD, i.e., $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ satisfies $\max\{\|\mathbf{U}^\top \mathbf{e}_i\|_2^2, \|\mathbf{V}^\top \mathbf{e}_i\|_2^2\} \le \frac{\mu d}{n}$ for all i. Unfortunately, in the worst case the incoherence parameter μ may be as large as $\mathsf{poly}(n)$, e.g., when $\mathbf{A}$ is a sparse matrix. In contrast, we show that it is possible to detect the rank inexpensively without polynomial dependence in n in the sample complexity.

Estimation of Schatten-p Norm. The Schatten-p norm has found many applications in differential privacy [14] and non-convex optimization [5, 12] for $p = 1$, and in numerical linear algebra [29] for $p \in \{2, \infty\}$. The paper of [21] studied the problem of sketching Schatten-p norms for various p under the bilinear sketch and general sketch models. Both the upper bounds and the lower bounds there depend polynomially on n. For even $p \ge 4$, they also proposed the first cycle estimator with a $(1 \pm \tau)$ approximation in the sketching model. More recently, Kong and Valiant [18] applied a similar cycle estimator to approximate the Schatten-p norm of the covariance matrix with computationally efficient algorithms. Khetan and Oh [17] estimated the Schatten-p norm in the sampling model by connecting the cycle estimator with the p-cyclic pseudograph, and showed that when $p \in \{3, 4, 5, 6, 7\}$, the estimator can be calculated in $\mathcal{O}(n^\omega)$ time , where $\omega < 2.373$ is the exponent of matrix multiplication. For the special case of $p = \infty$ (i.e., estimating the largest singular value), to obtain a $(1 \pm \epsilon)$ approximation, one would need to raise p to as large as $\widetilde{\Theta}(1/\epsilon)$. However, the sample complexity in prior work blows up if p goes beyond an absolute constant. Though Magdon-Ismail [28] showed that non-uniform sampling of rows of matrix provides a $(1 \pm \epsilon)$ approximation to the largest singular value with small samples, the sampling probability depends on the unknown ℓ_2 norm of each row. In contrast, we provide the first non-adaptive algorithm to estimate the largest singular value up to $(1 \pm \epsilon)$ relative error with sample complexity $\mathsf{poly}(d/\epsilon)$, under modest assumptions that the input matrix has stable rank d and a large Frobenius norm. For constant-factor approximation to the largest singular value $\|\cdot\|$, Rudelson and Vershynin [32] showed that uniformly sampling q rows of a matrix $\mathbf{A}$ gives a subsampled matrix $\mathbf{A}_{\mathsf{row}}$ such that $\|\mathbf{A}_{\mathsf{row}}\| \lesssim \sqrt{\frac{q}{n}}\|\mathbf{A}\| + \sqrt{\log q}\|\mathbf{A}\|_{(n/q)}$, where $\|\mathbf{A}\|_{(n/q)}$ is the average of n/q biggest Euclidean lengths of the columns of $\mathbf{A}$.

B New Operator Norm Estimators

In this section, we develop new $(1 \pm \tau)$-approximation estimators to the operator norm in sampling and sensing models.

B.1 Sampling Algorithms

We first discuss the sampling algorithms which are only allowed to read the entries of a matrix.

B.1.1 Estimation without Eigengap

Before proceeding, we first cite the following result from [28].

LEMMA B.1. (THEOREM 20, [28]) *Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ have rows $\{\mathbf{A}_{t,:}\}_{t=1}^n$. Independently sample q rows $\mathbf{A}_{t_1,:}, \ldots, \mathbf{A}_{t_q,:}$ with replacement from $\mathbf{A}$ according to the probabilities:*

$$p_t \ge \beta \frac{\|\mathbf{A}_{t,:}\|_2^2}{\|\mathbf{A}\|_F^2}$$

for $\beta < 1$. Let

$$\mathbf{A}_0 = \begin{bmatrix} \frac{\mathbf{A}_{t_1,:}}{\sqrt{qp_{t_1}}} \\ \vdots \\ \frac{\mathbf{A}_{t_q,:}}{\sqrt{qp_{t_q}}} \end{bmatrix}.$$

Then if $q \geq \frac{4\mathsf{srank}(\mathbf{A})}{\beta\tau^2}\log\frac{2n}{\delta}$, with probability at least $1-\delta$, we have

$$\|\mathbf{A}^\top\mathbf{A} - \mathbf{A}_0^\top\mathbf{A}_0\| \leq \tau\|\mathbf{A}\|^2.$$

REMARK 1. *Lemma B.1 implies that*

$$(1-\tau)\|\mathbf{A}\|^2 \leq \|\mathbf{A}_0\|^2 \leq (1+\tau)\|\mathbf{A}\|^2,$$

because

$$\begin{aligned} \left|\|\mathbf{A}\|^2 - \|\mathbf{A}_0\|^2\right| &= \left|\|\mathbf{A}^\top\mathbf{A}\| - \|\mathbf{A}_0^\top\mathbf{A}_0\|\right| \\ &\leq \|\mathbf{A}^\top\mathbf{A} - \mathbf{A}_0^\top\mathbf{A}_0\| \leq \tau\|\mathbf{A}\|^2. \end{aligned}$$

