New Tools for Smoothed Analysis: Least Singular Value Bounds for Random Matrices with Dependent Entries

Aditya Bhaskara
University of Utah
Salt Lake City, USA
bhaskaraaditya@gmail.com

Eric Evert
Northwestern University
Evanston, USA
eric.evert@northwestern.edu

Vaidehi Srinivas
Northwestern University
Evanston, USA
vaidehi@u.northwestern.edu

Aravindan Vijayaraghavan
Northwestern University
Evanston, USA
aravindv@northwestern.edu

ABSTRACT

We develop new techniques for proving lower bounds on the least singular value of random matrices with limited randomness. The matrices we consider have entries that are given by polynomials of a few underlying base random variables. This setting captures a core technical challenge for obtaining smoothed analysis guarantees in many algorithmic settings. Least singular value bounds often involve showing strong anti-concentration inequalities that are intricate and much less understood compared to concentration (or large deviation) bounds.

First, we introduce a general technique for proving anti-concentration that uses well-conditionedness properties of the Jacobian of a polynomial map, and show how to combine this with a hierarchical ε-net argument to prove least singular value bounds. Our second tool is a new statement about least singular values to reason about higher-order lifts of smoothed matrices and the action of linear operators on them.

Apart from getting simpler proofs of existing smoothed analysis results, we use these tools to now handle more general families of random matrices. This allows us to produce smoothed analysis guarantees in several previously open settings. These new settings include smoothed analysis guarantees for power sum decompositions and certifying robust entanglement of subspaces, where prior work could only establish least singular value bounds for fully random instances or only show non-robust genericity guarantees.

CCS CONCEPTS

• **Theory of computation** → **Design and analysis of algorithms**; **Randomness, geometry and discrete structures**.

KEYWORDS

smoothed analysis, tensors, random matrices, least singular values, matrix anticoncentration, unsupervised learning, quantum entanglement

STOC '24, June 24–28, 2024, Vancouver, BC, Canada

ACM ISBN 979-8-4007-0383-6/24/06
https://doi.org/10.1145/3618260.3649765

ACM Reference Format:
Aditya Bhaskara, Eric Evert, Vaidehi Srinivas, and Aravindan Vijayaraghavan. 2024. New Tools for Smoothed Analysis: Least Singular Value Bounds for Random Matrices with Dependent Entries. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC '24), June 24–28, 2024, Vancouver, BC, Canada.* ACM, New York, NY, USA, 30 pages. https://doi.org/10.1145/3618260.3649765

1 INTRODUCTION

Over the past two decades, there has been significant progress in using algebraic methods for high-dimensional statistical estimation (e.g., [2]). Techniques like tensor decomposition have been used for parameter estimation in mixture models [3, 10, 14], shallow neural networks [5, 25], stochastic block models [2], and more [26]. Recently, more sophisticated decomposition methods based on tensor networks [21], circuit complexity [12] and algebraic geometry [12, 19] have given to rise to new algorithms for many problems in high-dimensional geometry and parameter estimation. These algorithms start by building appropriate algebraic structures that "encode" the hidden parameters of interest. Then, they use the algebraic techniques described above for recovering the solution.

Unfortunately, in most of these applications, the recovery problem turns out to be NP hard in general. So the algorithms have provable recovery guarantees only under certain *algebraic* conditions. Typically, these conditions can be formulated in terms of appropriately defined matrices being well-conditioned, i.e., having a non-negligible least singular value. Furthermore, the least singular value determines the sample complexity and running time, and so it is important to obtain inverse polynomial bounds.

Now it is natural to ask: *do the algebraic conditions typically hold?* Due to NP hardness, we know there exist parameters for which the conditions do not hold. But how common or rare are such parameter settings/instances? A strong way to address this question is via the framework of smoothed analysis, developed in the seminal work of Spielman and Teng [23, 27, 28]. A condition is said to hold in a smoothed analysis setting if for *any* instance, a small random perturbation of magnitude, say $\rho = 1/n^2$, where n is the input size, results in an instance that satisfies the condition with high probability. Smoothed analysis guarantees show that any potential bad instance is isolated or degenerate: most other instances in a small ball around it have good guarantees. On the one hand, smoothed analysis gives a much stronger guarantee than

average case analysis, where one shows that the condition holds w.h.p. for a random choice of parameters from some distribution. On the other hand, it provides quantitative, robust analogs of *genericity* results in algebraic settings, which are needed in most algorithmic applications.

Considering the flavor of the algebraic non-degeneracy conditions, the problem of smoothed analysis boils down to the following: *given a matrix $\mathcal{M}$ whose entries are functions (typically polynomials) of some base variables, does randomly perturbing the variables result in $\mathcal{M}$ having a non-negligible least singular value with high probability?*

This question is non-trivial even in very specialized settings, as it is a statement about anti-concentration — a topic that is less understood in probability theory than concentration or large deviation bounds. For example when the underlying variables form a matrix $U \in \mathbb{R}^{n\times m}$, the structured matrix $\mathcal{M} = U \odot U = \left(u_i \otimes u_i\right)_{i\in[m]}$,[1] represents the Khatri-Rao product, and has been the subject of much past work [4, 9, 11] that developed intricate arguments specialized for this setting. Least singular value bounds of $\mathcal{M} = \widetilde{U} \odot \widetilde{U}$ for randomly perturbed $\widetilde{U}$ have lead to smoothed analysis guarantees for several problems including tensor decomposition [9], recovering assemblies of neurons [4], parameter estimation of latent variable models like mixtures of Gaussians [13], hidden Markov models [11], independent component analysis [15] and even learning shallow neural networks [5]. Another approach is to use concentration bounds to prove lower bounds on the least singular value [7, 22, 29**?**] for analyzing random instances; these techniques based on concentration bounds cannot handle smoothed instances. We lack a broader toolkit that allows us to analyze more general classes of random matrices that arise in many other smoothed analysis settings of interest.

Consider, for example the symmetric lift of the matrix $\widetilde{U}$ represented by

$$\widetilde{U}^{\circledast 2} := ((\tilde{u}_i \otimes \tilde{u}_j + \tilde{u}_j \otimes \tilde{u}_i) : 1 \le i \le j \le m),$$

where the columns (up to reshaping) give a basis for the space of all the symmetric matrices that are supported on the subspace $\tilde{U}$. Here $\circledast$ denotes the symmetrized Kronecker product.

Question 1.1. For a linear operator Φ acting on the space of symmetric $n \times n$ matrices (e.g., a projection matrix), can we obtain an inverse polynomial lower bound with high probability on the least singular value of the matrix

$$\mathcal{M} = \Phi(\widetilde{U}^{\circledast 2}) = \left(\Phi(\tilde{u}_i \otimes \tilde{u_j} + \tilde{u}_j \otimes \tilde{u}_i) : 1 \le i \le j \le m\right),$$

when $m \le cn$ for a sufficiently small $c \in (0, 1)$?

The new techniques developed in this paper, to our knowledge, give the first inverse polynomial lower bound on the least singular value of $\mathcal{M}$, and its higher order generalizations; see Theorem 1.4. As it turns out, this already captures the Khatri-Rao product $\widetilde{U} \odot \widetilde{U}$ setting as a special case by setting $m = 1$ and Φ appropriately. One interpretation of the statement is that $\widetilde{U} \circledast \widetilde{U}$ acts like "truly random" subspace in the lifted space $\mathsf{Sym}(\mathbb{R}^n \otimes \mathbb{R}^n)$ with the same dimension. With high probability, a random subspace of $\mathsf{Sym}(\mathbb{R}^n \otimes \mathbb{R}^n)$[2] with dimension $o(n^2)$ will not contain any vector near the kernel of Φ. The affirmative answer to the above question shows that the lifted space that corresponds to column space of $(\widetilde{U})^{\circledast 2}$ behaves similarly and is far from the kernel of Φ! In other words, it is rotationally well-spread; it is not too aligned with any specific subspace. Note that $\widetilde{U}$ only has about nm truly independent coordinates or "bits", whereas a random subspace of the same dimension has $c \cdot n^2m^2$ independent coordinates. Hence the lift $\mathcal{U}^{\circledast 2}$ of a smoothed subspace $\mathcal{U}$ acts "pseudorandom" – it acts like a random subspace in the lifted space with respect to all linear operators of reasonable rank.