THEOREM B.1. *Suppose that $\mathbf{A}$ is an $n \times n$ matrix satisfying that $\|\mathbf{A}\|_F^2 = \Omega(\tau n^2)$, $\|\mathbf{A}\|_\infty \leq 1$ and $\mathsf{srank}(\mathbf{A}) = \mathcal{O}(d)$. Then with probability at least 0.9, the output of Algorithm 5 satisfies $(1-\tau)\|\mathbf{A}\| \leq \|\widetilde{\mathbf{A}}\| \leq (1+\tau)\|\mathbf{A}\|$. The sample complexity is $\mathcal{O}(d^2\log^2(n)/\tau^4)$.*

Proof. We note that for any row $\mathbf{A}_{i,:}$ such that $|\mathbf{A}_{i,j}| \leq 1$ and $\eta \leq \|\mathbf{A}_{i,:}\|_2^2 \leq n$, uniformly sampling $\Theta(\frac{n}{\eta})$ entries of $\mathbf{A}_{i,:}$ suffices to estimate $\|\mathbf{A}_{i,:}\|_2^2$ within a constant multiplicative factor. To see this, we use Chebyshev's inequality. Let $s = \Theta(\frac{n}{\eta})$ be the number of sampled entries, Z_j be the square of the j-th sampled entry $\mathbf{A}_{i,l(j)}$ of vector $\mathbf{A}_{i,:}$, and $Z = \frac{n}{s}\sum_{j=1}^s Z_j$. So Z is an unbiased estimator:

$$\mathbb{E}[Z] = \frac{n}{s}s\mathbb{E}[Z_1] = n\sum_{j=1}^n \frac{1}{n}\mathbf{A}_{i,l(j)}^2 = \|\mathbf{A}_{i,:}\|_2^2.$$

For the variance, we have

$$\begin{aligned} \mathsf{Var}[Z] = \frac{n^2}{s^2}\sum_{j=1}^s \mathsf{Var}[Z_j] &\leq \frac{n^2}{s^2}\sum_{j=1}^s \mathbb{E}[Z_j^2] = \frac{n^2}{s}\,\mathbb{E}[Z_1^2] \\ &= \frac{n^2}{s}\sum_{j=1}^n \frac{1}{n}\mathbf{A}_{i,j}^4 \\ &\leq \frac{n}{s}\sum_{j=1}^n \mathbf{A}_{i,j}^2 \qquad (\text{since } |\mathbf{A}_{i,j}| \leq 1) \\ &= \Theta(\eta)\|\mathbf{A}_{i,:}\|_2^2 \\ &\leq \Theta(\|\mathbf{A}_{i,:}\|_2^4). \qquad (\text{since } \eta \leq \|\mathbf{A}_{i,:}\|_2^2) \end{aligned}$$

Therefore, by Chebyshev's inequality, we have

$$\Pr\left[\left|Z - \|\mathbf{A}_{i,:}\|_2^2\right| \geq 10\|\mathbf{A}_{i,:}\|_2^2\right] \leq \frac{1}{3}.$$

Note that in Step 5 of Algorithm 5, in total we sample $q_{\mathsf{row}} = \mathcal{O}(\frac{d\log n}{\tau^2})$ row indices, obeying the conditions in Lemma B.1 for a constant β. By concentration, with high probability $r = \mathcal{O}(\frac{\|\mathbf{A}\|_F^2}{\tau n})$ in Step 5, because in expectation we sample $\mathcal{O}(\frac{1}{\tau^2})$ entries to estimate r and we scale $\|\mathbf{x}\|_2^2$ by a τn factor in Steps 3 and 4, and that $\|\mathbf{A}\|_F^2$ is as large as $\Omega(\tau n^2)$. The probability that any given row i is sampled is equal to $\frac{1}{n\tau} \times \frac{r_i}{r} = \Omega(\frac{r_i}{\|\mathbf{A}\|_F^2})$. Suppose first that $\|\mathbf{A}_{i,:}\|_2^2 \leq \tau n$. Then we have $r_i = \tau n$. Consequently, for such i, the probability of sampling row i is at least $\Omega(\frac{\tau n}{\|\mathbf{A}\|_F^2}) \geq \Theta(\frac{\|\mathbf{A}_{i,:}\|_2^2}{\|\mathbf{A}\|_F^2})$, just as in Lemma B.1. Suppose next that $\|\mathbf{A}_{i,:}\|_2^2 \geq \tau n$. Then we have $r_i = \Theta(\|\mathbf{A}_{i,:}\|_2^2)$. Consequently, for such i, the probability of sampling row i is at least $\Omega(\frac{\|\mathbf{A}_{i,:}\|_2^2}{\|\mathbf{A}\|_F^2})$, just as in Lemma B.1. Therefore, in the followings we can set β in Lemma B.1 as an absolute constant.

It follows from Lemma B.1 that with probability at least 0.9,

$$(1-\tau)\|\mathbf{A}\|^2 \leq \|\mathbf{A}_{\mathsf{row}}\|^2 \leq (1+\tau)\|\mathbf{A}\|^2,$$

where $\mathbf{A}_{\mathsf{row}}$ is the *scaled* row sampling of $\mathbf{A}$ as in Lemma B.1. Conditioning on this event, by applying Lemma B.1 again to the column sampling of $\mathbf{A}_{\mathsf{row}}$, we have with high probability,

$$\begin{aligned} \text{(B.1)} \quad (1-\tau)^2\|\mathbf{A}\|^2 \leq (1-\tau)\|\mathbf{A}_{\mathsf{row}}\|^2 &\leq \|\widetilde{\mathbf{A}}\|^2 \\ &\leq (1+\tau)\|\mathbf{A}_{\mathsf{row}}\|^2 \leq (1+\tau)^2\|\mathbf{A}\|^2, \end{aligned}$$

where we have used the fact that $\mathsf{srank}(\mathbf{A}_{\mathsf{row}}) = \mathcal{O}(d)$. The statement $\mathsf{srank}(\mathbf{A}_{\mathsf{row}}) = \mathcal{O}(d)$ holds because $\mathbb{E}\,\|\mathbf{A}_{\mathsf{row}}\|_F^2 = \|\mathbf{A}\|_F^2$ and by the Markov bound, we have with constant probability that $\|\mathbf{A}_{\mathsf{row}}\|_F^2 \leq c\|\mathbf{A}\|_F^2$, so

$$\begin{aligned} \mathsf{srank}(\mathbf{A}_{\mathsf{row}}) &= \frac{\|\mathbf{A}_{\mathsf{row}}\|_F^2}{\|\mathbf{A}_{\mathsf{row}}\|^2} \\ &\leq \frac{c\|\mathbf{A}\|_F^2}{(1-\tau)\|\mathbf{A}\|^2} \leq C\mathsf{srank}(\mathbf{A}) \leq C'd. \end{aligned}$$

□

B.1.2 Estimation with Eigengap Let $\mathbf{A} \in \mathbb{R}^{n\times n}$. Suppose that $p = 2q$. We define a cycle σ to be an ordered pair of a sequence of length q: $\lambda = ((i_1, ..., i_q), (j_1, ..., j_q))$ such that $i_r, j_r \in [k]$ for all r. Now we associate with λ a scalar

$$\text{(B.2)} \qquad \mathbf{A}_\lambda = \prod_{\ell=1}^q \mathbf{A}_{i_\ell,j_\ell}\mathbf{A}_{i_{\ell+1},j_\ell},$$

Algorithm 5 The sampling algorithm to estimate $\|\mathbf{A}\|$ up to $(1 \pm \tau)$ relative error

```
    ▷ Lines 1-5 estimates the row norms of $\mathbf{A}$ and then sample rows non-uniformly.
 1: Sample each row of $\mathbf{A}$ by Bernoulli distribution with probability $\mathcal{O}(\frac{1}{n\tau})$. Denote by $\mathcal{S}_{\mathsf{row}}$ the sampled set and $q = |\mathcal{S}_{\mathsf{row}}|$.
 2: for $i \leftarrow 1$ to $q$ do
 3:     Uniformly sample $\mathcal{O}(\frac{1}{\tau})$ entries from $\mathbf{A}_{\mathcal{S}_{\mathsf{row}}(i),:}$, forming vector $\mathbf{x}$.
 4:     $r_i \leftarrow \max\{\tau n\|\mathbf{x}\|_2^2, \tau n\}$.
 5: Sample $q_{\mathsf{row}} = \mathcal{O}(\frac{d\log n}{\tau^2})$ indices in $\mathcal{S}_{\mathsf{row}}$ independently with replacement according to the probability $p_i = \frac{r_i}{r}$, where $r = \sum_{j=1}^{q} r_j$. Denote by $\mathcal{I}_{\mathsf{row}}$ the sampled row indices.

    ▷ Lines 6-10 estimates the column norms of $\mathbf{A}$ and then sample columns non-uniformly.
 6: Sample each row with probability $\mathcal{O}(\frac{1}{n\tau})$. Repeat the procedure $n$ times with replacement. Denote the sampled set by $\mathcal{S}_{\mathsf{col}}$ and $q' = |\mathcal{S}_{\mathsf{col}}|$.
 7: for $i \leftarrow 1$ to $q'$ do
 8:     Uniformly sample $\mathcal{O}(\frac{1}{\tau})$ entries from $\mathbf{A}_{\mathcal{I}_{\mathsf{row}},\mathcal{S}_{\mathsf{col}(i)}}$, forming vector $\mathbf{x}$.
 9:     $r'_i \leftarrow \max\{\tau q\|\mathbf{x}\|_2^2, \tau q\}$.
10: Sample $q_{\mathsf{col}} = \mathcal{O}(\frac{d\log n}{\tau^2})$ indices in $\mathcal{S}_{\mathsf{col}}$ independently with replacement according to the probability $p'_i = \frac{r'_i}{r'}$, where $r' = \sum_{j=1}^{q'} r'_j$. Denote by $\mathcal{I}_{\mathsf{col}}$ the sampled row indices.
11: $\widetilde{\mathbf{A}} \leftarrow \mathbf{A}_{\mathcal{I}_{\mathsf{row}},\mathcal{I}_{\mathsf{col}}}$. Rescale the rows of $\widetilde{\mathbf{A}}$ by $\left\{\sqrt{\frac{q}{p_i q_{\mathsf{row}}}}\right\}$ and the columns of $\widetilde{\mathbf{A}}$ by $\left\{\sqrt{\frac{q'}{p'_i q_{\mathsf{col}}}}\right\}$.
12: return index sets $\mathcal{I}_{\mathsf{row}}$, $\mathcal{I}_{\mathsf{col}}$, scaling factors $\left\{\sqrt{\frac{q}{p_i q_{\mathsf{row}}}}\right\}$, $\left\{\sqrt{\frac{q'}{p'_i q_{\mathsf{col}}}}\right\}$, $\widetilde{\mathbf{A}}$, and $\|\widetilde{\mathbf{A}}\|$.
```