Matrices of this flavor arise in open questions about the smoothed analysis of various algebraic algorithms for problems like robust certification of quantum entanglement in subspaces, certifying distance from varieties [19], and decomposition into sums of powers of polynomials [7, 12]. Specifically, rank-1 matrices (of unit norm) correspond to separable or non-entangled states in bipartite quantum systems. For a certain specific choice of Φ, the positive resolution of Question 1.1 certifies that a smoothed subspace of $n_1 \times n_2$ matrices of dimension cn_1n_2 (for some $c > 0$) is far from any rank-1 matrix of unit norm. Moreover, in the recent algebraic algorithms of [7, 12], they consider generic or random subspaces $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_t \subset \mathbb{R}^n$ and they need to argue that the corresponding dth order lifts $\mathcal{U}_1^{\circledast d}, \mathcal{U}_2^{\circledast d}, \ldots, \mathcal{U}_t^{\circledast d}$ are far from each other.

Our results give a novel and modular way to analyze such matrices. Our contributions are two fold:

- We give new tools for proving least singular value lower bounds via ε-nets. This involves identifying a key property that is sufficient for carrying forth net based arguments, and giving a new tool for proving such a property.
- We consider matrices that have the structure of a linear operator applied to higher-order lifts corresponding to the Kronecker product, and give new techniques to reason about the least singular value. This resolves open questions raised in [7, 12, 19].

1.1 Our Results

1.1.1 Hierarchical Nets. Our first set of results focus on ε-net based arguments for proving bounds for least singular values. Suppose we have a random matrix $\mathcal{M}$, the idea is to consider a fixed "test" vector α, prove that $\|\mathcal{M}\alpha\|$ is large enough with high probability, and then take a union bound over "all possible vectors α". As the set of candidate α is infinite, the idea is to take a fine enough net over possible vectors α. The challenge when dealing with structured matrices (of the kind discussed above) is that for a single test vector α, we do not obtain a sufficiently strong probability guarantee. This is because the individual columns of $\mathcal{M}$ may not have "sufficient randomness", and since we do not know how α spreads its mass across columns, the bound will be weak. Our main observation is that in the matrices we consider for our application, as long as α is *well spread*, we can obtain a much stronger bound. We refer to this as a "combination amplifies anticoncentration" (CAA) property of $\mathcal{M}$.

CAA Property (Informal Definition). We say that $\mathcal{M}$ has the CAA property if for every $k \ge 1$, for any test vector α that has k entries of magnitude $\ge \delta$, we have that $\|\mathcal{M}\alpha\| \ge \Omega(\delta)$, with probability $1 - \exp(-\omega(k))$.

[1] Here, $\otimes$ represents the standard tensor product or Kronecker product.
[2] $\mathsf{Sym}(\mathbb{R}^n \otimes \mathbb{R}^n)$ is the space of all symmetric $n \times n$ matrices.

Formally, to capture the $\omega(k)$ term, we have a parameter β. See Definition 4.1 for details. Our first result is that for any matrix with this property, we have a bound on $\sigma_{\min}(\mathcal{M})$.

Informal Theorem 1.2. *Suppose $\mathcal{M}$ is a random matrix with m columns and that $\mathcal{M}$ satisfies the CAA property with parameter $\beta > 0$. Then with high probability (indeed, exponentially small probability of failure), we have $\sigma_{\min}(\mathcal{M}) > \text{poly}(1/m)$. (See Theorem 4.2 for the formal statement.)*

The proof uses a novel ε-net construction. Nets that use structural properties of the test vector α have been used in prior works in the context of proving least singular value bounds, notably in the celebrated work of Rudelson and Vershynin [24]. In proving our result, the natural approach of constructing a hierarchy of nets based on increasing k (and using some threshold δ) does not work. Informally, this is because the error from ignoring terms that are slightly smaller than δ can add up significantly, causing the argument to fail. We introduce a new hierarchical construction that overcomes this problem.

The next question we consider is how to prove that the CAA property holds in a particular context. This can be shown via a direct argument when $\mathcal{M}$ is simple, e.g., a random matrix with independent entries. However, for matrices with more structured entries, it can need a careful analysis. To handle this, we develop a new tool for proving anticoncentration that we believe is of independent interest.

1.1.2 Anti-concentration of a Vector of Polynomials. Consider $P(x) := (p_1(x), p_2(x), \ldots, p_N(x))$, where each p_i is a polynomial of n "base" random variables. Suppose we wish to show anti-concentration bounds for $P(\tilde{x})$, where $\tilde{x}$ is a perturbation of some x (i.e., we wish to bound the probability that $P(\tilde{x})$ is within a small ball of a point y is small, for all y). One hope is to use a coordinate-wise bound (e.g., using known results like [30]) and take the product over $1, 2, \ldots, N$. It is easy to see that this is too good to be true: consider an example where p_i are all equal; here having N coordinates is the same as having just one. So we need a good metric for "how different" the polynomials p_i are for a *typical* x. We capture this notion using the Jacobian of the polynomial map P. Recall that in this case, the Jacobian $J(x)$ is a matrix with one column per p_i, containing the vector of partial derivatives, $\nabla p_i(x)$.

Jacobian rank property (Informal Definition). We say that $P(x)$ has the Jacobian rank property if for every x, at a slightly perturbed point $\tilde{x}$, $J(\tilde{x})$ has at least k singular values that are *large enough* (where k is a parameter).

We refer to Definition B.1 for the formal statement. Our result here is that this property implies anticoncentration:

Informal Theorem 1.3. *Suppose $P(x)$ defined as above satisfies the Jacobian rank property with parameter k. Then for a perturbation of any point x, we have that $\forall y, \mathbb{P}[\|P(\tilde{x}) - y\| < \varepsilon] < \exp(-\Omega(k))$. (Here, ε is a quantity that depends on the dimensions, k, the perturbation, and the singular value guarantee; see Theorem 4.7 for the formal statement.)*

Intuitively, the Jacobian having several large singular values must result in anticoncentration (because $P(x)$ locally behaves linearly). However, the challenging aspect is that the Jacobian need not *always* have many large singular values. Our assumption (Jacobian rank property) is itself made for a perturbed vector, i.e., we assume that $J(\tilde{x})$ has many high singular values with high probability. Further, the magnitude of these singular values will depend on the perturbation: if a "bad" x was perturbed by ρ, $J(\tilde{x})$ will have most of the large singular values being $\approx \rho$. Dealing with this issue turns out to be the main challenge in proving the theorem (see Theorem 4.7 for a formal statement).

As an application of the Jacobian rank method, we re-prove the main result of [9] and [4]. They consider random matrices $\mathcal{M}$ where the ith column is $\tilde{u}_i \otimes \tilde{v}_i$, and $\tilde{u}_i, \tilde{v}_i$ are perturbed vectors in $\mathbb{R}^n$. We show that this $\mathcal{M}$ satisfies the CAA property, and thus our first result (above) implies a condition number lower bound. In order to prove the CAA property, we consider a combination of the columns $\sum_i \alpha_i(\tilde{u}_i \otimes \tilde{v}_i)$ and prove that if α has k entries $\geq \delta$, then the Jacobian has $nk/2$ large singular values. Using our second result, we obtain a strong anticoncentration bound, thus completing the proof. This technique also lets us tackle Question 1.1 described above, but in what follows, we describe a different technique that also generalizes to higher orders.

1.1.3 Structured Matrices from Kronecker Products. Next, we consider a general class of structured matrices that are obtained by taking the symmetrized Kronecker product of some ρ-perturbation $\tilde{U}$ of an underlying matrix U and applying a linear operator Φ. Here, $\tilde{U}$ is a ρ-perturbation of U means $\tilde{U} = U + \mathcal{N}(0, \rho^2)$. In other words, the matrix of interest is $\mathcal{M} = \Phi\widetilde{U}^{\circledast d}$, where d is a constant. For such a matrix, we can ask the question: are there conditions on Φ under which we can prove that $\sigma_{\min}(\mathcal{M})$ is large, with high probability over the perturbation? We provide an affirmative answer to this question in terms of the rank of Φ.

This question captures a variety of settings studied previously. For example, [11] studies matrices $\mathcal{M}$ whose columns are tensor products of some underlying vectors (i.e., the columns have the form $u_{i_1} \otimes u_{i_2} \otimes \cdots \otimes u_{i_d}$). This turns out to be a special case of our setting above. Likewise, in the work of [7], one of the matrices they consider is an $\mathcal{M}$ formed by concatenating the Kronecker products of a collection of underlying matrices, and the analysis of their algorithm relies on $\sigma_{\min}(\mathcal{M})$ being non-negligible. This also falls into our setting by choosing Φ appropriately (as we show in Corollary 5.3). Finally, as we discuss in our applications, the setting $\mathcal{M} = \Phi\widetilde{U}^{\circledast d}$ also directly appears in the work of [19].

The following is an informal statement of our result. $\text{Sym}_d(\mathbb{R}^n)$ will refer to a symmetrization of $(\mathbb{R}^n)^{\otimes d}$.[3] Also, as before, $\sigma_{\min}$ corresponds to right singular vectors.

Informal Theorem 1.4. *Suppose Φ is a matrix of rank $\delta\binom{n+d-1}{d}$ for some constant $\delta > 0$, and let U be any $n \times m$ matrix. Let $\widetilde{U}$ be a ρ-perturbation of U. Then as long as $m \leq cn$ for some constant c, we have $\geq 1 - \exp(-\Omega(n))$,*

$$\sigma_{\min}(\Phi\widetilde{U}^{\circledast d}) \geq \text{poly}\left(\rho, \frac{1}{n}\right).$$

(See Theorem 5.1 for a formal statement.)

[3] The latter can be viewed as having a coordinate for all "ordered" monomials of degree d in n variables (e.g., $x_i x_j$ and $x_j x_i$ correspond to different coordinates), while the former collects the terms with the same product. See Section 3 for a formal description.

Note that the above Theorem 1.4 with $d = 2$ answers Question 1.1 affirmatively. It also proves a similar statement about how the column space of a dth order lift $\widetilde{U}^{\circledast d}$ behaves like a random subspace of the lifted space of the same dimension with respect to linear operators in the lifted space of reasonable rank, even though we have only dnm random "bits" as opposed to $\Omega_d((mn)^d)$. As we describe in Section 2, the proof relies on first moving to non-symmetric products via a new decoupling argument. In the case of non-symmetric products, we end up having to analyze the least singular value of a matrix of the form $\Phi(\widetilde{U}^{(1)} \otimes \widetilde{U}^{(2)} \otimes \cdots \otimes \widetilde{U}^{(d)})$. This can be interpreted as a "modal contraction" (or dimension reduction of the mode) defined by $\{\widetilde{U}^{(i)}\}$ applied to the tensor Φ. We then show how to analyze such *smoothed modal contractions*, which ends up being one of our technical contributions (see Section 2.3 and Theorem 5.2).

1.1.4 *Applications.*

Certifying distance from variety and quantum entanglement. Our first application is to the problem of certifying that a variety is "far" from a *generic* linear subspace. As a simple motivation, suppose we have a linear subspace $\mathcal{X}$ of dimension δn in $\mathbb{R}^n$ (assume $\delta < 1/2$). Then for a randomly ρ-perturbed subspace $\widetilde{\mathcal{U}}$ of dimension $< n/2$, we can show that the two spaces have no overlap in a strong sense: every unit vector $u \in \mathcal{X}$ is at a distance $\Omega(\rho)$ from $\widetilde{\mathcal{U}}$. It is natural to ask if a similar statement holds when $\mathcal{X}$ is an algebraic variety (as opposed to a subspace). This problem also has applications to quantum information (see [19] and references therein). Furthermore, we can ask if there is an efficient algorithm that can *certify* that every unit vector in $\mathcal{X}$ is far from $\widetilde{\mathcal{U}}$.

We answer both these questions in the affirmative.

Informal Theorem 1.5. *Suppose $\mathcal{X} \subset \mathbb{R}^n$ is an irreducible variety cut out by $\delta\binom{n+d-1}{d}$ homogeneous degree d polynomials. There exists a $c > 0$ such that for any ρ-perturbed subspace $\widetilde{\mathcal{U}}$ of dimension at most cn, with probability $1 - \exp(-\Omega(n))$, every unit vector in $\mathcal{X}$ has distance $\geq \mathrm{poly}\left(\rho, \frac{1}{n}\right)$ to $\widetilde{\mathcal{U}}$. Further, this can be certified by an efficient algorithm. (See Theorem D.1 for the formal statement.)*

The recent work of [19] gave an algorithm that we also use, but our new least singular value bounds imply the quantitative distance lower bound stated above. Applying this theorem with the variety of rank-1 matrices gives the following direct corollary.

Corollary 1.6. *There is a polynomial time algorithm that given a random ρ-perturbed subspace $\widetilde{\mathcal{U}}$ of $n_1 \times n_2$ matrices of dimension $m \leq cn_1n_2$ (for some universal constant $c > 0$) certifies w.h.p. that $\widetilde{\mathcal{U}}$ is at least $\mathrm{poly}(\rho, 1/n)$ far from every rank-1 matrix of unit norm.*

The above theorem also has a direct implication to robustly certifying entanglement of different kinds, which we describe in Section D.

Decomposing sums of powers of polynomials. Our second application is to the problem of "decomposing power sums" of polynomials, a question that has applications to learning mixtures of distributions. In the simplest setting, [12] and [7] consider the following problem: given a polynomial $p(\mathbf{x})$ that can be expressed as

$$p(\mathbf{x}) = \sum_{t \in [m]} a_t(\mathbf{x})^3 + e(\mathbf{x})$$

where a_t are quadratic polynomials and $e(\mathbf{x})$ is a small enough error term, the goal is to recover $\{a_t(\mathbf{x})\}_{t\in[m]}$.[4] The work of [7] gave an algorithm for this problem, but their analysis relies on certain *non-degeneracy* conditions, which can be formulated as a lower bound on the least singular value of appropriate matrices. They prove that these conditions hold if the instances (i.e., the polynomials a_t) are *random*, using the machinery of graph matrices [1]. However, the question of obtaining a smoothed analysis guarantee is left open. As discussed earlier, a smoothed analysis guarantee is much stronger than a guarantee for random instances, as it shows that even in the neighborhood of hard instances, most instances are easy.

Their analysis requires least singular value bounds for various matrices that arise from higher order lifts and polynomials of some underlying random variables. For example, they require least singular value bounds on matrices of the form $\Phi(\tilde{U}^{\circledast 3})$, for a specific symmetrization operator Φ that acts on the lifted space. Another type of matrix that they analyze are *block Kronecker products*, of the form $V = [\tilde{U}_1^{\circledast 2} \ldots \tilde{U}_m^{\circledast 2}]$ that arise from different partial derivatives.[5] These kinds of matrices are ideal candidates for our techniques.