where for convention we define that $i_{q+1} = i_1$. Denote by

$$Z = \frac{1}{N}\sum_{i=1}^{N} \mathbf{A}_{\lambda_i}. \tag{B.3}$$

Our goal is to estimate $\sigma_1(\mathbf{A})$ up to $(1 \pm \tau)$ relative error, which is an $(1 \pm \tau)$ approximation to $\|\mathbf{A}\|$.

Algorithm 6 Estimate $\|\mathbf{A}\|$ up to $(1 \pm \tau)$ relative error

```
Input: Cycle length $q$, matrix size $n$.
Output: $(1 \pm \tau)$-approximation estimator.
for $i = 1$ to $N$ do
    Uniformly sample a cycle $\lambda_i$ of length $q$.
    Compute $\mathbf{A}_{\lambda_i}$ by Eqn. (B.2).
Compute $Z$ as defined in (B.3).
return $Z^{1/(2q)}n$.
```

THEOREM B.2. *Let $\tau \in (0, \frac{1}{2})$ be the accuracy parameter and suppose that the input matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ satisfies*

- $\|\mathbf{A}\|_\infty \leq 1$*;*
- $\|\mathbf{A}\|_F \geq cn$ *for some absolute constant* $c > 0$*;*
- $\sigma_2(\mathbf{A})/\sigma_1(\mathbf{A}) \leq \tau^\gamma$ *for some absolute constant* $\gamma > 0$*;*
- $\mathsf{srank}(\mathbf{A}) = \mathcal{O}(1)$*.*

Let $N = \frac{C_1}{\tau^2}\exp(\frac{c_1}{\gamma})$ and $q = \frac{C_2}{\gamma}$ for some large constants $C_1, C_2 > 0$ and some small constant $c_1 > 0$. Then with probability at least 0.9, the estimator returned by Algorithm 6 satisfies $(1-\tau)\|\mathbf{A}\| \leq Z^{1/(2q)}n \leq (1+\tau)\|\mathbf{A}\|$. The sample complexity is $\Theta(Nq) = \Theta\left(\frac{1}{\gamma\tau^2}\exp(\frac{c_1}{\gamma})\right)$.

Proof. [Proof of Theorem B.2] We show that the cycle estimator approximates $\|\mathbf{A}\|$ within a $(1 \pm \tau)$ relative error. Let $\lambda = (\{i_s\}, \{j_s\})$ which is chosen uniformly with replacement. Recall that $\mathbf{A}_\lambda = \prod_{\ell=1}^{q} \mathbf{A}_{i_\ell,j_\ell}\mathbf{A}_{i_{\ell+1},j_\ell}$. Hence

$$\begin{aligned}\mathbb{E}\mathbf{A}_\lambda &= \mathbb{E}\left[\prod_{\ell=1}^{q} \mathbf{A}_{i_\ell,j_\ell}\mathbf{A}_{i_{\ell+1},j_\ell}\right] \\ &= \frac{1}{n^{2q}}\left[\sum_{i_1,i_2,\ldots,i_q,j_1,j_2,\ldots,j_q}\prod_{\ell=1}^{q} \mathbf{A}_{i_\ell,j_\ell}\mathbf{A}_{i_{\ell+1},j_\ell}\right].\end{aligned}$$