Informal Theorem 1.7. *For the matrices $\mathcal{M}$ arising in the analysis of [7], a ρ-perturbation of the parameters of a_t results in $\sigma_{\min}(\mathcal{M}) \geq \mathrm{poly}(\rho, 1/n)$, with probability $1 - \exp(-\mathrm{poly}(m, n))$. (This corresponds to the formal statements of propositions E.1, E.2, and E.3.)*

These least singular bounds allow us to conclude that the algorithm of [7] indeed has a smoothed analysis guarantee. In Section E, we outline the algorithm of [7], identify the different non-degeneracy conditions required and show that each of these conditions holds for *smoothed*/perturbed polynomials a_t. Interestingly, we can avoid the technically heavy machinery of graph matrices, while obtaining stronger (smoothed) results. We hope our new techniques can also help obtain smoothed analysis guarantees for other algebraic methods like the framework of [12].

2 PROOF OVERVIEW AND TECHNIQUES

2.1 Improved Net Analyses

ε-Nets and limitations. The classic approach to proving least singular value bounds is an ε-net argument. The argument proceeds by trying to prove that $\|\mathcal{M}\alpha\|$ is large for all α in the unit sphere. It does so by constructing a fine "net" over points in the sphere with the properties that (a) the net has a small number of points, and hence a union bound can establish the desired bound for points in the net, and (b) for every other point α in the sphere, there is a point α' in the net that is close enough, and hence the bound for α' "translates" to a bound for α. However, in settings where the columns $\widetilde{X}_i$ of $\mathcal{M}$ have "limited randomness", this approach cannot be applied in many parameter regimes of interest. The simplest example is one where each $\widetilde{X}_i$ is of the form $\tilde{u}_i \otimes \tilde{u}_i$, where $\tilde{u}_i \in \mathbb{R}^n$ and we have around $m = n^2/4$ such vectors. In this case, (a) above

[4] This corresponds to the setting $K = 2, D = 1$ in their framework. We focus only on this setting, as it turns out to be representative of their techniques.

[5] The actual matrix is slightly different, and is described in detail in Section E.

causes a problem: the size of a net for unit vectors in a sphere in $\mathbb{R}^m$ is $\exp(m) = \exp(n^2/4)$. This is much too big for applying a union bound, since each column only has "n bits" of randomness, so the failure probability we can obtain for a general α is $\exp(-n)$. For this specific example, the works [4, 9] overcome this limitation by considering more ad-hoc methods for showing least singular value bounds, not based on ε-nets.

Main idea from Section 4.1. As described above, the limited randomness in each column $\widetilde{X}_i$ limits the probability with which we can show that $\mathbb{P}[\|\mathcal{M}\alpha\|]$ is large. However, we observe that in many settings, as long as we consider an α that is *spread out*, we can show that $\mathbb{P}[\|\mathcal{M}\alpha\|]$ is large with a significantly better probability. Informally, in this case, the randomness across many different columns gets "accumulated", thus amplifying the resulting bound. We refer to this phenomenon as *combination amplifies anticoncentration* (CAA) (described informally in Section 1.1; see Definition 4.1). Our first theorem states that the CAA property automatically implies a lower bound on $\sigma_{\min}(\mathcal{M})$ with high probability.

To outline the proof of the theorem, let us consider some unit vector $\alpha \in \mathbb{R}^m$. If α has say $m/2$ "large enough" entries, then the CAA property implies that $\|\mathcal{M}\alpha\|$ is non-negligible with probability $1 - \exp(-m)$ (roughly), and so we can take a union bound over a (standard) ε-net, and we would be done. However, suppose α had only k entries that are large enough (defined as $> \delta$ for some threshold), and $k \ll m$. In this case, the CAA property implies that $\|\mathcal{M}\alpha\| \geq c\delta$ with probability roughly $1 - \exp(-k)$. While this is large enough to allow a union bound over just the large entries of α (placing a zero in the other entries), the problem is that there can be *many* entries in α that are just slightly smaller than δ. In this case, having $\|\mathcal{M}\alpha_{\geq\delta}\| \geq c\delta$ (where $\alpha_{\geq\delta}$ is the vector α restricted to the entries $\geq \delta$ in magnitude, and zeros everywhere else) does not let us conclude that $\|\mathcal{M}\alpha\| > 0$, unless c is very large. Since we cannot ensure that c is large, we need a different argument.

The idea will be to use the fact that our definition of the CAA comes with a slack parameter β. In particular, for α as above with k values of magnitude $\geq \delta$, it allows us to take a union bound over $k \cdot m^\beta$ parameters. Thus, if we knew that there are at most $k \cdot m^\beta$ entries that are "slightly smaller" (by a factor roughly θ) than δ, we can include them in the ε-net. Defining θ appropriately, we can ensure that the problem described above (where the slightly smaller entries cancel out the $\mathcal{M}\alpha_{\geq\delta}$) does not occur. The problem now is when α has $> k \cdot m^\beta$ entries of magnitude between $\theta\delta$ and δ. While this is indeed a problem for this value of δ, it turns out that we can try to work with $\theta\delta$ instead. Now the problem can recur, but it cannot recur more than $(1/\beta)$ times (because each time, k grows by an m^β factor). This allows to define a hierarchical net, which helps us identify the threshold δ for which the ratio of the number of entries $\geq \theta\delta$ and $\geq \delta$ is smaller than m^β.

By carefully bounding the sizes of all the nets and setting θ appropriately, Theorem 4.2 follows.

2.2 Jacobian Based Anticoncentration

As described in Section 1.1, proving smoothed analysis bounds often requires dealing with a vector of polynomials

$$P(x) = (p_1(x), \ldots, p_N(x))$$

in some underlying variables x. The goal is to show that for every x, evaluating P at a ρ-perturbed point $\tilde{x}$ gives a vector that is not too small in magnitude. (A slight generalization is to show that $P(\tilde{x})$ is not too close to any fixed y.)

We first observe that such a statement is not hard to prove if we know that the Jacobian $J(x)$ of $P(x)$ has many large singular values at *every* x, and if the perturbation ρ is small enough. This is because around the given point x, we can consider the linear approximation of $P(\tilde{x})$ given by the Jacobian. Now as long as the perturbation has a high enough projection onto the span of the corresponding singular vectors of $J(x)$, $P(\tilde{x})$ can be shown to have desired anti-concentration properties (by using the standard anticoncentration result for Gaussians). Finally, if $J(x)$ has k large singular values, a random ρ-perturbation will have a large enough projection to the span of the singular vectors with probability $1 - \exp(-k)$.

Now, in the applications we are interested in, the polynomials P tend to have the Jacobian property above for "typical" points x, but not all x. Our main result here is to show that this property suffices. Specifically, suppose we know that for every x, the Jacobian at a ρ perturbed point has k singular values of magnitude $\geq c\rho$ with high probability. Then, in order to show anticoncentration, we view the ρ perturbation of x as occurring in two independent steps: first perturb by $\rho\sqrt{1-z^2}$ for some parameter z, and then perturb by ρz. The key observation is that for Gaussian perturbations, this is identical to a ρ perturbation!

This gives an approach for proving anticoncentration. We use the fact that the first perturbation yields a point with sufficiently many large Jacobian singular values with high probability, and combine this with our earlier result (discussed above) to show that if z is small enough, the linear approximation can indeed be used for the second perturbation, and this yields the desired anticoncentration bound.

Applications. The simplest application for our framework is the setting where $\mathcal{M}$ has columns being $\tilde{u}_i \otimes \tilde{v}_i$, for some ρ-perturbations of underlying vectors u_i, v_i. (This setting was studied in [4, 9] and already had applications to parameter recovery in statistical models.) Here, we can show that $\mathcal{M}$ has the CAA property. To show this, we consider some combination $\sum_i \alpha_i(\tilde{u}_i \otimes \tilde{v}_i)$ with k "large" coefficients in α, and show that in this case, the Jacobian property holds. Specifically, we show that the Jacobian has $\Omega(kn)$ large singular values. This establishes the CAA property, which in turn implies a lower bound on $\sigma_{\min}(\mathcal{M})$. This gives an alternative proof of the results of the works above.

2.3 Structured Matrices from Kronecker Products and Higher-Order Lifts

Our second set of techniques allow us to handle structured matrices that arise from the action of a linear operator on Kronecker products, as described in Question 1.1. For simplicity let us focus on the setting when $d = 2$, and let $\Phi : \mathrm{Sym}(\mathbb{R}^n \otimes \mathbb{R}^n) \to \mathbb{R}^k$ be an (orthogonal) projection matrix of rank $R \geq 0.01n^2$ acting on the space of symmetric matrices $\mathrm{Sym}(\mathbb{R}^n \otimes \mathbb{R}^n)$ (in general Φ can also be any linear operator of large rank). Let $m = o(n)$ and $\widetilde{U} \in \mathbb{R}^{n \times m}$ be a small random ρ-perturbation of arbitrary matrix $U \in \mathbb{R}^{n \times m}$. The columns of the matrix $\widetilde{U}^{\otimes 2}$ are linearly independent with high probability, and span the symmetric lift of the column space of $\widetilde{U}$.

An arbitrary subspace of $\mathrm{Sym}(\mathbb{R}^n \otimes \mathbb{R}^n)$ of the same dimension may intersect non-trivially, or lie close to the kernel of Φ. Theorem 1.4 shows that the column space of $\widetilde{U}^{\otimes 2}$ for a smoothed $\widetilde{U}$ is in fact far from the kernel of Φ with high probability. Note that $\widetilde{U}$ only has about nm truly independent coordinates or "bits", whereas a random subspace (matrix) of the same dimension has $c \cdot n^2m^2$ independent coordinates.

Challenge with existing approaches. This setting captures many kinds of random matrices that have been studied earlier including [4, 9, 11]. For example, [11] studies the setting when a fixed polynomial map $f : \mathbb{R}^n \to \mathbb{R}^k$ applied to a randomly perturbed vector $\tilde{u}_i$ to produce the ith column $f(\tilde{u}_i)$. It turns out to be a special case of our setting above when $m = 1$. These works use the *leave-one-out approach* to lower bound the least singular value, where they establish that every column has a non-negligible component orthogonal to the span of the rest of the columns (see Lemma 3.1). However this approach crucially relies on the columns bringing in independent randomness.[6] This does not hold in our setting, since every column share randomness with $\Omega(m)$ other columns.

In the recent algebraic algorithms of [7, 12] for decomposing sum of powers of polynomials, the analysis of the algorithm involves analyzing the least singular value of different random matrices. One such matrix $\mathcal{M}$ is formed by concatenating the Kronecker products of a collection of underlying matrices. This allows us to reason about that the non-overlap or distance between the lifts of a collection of subspaces. The work of [7] analyzed the *fully random* setting and proves least singular value bounds with intricate arguments involving graph matrices, matrix concentration, and other ideas. Specifically, like in [29], they show that $\mathbb{E}[\mathcal{M}]$ has good least singular value, and then prove deviation bounds on the largest singular value of $\mathcal{M} - \mathbb{E}[\mathcal{M}]$ to get a bound of $\sigma_{\min}(\mathbb{E}[\mathcal{M}]) - \|\mathcal{M} - \mathbb{E}[\mathcal{M}]\|$. But this approach does not extend to the smoothed setting, since the underlying arbitrary matrix U makes it challenging to get good bounds for $\|\mathcal{M} - \mathbb{E}[\mathcal{M}]\|$.

For the smoothed case, when $d = 2$, it turns out that we can use ideas similar to those described in Sections 2.1 and 2.2 to show Theorem 1.4. However, the approach runs into technical issues for larger d. Thus, we develop an alternate technique to analyze higher-order lifts that proves Theorem 1.4 for all constant d. In order to prove Theorem 1.4 we first move to a decoupled setting where we are analyzing the action of a linear operator on decoupled products of the form

$$\Phi(\widetilde{U} \otimes \widetilde{V}),$$

where $\widetilde{V}$ has a random component that is independent of $\widetilde{U}$. This new decoupling step leverages symmetry and the Taylor expansion and carefully groups together terms in a way that decouples the randomness. The main technical statement we prove is the following non-symmetric version of Theorem 1.4 which analyzes a linear operator acting on a Kronecker product of different smoothed matrices.

Informal Theorem 2.1 (Non-symmetric version for $d = 2$ and modal contractions). *Suppose $\Phi \in \mathbb{R}^{R \times n^d}$ is a matrix with at least $\Omega(n^2)$ singular values larger than 1, and let $\widetilde{U}, \widetilde{V}$ be random ρ-perturbations of arbitrary matrices U, V. Then if $m \le cn$ for an appropriate small constant $c > 0$, we have with probability $\ge 1 - \exp(-\Omega(n))$ that*

$$\sigma_{\min}\left(\Phi(\widetilde{U} \otimes \widetilde{V})\right) \ge \mathrm{poly}\left(\rho, \frac{1}{n}\right).$$

(See Theorem 5.2 for the formal statement for general d.)

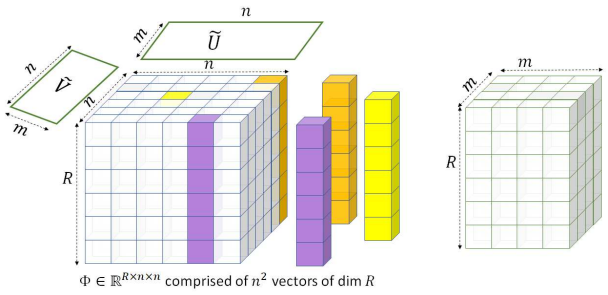

Figure 1: The figure shows the setting of Theorem 2.1 with $d = 2$. *Left*: The linear operator $\Phi : \mathbb{R}^{n \times n} \to \mathbb{R}^R$ interpreted as a tensor consisting of a $n \times n$ array of R-dimensional vectors. There are *smoothed* or random contractions applied using matrices $\widetilde{U}, \widetilde{V} \in \mathbb{R}^{n \times m}$. *Right*: The operator $\Phi(\widetilde{U} \otimes \widetilde{V}) : \mathbb{R}^{m \times m} \to \mathbb{R}^R$ interpreted as an m^2 array of R-dimensional vectors. Theorem 2.1 shows that under the conditions of the theorem, with high probability the robust rank is m^2.

Smoothed modal contractions. While Φ is specified as a linear operator or a matrix of dimension $R \times n^2$ in Theorem 2.1, one can alternately view Φ as a order-3 tensor of dimensions $R \times n \times n$ as shown in Figure 1. Theorem 2.1 then gives a lower bound for the multilinear rank[7] (or its robust analog) under smoothed modal contractions (dimension reduction) along the modes of dimension n each. The proof of this theorem is by induction on the order d. We perform each modal contraction one at a time. As shown in Figure 2, we first do modal contraction by $\widetilde{V}$ to obtain a $R \times n \times m$ tensor W and then by $\widetilde{U}$ to form the final $R \times m \times m$ tensor. We need to argue about the (robust) ranks of the matrix slices (we also call them blocks) and tensors obtained in intermediate steps. For any matrix M (potentially a matrix slice of the tensor Φ) of large (robust) rank $k > 1.1m$, a smoothed contraction $M\widetilde{U}$ has full rank m (i.e., non-negligible least singular value) with probability $1 - \exp(-\Omega(k))$. To argue that the final tensor (when flattened) has full rank m^2, we need to argue that for the tensor in the intermediate step W, each of the m slices (along the contracted mode) has rank at least $\Omega(n)$. The original rank of Φ was large, so we know that a constant fraction of the slices $\Phi_1, \ldots, \Phi_n$ must have rank $\Omega(n)$. But this alone may not be enough since many of the slices can be identical, in which case the m slices are not sufficiently different from each other.