Note that (see, e.g., [24])

$$\sum_{i_1,i_2,\ldots,i_q,j_1,j_2,\ldots,j_q}\prod_{\ell=1}^{q} \mathbf{A}_{i_\ell,j_\ell}\mathbf{A}_{i_{\ell+1},j_\ell} = \|\mathbf{A}\|_{2q}^{2q},$$

and by the assumption on the singular values and the stable rank, $\sigma_1(\mathbf{A})^{2q} \leq \|\mathbf{A}\|_{2q}^{2q} \leq (1+\tau)\sigma_1(\mathbf{A})^{2q}$, provided that $q \geq \frac{1}{2\gamma}(\frac{\log \mathsf{srank}(\mathbf{A})}{\log(1/\tau)} + 1)$, and thus it suffices to take $q = \Theta(\frac{1}{\gamma})$.

Therefore, noting that $\mathbb{E}[Z] = \mathbb{E}[\mathbf{A}_\lambda]$,

$$\mathbb{E}[Z] \le \frac{1+\tau}{n^{2q}}\sigma_1(\mathbf{A})^{2q} \le 1+\tau, \tag{B.4}$$

$$\begin{aligned}\mathbb{E}[Z] &\ge \frac{1}{n^{2q}}\sigma_1(\mathbf{A})^{2q} \ge \frac{1}{n^{2q}}\left(\frac{\|\mathbf{A}\|_F^2}{\mathsf{srank}(\mathbf{A})}\right)^q \\ &\ge \left(\frac{c^2}{\mathsf{srank}(\mathbf{A})}\right)^q = \exp\left(\frac{c_1}{\gamma}\right).\end{aligned} \tag{B.5}$$

We now bound the variance of $\mathbf{A}_\lambda$. Observe that $\mathsf{Var}[\mathbf{A}_\lambda] \le \mathbb{E}[\mathbf{A}_\lambda^2] \le 1$, because $|\mathbf{A}_{i,j}| \le 1$ for all $i,j \in [n]$. Thus by repeating the procedure $N = \frac{C_1}{\tau^2}\exp\left(\frac{2c_1}{\gamma}\right)$ times, we have $\mathsf{Var}[Z] = \frac{1}{N}\mathsf{Var}[\mathbf{A}_\lambda] \le \frac{1}{10}\tau^2\exp\left(\frac{2c_1}{\gamma}\right)$, by choosing C_1 sufficiently large. It follows from the Chebyshev inequality that $\Pr\left[|\mathbb{E}[Z]-Z| > \tau\mathbb{E}[Z]\right] \le \frac{\mathsf{Var}[Z]}{\tau^2\mathbb{E}[Z]^2} \le \frac{1}{10}$, where we have used the lower bound (B.5). This together with (B.4) and (B.5) implies that

$$\Pr\left[(1-\tau)\frac{1}{n^{2q}}\sigma_1(\mathbf{A})^{2q} \le Z \le (1+\tau)^2\frac{1}{n^{2q}}\sigma_1(\mathbf{A})^{2q}\right] > \frac{9}{10}.$$

So

$$\Pr\left[(1-\tau)\,\sigma_1(\mathbf{A}) \le Z^{1/(2q)}n \le (1+\tau)\,\sigma_1(\mathbf{A})\right] > \frac{9}{10}.$$

□

B.2 Sensing Algorithms

Theorem B.3. *Suppose that $\mathbf{A}$ is an $n \times n$ matrix such that $\|\mathbf{A}\|_F^2 = \Omega(\tau n^2)$, $\|\mathbf{A}\|_\infty \le 1$ and $\mathsf{srank}(\mathbf{A}) = \mathcal{O}(d)$. Then Algorithm 7 outputs a value Z, which satisfies $(1-\tau)\|\mathbf{A}\| \le Z \le (1+\tau)\|\mathbf{A}\|$ with probability at least 0.9. The sketching complexity is $\mathcal{O}(\max\{\log^2(d\log(n)/\tau), d^2\log(n)\}/\tau^2)$.*

Before proving Theorem B.3, we introduce a new estimator of operator norm under the sensing model, which approximates the operator norm by the Schatten-p norm of large p.

Specifically, let $\mathbf{A}$ be an $n \times n$ matrix. We define a cycle σ to be an ordered pair of a sequence of length q with $p = 2q$: $\lambda = ((i_1,\ldots,i_q),(j_1,\ldots,j_q))$ such that $i_r, j_r \in [k]$ for all r, $i_r \ne i_s$ and $j_r \ne j_s$ for $r \ne s$. Now we associate with λ a scalar

$$\mathbf{A}_\lambda = \prod_{\ell=1}^{q}\mathbf{A}_{i_\ell,j_\ell}\mathbf{A}_{i_{\ell+1},j_\ell}, \tag{B.6}$$

where for convention we define that $i_{q+1} = i_1$. Denote by $\mathcal{C}$ the set of cycles. We define

$$Y = \frac{1}{|\mathcal{C}|}\sum_{\lambda\in\mathcal{C}}(\mathbf{G}\mathbf{A}\mathbf{H}^\top)_\lambda \tag{B.7}$$

for even p, where $\mathbf{G} \sim \mathcal{G}(k,n)$, $\mathbf{H} \sim \mathcal{G}(k,n)$, and $k \ge q$. This estimator, akin to that in [21], approximates the Schatten-p and thus the operator norm, as we shall show below.