We can use the large rank of Φ to argue that a constant fraction of the matrix slices should have large "marginal rank" i.e., they have large rank even if we project out the column spaces of the slices

[6] The work of [11] also handles some specific settings with a small overlap across columns, but these specialized ideas do not extend more generally to our setting.

[7] The multilinear rank(s) of a tensor is the rank of the matrix after flattening all but one mode of the tensor.

that were chosen before it. While this strategy may work in the non-robust setting, this incurs an exponential blowup in the least singular value. Instead we use the following *randomized* strategy of finding a collection of blocks or slices $S_1 \subset [n]$, each of which has a *large "relative rank"*, even after we project out the column spaces of all the other blocks in S_1 (we show these statements in a robust sense, formalized using appropriate least singular values).

Finding many blocks with large relative rank. We note that while the idea is quite intuitive, the proof of the corresponding claim (Lemma 5.4) is non-trivial because we require that in any selected block, there must be many vectors with a large component orthogonal to the *entire span* of the other selected blocks. As a simple example, consider setting $n_2 = 2t$ and

$$\Phi_1 = \{e_1, e_2, \ldots, e_t, \varepsilon e_{t+1}, \varepsilon e_{t+2}, \ldots, \varepsilon e_{2t}\},$$
$$\text{and } \Phi_2 = \{\varepsilon e_1, \varepsilon e_2, \ldots, \varepsilon e_t, e_{t+1}, e_{t+2}, \ldots, e_{2t}\}.$$

In this case, even if ε is tiny, we cannot choose both the blocks, because the span of the vectors in Φ_2 contains all the vectors in Φ_1.

The proof will proceed by first identifying a set of roughly $R = \Omega(n^2)$ vectors (spread across the blocks) that form a well conditioned matrix, followed by randomly restricting to a subset of the blocks. We start with the following claim, which gives us the first step.

Claim 2.2 (Same as Lemma C.2)**.** *Suppose A is an $m \times n$ matrix such that $\sigma_k(A) \geq \theta$. Then there exists a submatrix A_S with $|S| = k$ columns, such that $\sigma_k(A_S) \geq \theta/\sqrt{nk}$.*

The lemma is a robust version of the simple statement that if $\sigma_k(A) > 0$, then there exist k linearly independent columns. The proof of the claim is elegant and uses the choice of a so-called Auerbach basis or a well-conditioned basis for the column span.

The outline of the main argument is as follows:

(1) First find a submatrix M of $R = \delta n^2$ columns of Φ such that $\sigma_R(M)$ is large
(2) Randomly sample a subset $T \subseteq [n]$ of the blocks.
(3) Discard any block $j \in T$ that has fewer than $\delta n/6$ vectors with a non-negligible component orthogonal to the span of $\cup_{r \in (T \setminus \{j\})} \Phi_r$; argue that there are $\Omega(\delta n)$ blocks remaining.

We remark that the above idea of a random restriction to obtain many blocks with large relative rank (in a robust sense) seems of independent interest and also comes in handy in the application to power sum decompositions (Claim E.5).

Finishing the inductive argument. As shown in Figure 2, after modal contraction along $\tilde{V} \in \mathbb{R}^{n \times m}$, we get $W \in \mathbb{R}^{R \times n \times m}$ with slices $W_1, \ldots, W_n$.

Now we would like to argue that when we perform a smoothed contraction with $\widetilde{U}$, the contracted slices have large rank, while simultaneously preserving the relative rank across the slices. Let $W_{S_1} \in \mathbb{R}^{R \times S_1 \times m}$ represent the subtensor corresponding to the slices obtained from the "good" blocks $S_1 \subset [n]$ (which have large relative rank), and let $W_{[n]\setminus S_1} \in \mathbb{R}^{R \times ([n]\setminus S_1) \times m}$ represent the remaining slices. Also let $W^{(j)} \in \mathbb{R}^{R \times n}$ denote the matrix slices along the alternate mode for each $j \in [m]$. We can show that the randomly contracted matrices $W^{(j)}_{S_1}$ have large relative rank with respect to each other. The random modal contraction $\widetilde{U}$ can also now be

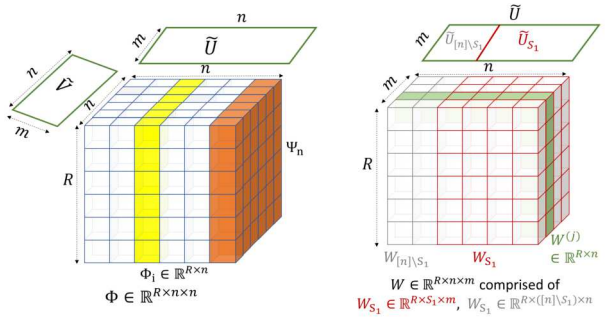

Figure 2: *Left*: The setting of $d = 2$ with linear operator $\Phi : \mathbb{R}^{n \times n} \to \mathbb{R}^R$ having slices $\Phi_1, \ldots, \Phi_n \in \mathbb{R}^{R \times n}$. The modal contractions $\widetilde{U}, \widetilde{V} \in \mathbb{R}^{n \times m}$ have not yet been applied. *Right*: After modal contraction along $\widetilde{V} \in \mathbb{R}^{n \times m}$, we get $W \in \mathbb{R}^{R \times n \times m}$ with slices $W_1, \ldots, W_n$. $W_{S_1} \in \mathbb{R}^{R \times S_1 \times m}$ represents the slices obtained from the "good" blocks $S_1 \subset [n]$, and $W_{[n]\setminus S_1} \in \mathbb{R}^{R \times ([n]\setminus S_1) \times m}$ represents the remaining slices. The random modal contraction $\widetilde{U}$ is also split into $\widetilde{U}_{S_1} \in \mathbb{R}^{S_1 \times m}, \widetilde{U}_{[n]\setminus S_1} \in \mathbb{R}^{[n]\setminus S_1 \times m}$.

split into $\widetilde{U}_{S_1} \in \mathbb{R}^{S_1 \times m}, \widetilde{U}_{[n]\setminus S_1} \in \mathbb{R}^{[n]\setminus S_1 \times m}$. The final matrix slice obtained for each $j \in [m^{d-1}]$ can be written as

$$M^{(j)} = W^{(j)}_{S_1} \widetilde{U}_{S_1} + W^{(j)}_{[n]\setminus S_1} \widetilde{U}_{[n]\setminus S_1},$$

where the randomness in the two summands is independent. Arguing that the high relative rank across the slices is preserved involves some work, and this is achieved in Lemma 5.5. The lemma proves that with high probability, every test unit vector $\alpha \in \mathbb{R}^{m \cdot m}$ has non-negligible value of $\|M\alpha\|_2$. A standard argument would consider a net over all potential unit vectors $\alpha \in \mathbb{R}^{m \cdot m}$. However this approach fails here, since we cannot get high enough concentration (of the form $e^{-\Omega(m^2)}$) that is required for this argument. Instead, we argue that if there were such a test vector $\alpha \in \mathbb{R}^{m \cdot m}$, there exists a block $j^* \in [m]$ where we get a highly unlikely event. This allows us to conclude the inductive proof that establishes Theorem 2.1.

3 PRELIMINARIES

We now introduce our basic definitions and notation. For a matrix $U \in \mathbb{R}^{n \times m}$, let $\|U\|$ and $\|U\|_F$ denote the operator and Frobenius norms of U, respectively. Central to the paper are ρ-smoothed matrices. In particular, given a matrix $U \in \mathbb{R}^{n \times m}$, we let $\tilde{U} = U + E$ where $E \in \mathcal{N}(0, \rho^2)$. We commonly call $\tilde{U}$ a ρ-smoothing of U or a ρ-perturbation of U. Similar notation is used for vector inputs $x = (x_1, \ldots, x_n)$ to a polynomial $p : \mathbb{R}^n \to \mathbb{R}^m$. I.e., $\tilde{x} = x + \eta$ where $\eta \in \mathcal{N}(0, \rho^2)$. Thus, for example, $p(\tilde{x})$ is the evaluation of p on a ρ-smoothed x.