Algorithm 7 The sketching/sensing algorithm to estimate $\|\mathbf{A}\|$ up to $(1\pm\tau)$ relative error

```
Obtain indices $\mathcal{I}_{\mathsf{row}}$, $\mathcal{I}_{\mathsf{col}}$ and scaling factors $\left\{\sqrt{\frac{q}{p_i q_{\mathsf{row}}}}\right\}$, $\left\{\sqrt{\frac{q'}{p'_i q_{\mathsf{col}}}}\right\}$ by Algorithm 5 with $|\mathcal{I}_{\mathsf{row}}| = |\mathcal{I}_{\mathsf{col}}| = \mathcal{O}(d\log(n)/\tau^2)$.
Let $\mathbf{G}$ and $\mathbf{H}$ be $\Theta(\frac{\max\{\log(d\log(n)/\tau),d\}}{\tau}) \times \mathcal{O}(\frac{d\log n}{\tau^2})$ matrices with i.i.d. $\mathcal{N}(0,1)$ entries. Scale the columns of $\mathbf{G}$ by $\left\{\sqrt{\frac{q}{p_i q_{\mathsf{row}}}}\right\}$ and the columns of $\mathbf{H}$ by $\left\{\sqrt{\frac{q'}{p'_i q_{\mathsf{col}}}}\right\}$.
Maintain $\mathbf{G}\mathbf{A}_{\mathcal{I}_{\mathsf{row}},\mathcal{I}_{\mathsf{col}}}\mathbf{H}^\top$.
Compute $Y$ defined in Eqn. (B.7).
return $Y^{\tau/(2\log(d\log(n)/\tau^2))}$.
```

Lemma B.2. *Suppose that $\mathbf{A}$ is a $n \times n$ matrix of stable rank at most d. Let $k = \Theta(\max\{\sqrt{nd}, \log n\})$ and Y be the estimator defined in (B.7). With probability at least 0.9, it holds that $(1-\tau)\|\mathbf{A}\| \le Y^{\tau/(2\log(n))} \le (1+\tau)\|\mathbf{A}\|$. The sketching complexity is $\mathcal{O}(k^2) = \mathcal{O}(\max\{nd, \log^2 n\})$.*

Proof. [Proof of Lemma B.2] We first show that $\|\mathbf{A}\|_{\mathcal{S}_p}$ and $\|\mathbf{A}\|$ differ at most a $(1\pm\tau)$ factor for $p = 2\lceil\log(n)/\tau\rceil$. To see this,

$$1 \le \frac{\|\mathbf{A}\|_{\mathcal{S}_p}^p}{\|\mathbf{A}\|^p} = \frac{\sigma_1^p(\mathbf{A}) + \sigma_2^p(\mathbf{A}) + \cdots + \sigma_n^p(\mathbf{A})}{\sigma_1^p(\mathbf{A})} \le n,$$

and therefore $1 \le \frac{\|\mathbf{A}\|_{\mathcal{S}_p}}{\|\mathbf{A}\|} \le n^{1/p} \le 1 + \frac{1}{2}\tau$.

We now show that the cycle estimator $Y^{1/p}$ approximates $\|\mathbf{A}\|_{\mathcal{S}_p}$ within a $(1\pm\frac{1}{2}\tau)$ relative error. We say that two cycles $\lambda = (\{i\},\{j\})$ and $\tau = (\{i'\},\{j'\})$ are (a_1,a_2)-disjoint if $|i\Delta i'| = 2a_1$ and $|j\Delta j'| = 2a_2$, denoted by $|\lambda\Delta\tau| = (a_1,a_2)$. Here Δ is the symmetric difference. Denote by $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ the skinny SVD of $\mathbf{A}$. Let $\mathbf{G}$ and $\mathbf{H}$ be random matrices with i.i.d. $\mathcal{N}(0,1)$ entries. Note that $\mathbf{G}\mathbf{A}\mathbf{H}^\top$ is identically distributed as $\mathbf{G}\mathbf{\Sigma}\mathbf{H}^\top$ by rotational invariance. Let $\widetilde{\mathbf{A}}$ be the $k \times k$ matrix $\mathbf{G}\mathbf{\Sigma}\mathbf{H}^\top$, where $k \ge q$. It is clear that $\widetilde{\mathbf{A}}_{s,t} = \sum_{i=1}^n \sigma_i\mathbf{G}_{s,i}\mathbf{H}_{t,i}$. Define $Y = \frac{1}{|\mathcal{C}|}\sum_{\lambda\in\mathcal{C}}\widetilde{\mathbf{A}}_\lambda$. Let $\lambda = (\{i_s\},\{j_s\})$. Then