Products. We also frequently use the Kronecker product, denoted $\otimes$, and the Khatri-Rao product, denoted $\odot$. Given matrices, $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{k \times \ell}$, the Kronecker product $A \otimes B$ is the block

matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1m_1}B \\ \vdots & \ddots & \vdots \\ a_{n_1 1}B & \dots & a_{n_1 m_1}B \end{bmatrix} \in R^{nk \times m\ell}.$$

We let $A^{\otimes d} \in R^{n^d \times m^d}$ denote the Kroncker product of a total of d copies of A. In the case that $m = \ell$, the Khatri-Rao product $A \odot B$ is defined by

$$A \odot B = \begin{bmatrix} \uparrow & & \uparrow \\ a_1 \otimes b_1 & \dots & a_m \otimes b_m \\ \downarrow & & \downarrow \end{bmatrix} \in \mathbb{R}^{nk \times m}.$$

Here a_j and b_j denote the jth column of A and B, respectively, and $a_j \otimes b_j$ is the Kronecker product (or simply the tensor product) of these columns.

For vector spaces $\mathcal{U}, \mathcal{V}$, the tensor product space $\mathcal{U} \otimes \mathcal{V} = \{u \otimes v : u \in \mathcal{U}, v \in \mathcal{V}\}$. When $\mathcal{U} = \mathcal{V}$, we also call $\mathcal{U}^{\otimes 2} = \mathcal{U} \otimes \mathcal{U}$ a lift of the space $\mathcal{U}$ (of degree/order 2). This can also be generalized to d-wise products and lifts. When $\mathcal{U} = \mathbb{R}^n$, the space $(\mathbb{R}^n)^{\otimes d}$ corresponds to the space of all d-th order tensors of dimensions $n \times n \cdots \times n$. This is isomorphic to the space $\mathbb{R}^{n^d}$; each tensor can be flattened to form a vector in n^d dimensions i.e., $(\mathbb{R}^n)^{\otimes d} \cong \mathbb{R}^{n^d}$.

Symmetrized products. We are often concerned with symmetrized versions of matrix products. To handle these, we introduce a (partially) symmetrized Kronecker product $\circledast$ which is defined for tuples of matrices $(U^{(1)}, \dots, U^{(d)})$ where $U^{(j)} \in \mathbb{R}^{n_j \times m}$. We define $U^{(1)} \circledast U^{(2)} \circledast \dots \circledast U^{(d)} \in \mathbb{R}^{\Pi_{i=1}^d n_j \times \binom{m+d-1}{d}}$ to be the matrix with columns indexed by tuples $(i_1, i_2, \dots, i_d)$ with $1 \le i_1 \le i_2 \le \dots \le i_d \le m$ where the column corresponding to $(i_1, i_2, \dots, i_d)$ is

$$\frac{1}{|S_d|} \sum_{\pi \in S_d} u^{(1)}_{i_{\pi(1)}} \otimes u^{(2)}_{i_{\pi(2)}} \otimes \dots \otimes u^{(d)}_{i_{\pi(d)}}.$$

Here S_d denotes the symmetric group on $[d]$ and $u^{(j)}_{i_{\pi(j)}}$ denotes the $i_{\pi(j)}$th column of $U^{(j)}$. For example, for matrices $U, V \in \mathbb{R}^{n \times m}$, the column of $U \circledast V$ corresponding to a tuple (i, j) with $i \le j$ is

$$\frac{1}{2}(u_i \otimes v_j + u_j \otimes v_j).$$

In the case that $i = j$, this reduces to $u_i \otimes v_i$. For a matrix $U \in \mathbb{R}^{n \times m}$, we let $U^{\circledast d} \in \mathbb{R}^{n^d \times \binom{m+d-1}{d}}$ denote the $\circledast$ product of a total of d copies of U. The product $\circledast$ can be viewed as a partially symmetrized version of the Kronecker product since all columns of $U^{\circledast d}$ are symmetric with respect to the natural symmetrization of $\mathbb{R}^{n^d} \cong (\mathbb{R}^n)^{\otimes d}$.

Along these lines, we introduce the operator $\mathsf{Sym}_d : \mathbb{R}^{n^d} \to \mathbb{R}^{n^d}$ which symmetrizes elements of $\mathbb{R}^{n^d}$ with respect to the identification $\mathbb{R}^{n^d} \cong (\mathbb{R}^n)^{\otimes d}$. With this notation, we have that

$$\mathsf{Sym}_d(U^{\circledast d}) = U^{\otimes d}.$$

Moreover, the columns of the matrix $U^{\circledast d}$ are precisely the *unique* columns of the matrix $\mathsf{Sym}_d(U^{\otimes d})$.

Finally, for a vector space $\mathcal{U}$, we have that $\mathcal{U}^{\circledast d} = \mathsf{Sym}_d(\mathcal{U}^{\otimes d})$ is the space of symmetric dth tensors over the spacd $\mathcal{U}$. We also call this the symmetric dth order left of the space $\mathcal{U}$.

Leave-one-out distance. The leave-one-out distance of a matrix U is a useful tool for analyzing least singular values. Given $U \in \mathbb{R}^{n \times m}$, define the leave-one-out distance $\ell(U)$ by

$$\ell(U) = \min_i \operatorname{dist}\left(u_i, \operatorname{Span}\{u_j : j \ne i\}\right).$$

The least singular value of U is related to the leave-one-out distance of U through the following lemma [24].

Lemma 3.1 (Leave one out distance)**.** *Let $U \in \mathbb{R}^{n \times m}$. Then*

$$\frac{\ell(U)}{\sqrt{m}} \le \sigma_{\min}(U) \le \ell(U).$$

See also Lemma A.2 for a block-version of leave-one-out singular value bounds.

In our work we also encounter the Jacobian of a polynomial map. Given a vector valued function $P(x) = (p_1(x), p_2(x), \dots, p_N(x))$ over underlying variables $x = (x_1, x_2, \dots, x_n)$, the Jacobian is defined as the $(n \times N)$ matrix of partial derivatives where the (i, j)th entry is $\frac{\partial p_j}{\partial x_i}$. Thus, the linear approximation of $P(x)$ around a point x is simply $P(x + \eta) = P(x) + J(x)^T \eta$.

4 HIERARCHICAL NETS AND ANTICONCENTRATION FROM JACOBIAN CONDITIONING

A complete version of this section, including all deferred proofs, can be found in Appendix B. In this section, we will primarily deal with a matrix $\mathcal{M}$ of dimensions $N \times m$ where $m < N$. The columns will be denoted by $\widetilde{X}_i$, and we wish to show a lower bound on $\sigma_m(\mathcal{M})$.

In this section, we describe the finer ε-net argument outlined in Section 2. We begin with a formal definition of the CAA property.

Definition 4.1 (CAA property)**.** We say that a random matrix $\mathcal{M}$ with m columns has the CAA property with parameter $\beta > 0$, if for all $k \ge 1$, for all test vectors $\alpha \in \mathbb{R}^m$ with at least k coordinates of magnitude δ, there exist $\lambda > 0$ and $c \ge \frac{8}{\beta}$ (dependent only on $\mathcal{M}$) such that

$$\forall h \in (0, 1), \quad \mathbb{P}[\|\mathcal{M}\alpha\| < \delta h/\lambda] \le \exp\left(-c \min(m, km^\beta) \log(1/h)\right).$$

Remark. We note that the condition $c \ge 8/\beta$ may seem strong; however, as we will see in applications, it is satisfied as long as m is small enough compared to N, the number of rows of the matrix.

4.1 Hierarchical Nets

The following shows that the CAA property implies a least singular value guarantee.