$$\widetilde{\mathbf{A}}_\lambda = \sum_{\substack{\ell_1\in[n],\ldots,\ell_q\in[n]\\ m_1\in[n],\ldots,m_q\in[n]}}\prod_{s=1}^{q}\sigma_{\ell_s}\sigma_{m_s}\mathbf{G}_{i_s,\ell_s}\mathbf{H}_{j_s,\ell_s}\mathbf{G}_{i_{s+1},m_s}\mathbf{H}_{j_s,m_s}.$$

We note that

$$\mathbb{E} Y = \mathbb{E}\,\widetilde{\mathbf{A}}_\lambda = \sum_{i=1}^{n} \sigma_i^{2q} = \|\mathbf{A}\|_{\mathcal{S}_p}^p.$$

We now bound the variance of Y. Let $\tau = (\{i'_s\}, \{j'_s\})$. Observe that

$$\mathbb{E}Y^2 = \frac{1}{|\mathcal{C}|^2} \sum_{a_1=0}^{q} \sum_{a_2=0}^{q} \sum_{\substack{\lambda,\tau\in\mathcal{C}\\ |\lambda\Delta\tau|=(a_1,a_2)}} \mathbb{E}(\widetilde{\mathbf{A}}_\lambda \widetilde{\mathbf{A}}_\tau),$$

where

(B.8)

$$\begin{aligned}\mathbb{E}(\widetilde{\mathbf{A}}_\lambda \widetilde{\mathbf{A}}_\tau) = & \sum_{\substack{\ell_1\in[n],\dots,\ell_q\in[n]\\ \ell'_1\in[n],\dots,\ell'_q\in[n]\\ m_1\in[n],\dots,m_q\in[n]\\ m'_1\in[n],\dots,m'_q\in[n]}} \left(\prod_{i=1}^{q} \sigma_{\ell_i}\sigma_{m_i}\sigma_{\ell'_i}\sigma_{m'_i}\right)\\ & \times \mathbb{E}\left(\prod_{s=1}^{q} \mathbf{G}_{i_s,\ell_s}\mathbf{G}_{i_{s+1},m_s}\mathbf{G}_{i'_s,\ell'_s}\mathbf{G}_{i'_{s+1},m'_s}\right)\\ & \times \mathbb{E}\left(\prod_{s=1}^{q} \mathbf{H}_{j_s,\ell_s}\mathbf{H}_{j_s,m_s}\mathbf{H}_{j'_s,\ell'_s}\mathbf{H}_{j'_s,m'_s}\right).\end{aligned}$$

For any fixed cycles $\lambda = (\{i_s\}, \{j_s\})$ and $\tau = (\{i'_s\}, \{j'_s\})$ such that $|\lambda\Delta\tau| = (a_1, a_2)$, we notice that

$$\text{(B.9)} \qquad \mathbb{E}(\widetilde{\mathbf{A}}_\lambda \widetilde{\mathbf{A}}_\tau) \le (2cnd)^p \|\mathbf{A}\|_{\mathcal{S}_p}^{2p},$$

for an absolute constant c. To see this, we observe that for the expectation $\mathbb{E}(\widetilde{\mathbf{A}}_\lambda \widetilde{\mathbf{A}}_\tau)$ to be non-zero, we must have that each appeared $\mathbf{G}$ and $\mathbf{H}$ in Eqn. (B.8) repeats an even number of times. Though there are totally n^{4q} many of configurations for $\{\ell_s\}$, $\{\ell'_s\}$, $\{m_s\}$ and $\{m'_s\}$, there are at most $n^{2q}3^q$ non-zero terms among the summation in Eqn. (B.8). This is because each $\mathbf{G}$ and $\mathbf{H}$ must have power 2 or 4 by the construction of the cycle. We know that for each fixed configuration of blocks there are at most n^{2q} free variables, and there are at most 16^q different kinds of configurations of blocks because the size of each block is at most 4. So the number of non-zero terms is at most $(4n)^{2q}$. This is true no matter whether there exists some i_r, i'_s or j_r, j'_s such that $i_r = i'_s$ or $j_r = j'_s$. We also claim that for each non-zero term in the summation of Eqn. (B.8),

$$\begin{aligned}\mathbb{E}&\left(\prod_{s=1}^{q} \mathbf{G}_{i_s,\ell_s}\mathbf{G}_{i_{s+1},m_s}\mathbf{G}_{i'_s,\ell'_s}\mathbf{G}_{i'_{s+1},m'_s}\right)\\ &\cdot \mathbb{E}\left(\prod_{s=1}^{q} \mathbf{H}_{j_s,\ell_s}\mathbf{H}_{j_s,m_s}\mathbf{H}_{j'_s,\ell'_s}\mathbf{H}_{j'_s,m'_s}\right) \le 25^q.\end{aligned}$$

This is because $\mathbb{E}\,\mathbf{G}^2 = \mathbb{E}\,\mathbf{H}^2 = 1$ and $\mathbb{E}\,\mathbf{G}^4 = \mathbb{E}\,\mathbf{H}^4 = 3$. Therefore, for a certain configuration in which $p_1, \dots, p_w$ are free variables with multiplicity $r_1, \dots, r_w \ge 2$, the summation in Eqn. (B.8) is bounded by

$$\begin{aligned}4n^{2q}100^q \sum_{p_1,\dots,p_w} \sigma_{p_1}^{r_1}\cdots\sigma_{p_w}^{r_w} &\le (2n)^p \|\mathbf{A}\|_{\mathcal{S}_{r_1}}^{r_1}\cdots\|\mathbf{A}\|_{\mathcal{S}_{r_w}}^{r_w}\\ &\le (2nd)^p \|\mathbf{A}\|_{\mathcal{S}_p}^{2p},\end{aligned}$$

where the last inequality follows from the facts that $\sum_{i=1}^{w} r_i = 2p$ and, by the assumption $\mathsf{srank}(A) \le d$, that $\|\mathbf{A}\|_{\mathcal{S}_r} \le \|\mathbf{A}\|_F \le \sqrt{d}\|\mathbf{A}\|_{\mathcal{S}_p}$ for any $r \ge 2$. Thus we obtain Eqn. (B.9).

We now bound $\mathbb{E}Y^2$. Note that $|\mathcal{C}| = \Theta(k^p)$ and there are

$$\binom{k}{q}\binom{q}{q-a_1}\binom{k-(q-a_1)}{a_1}\binom{k}{q}\binom{q}{q-a_2}\binom{k-(q-a_2)}{a_2}$$

pairs of (a_1, a_2)-disjoint cycles, which can be upper bounded by $\mathcal{O}(10^q)$. Hence

$$\begin{aligned}\mathbb{E}Y^2 &= \frac{1}{|\mathcal{C}|^2} \sum_{a_1=0}^{q} \sum_{a_2=0}^{q} \sum_{\substack{\lambda,\tau\in\mathcal{C}\\ |\lambda\Delta\tau|=(a_1,a_2)}} \mathbb{E}(\widetilde{\mathbf{A}}_\lambda \widetilde{\mathbf{A}}_\tau)\\ &\le C' \frac{1}{k^{2p}} q^2 10^q (2nd)^p \|\mathbf{A}\|_{\mathcal{S}_p}^{2p} \le \|\mathbf{A}\|_{\mathcal{S}_p}^{2p}\end{aligned}$$

by the assumption that $k = \Omega(\sqrt{nd})$.

It follows that $\mathsf{Var}[Y] \le \mathbb{E}\,Y^2 \le \|\mathbf{A}\|_{\mathcal{S}_p}^{2p}$. Then by the Chebyshev inequality,

$$\Pr\left[\left|\|\mathbf{A}\|_{\mathcal{S}_p}^p - Y\right| > \frac{1}{2}\|\mathbf{A}\|_{\mathcal{S}_p}^p\right] \le \frac{\mathsf{Var}[Y]}{4\|\mathbf{A}\|_{\mathcal{S}_p}^{2p}} \le \frac{1}{10},$$

i.e., $\Pr\left[\left(1-\frac{1}{2}\tau\right)\|\mathbf{A}\|_{\mathcal{S}_p} \le Y^{1/p} \le \left(1+\frac{1}{2}\tau\right)\|\mathbf{A}\|_{\mathcal{S}_p}\right] > \frac{9}{10}$. This together with the fact that $\|\mathbf{A}\| \le \|\mathbf{A}\|_{\mathcal{S}_p} \le (1+\frac{1}{2}\tau)\|\mathbf{A}\|$ implies that $\Pr\left[(1-\tau)\|\mathbf{A}\| \le Y^{1/p} \le (1+\tau)\|\mathbf{A}\|\right] > \frac{9}{10}$, as desired. This completes the proof of Lemma B.2. □

We are now ready to prove Theorem B.3. Recall that we have shown that by focusing on an $\mathcal{O}(\frac{d\log n}{\tau^2}) \times \mathcal{O}(\frac{d\log n}{\tau^2})$ submatrix (without sampling it), we can achieve guarantee (B.1) when $\|\mathbf{A}\|_F^2 = \Omega(\tau n^2)$ and $\|\mathbf{A}\|_\infty \le 1$. Letting $d \leftarrow c_1 d$ and $n \leftarrow \mathcal{O}(\frac{d\log n}{\tau^2})$ in Lemma B.2 concludes the proof of Theorem B.3.