Theorem 4.2. *Suppose $\mathcal{M}$ is a random matrix with m columns and that $\mathcal{M}$ satisfies the CAA property with some parameter $\beta > 0$. Suppose additionally that we have the spectral norm bound $\|\mathcal{M}\| \le L$ with probability $1 - \eta$. Then with probability at least $1 - \exp(-m^\beta) - \eta$, we have*

$$\sigma_m(\mathcal{M}) \ge \frac{1}{(Lm\lambda)^{2\lceil \frac{1}{\beta} \rceil}},$$

where λ comes from the CAA property.

As discussed in Section 2, the natural approach to proving such a result would be to take nets based on the sparsity of the test vector α. In other words, if there are k nonzero values of magnitude $\delta > 0$,

the CAA property yields a least singular value lower bound of δ/λ (choosing h to be a small constant), and we can take a union bound over a net of size $\exp(k)$. The issue with this argument is that α might have many other non-zero values that are *slightly* smaller than δ, and these might lead to a zero singular value (unless it so happened that $\lambda < 1/m$, which we do not have a control of). Of course, in this case, we should have worked with a slightly smaller value of δ, but this issue may recur, so we need a more careful argument.

The rest of this subsection will focus on proving Theorem 4.2. For defining the nets, we will use threshold values $\tau_1 = 1/m$, $\tau_2 = \theta/m$, and so on (more generally, $\tau_j = \theta^{j-1}/m$). θ is a parameter that will be chosen appropriately; for now we simply use $\theta \in (0, 1/m)$.

We construct a sequence of nets $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_{s-1}$ as follows. The net $\mathcal{N}_1$ is a set of vectors parametrized by pairs $(r_1, r_2) \in \mathbb{N}^2$, where: (a) $1 \le r_1 \le m^{1-\beta}$, (b) $r_2 \le m^\beta r_1$. For each pair (r_1, r_2), we include all the vectors whose entries are integer multiples of $\frac{\theta}{m}$ with have exactly $(r_1 + r_2)$ non-zero entries, of which r_1 entries are in $(\tau_1, 1]$ and r_2 entries are in $[\tau_2, \tau_1]$. Thus, the number of vectors in $\mathcal{N}_1$ for a single pair (r_1, r_2) is bounded by:

$$\binom{m}{r_1}\binom{m}{r_2}\left(\frac{m}{\theta}\right)^{r_1}\left(\frac{m}{\theta}\right)^{r_2} < \left(\frac{m}{\theta}\right)^{2(r_1+r_2)}.$$

The next net $\mathcal{N}_2$ has vectors parametrized by $(r_1, r_2, r_3) \in \mathbb{N}^3$, where (a) $r_2 \le m^{1-\beta}$, (b) $r_3 \le m^\beta r_2$, and additionally, (c) $r_2 \ge m^\beta r_1$. For each such tuple, we include vectors that have exactly $(r_1+r_2+r_3)$ non-zero entries (in the corresponding τ ranges as above), and have values that are all integer multiples of θ^2/m.

More generally, the vectors of $\mathcal{N}_j$ will be parametrized by $(r_1, r_2, \ldots, r_{j+1}) \in \mathbb{N}^{j+1}$, where (a) $r_j \le m^{1-\beta}$, (b) $r_{j+1} \le m^\beta r_j$, and additionally, (c) for $1 \le i < j$, we have $r_{i+1} > m^\beta r_i$. In other words, r_{j+1} is the first value that does not grow by a factor m^β. For every such tuple, $\mathcal{N}_j$ includes all vectors that have exactly $(r_1 + \cdots + r_{j+1})$ non-zero entries, each of which is an integer multiple of $\frac{\theta^j}{m}$, and exactly r_i of them in the range $(\tau_i, \tau_{i-1}]$ for all $i \le j+1$.

We have nets of this form for $j = 1, 2, \ldots, s-1$, where $s = \lceil \frac{1}{\beta} \rceil$. We now have the following claim.

Claim 4.3. *Fix any $1 \le j < s$. We have*

$$\mathbb{P}\left[\exists \alpha \in \mathcal{N}_j, \|\mathcal{M}\alpha\| < \frac{\theta^{j-\frac{1}{2}}}{m\lambda}\right] < \exp\left(-\tfrac{1}{2}cm^{j\beta}\right).$$

Finally, we have a bigger net for all "dense" vectors α, that have at least $m^{1-\beta}$ coordinates of magnitude $\ge \frac{\theta^{s-1}}{m}$. This net consists of vectors $\in \mathbb{R}^m$ for which (a) every coordinate is an integer multiple of θ^s/m (between 0 and 1), and (b) at least $m^{1-\beta}$ coordinates are $\ge \frac{\theta^{s-1}}{m}$. Call this net $\mathcal{N}_0$. An easy upper bound for the size is

$$|\mathcal{N}_0| \le \left(\frac{m}{\theta^s}\right)^m.$$

Using this, we have the following:

Claim 4.4.

$$\mathbb{P}\left[\exists \alpha \in \mathcal{N}_0 : \|\mathcal{M}\alpha\| < \frac{\theta^{s-\frac{1}{2}}}{m\lambda}\right] < \exp\left(-\frac{c}{2}m\right).$$

One of the advantages of our ε-net argument is that if we only care about "well spread" vectors, we can obtain a much stronger concentration bound (Eq (10)).

Observation 4.5. *Suppose $\mathcal{M}$ is a random matrix that satisfies the CAA property with parameter β. Let us call a test vector α (of length ≤ 1) "dense" if it has at least $m^{1-\beta}$ coordinates of magnitude $> \delta$. Then*

$$\mathbb{P}\left[\exists \text{ dense } \alpha : \|\mathcal{M}\alpha\| < \frac{1}{(Lm\lambda)^{2\lceil \frac{1}{\beta} \rceil}}\right] < \exp\left(-\tfrac{1}{2}cm\right).$$

Note that in the above claim, m could be quite large compared to n. The observation follows immediately from (10), but we will use it later in Section 4.3.

4.2 Anticoncentration of a Vector of Homogeneous Polynomials

We consider the following setting: let $p_1, p_2, \ldots, p_N$ be a collection of homogeneous polynomials over n variables $(x_1, x_2, \ldots, x_n)$, and define

$$P(x) = \begin{bmatrix} p_1(x) \\ p_2(x) \\ \vdots \\ p_N(x) \end{bmatrix} \tag{1}$$

Our goal will be to show anticoncentration results for P. Specifically, we want to prove that $\mathbb{P}[\|P(\tilde{x}) - y\| < \varepsilon]$ is small for all y, where $\tilde{x}$ is a perturbation of some (arbitrary) vector $x \in \mathbb{R}^n$. We give a sufficient condition for proving such a result, in terms of the Jacobian of P. (See Section 3 for background.)

Definition 4.6 (Jacobian rank property)**.** We say that P has the Jacobian rank property with parameters (k, c, γ) if for all $\rho > 0$ and for all x, the matrix $J(\tilde{x})$ has at least k singular values of magnitude $\ge c\rho$, with probability at least $1 - \gamma$. Here, $\tilde{x} = x + \eta$, where $\eta \sim \mathcal{N}(0, \rho^2)$ is a perturbation of the vector x.

Comment. Indeed, all of our results will hold if we only have the required condition for *small enough* perturbations ρ. To keep the results simple, we work with the stronger definition.

For many interesting settings of P, the Jacobian rank property turns out to be quite simple to prove. Our main result now is that the property above implies an anticoncentration bound for P.

Theorem 4.7. *Suppose $P(x)$ defined as above satisfies the Jacobian rank property with parameters (k, c, γ), and suppose further that the Jacobian P' is M-Lipschitz in our domain of interest. Let x be any point and let $\tilde{x}$ be a ρ-perturbation. Then for any $h > 0$, we have*

$$\forall y \in \mathbb{R}^N,\ \mathbb{P}\left[\|P(\tilde{x}) - y\| < \frac{c\rho^2 h}{64Mnk}\right] \le \gamma + \exp(-\frac{1}{4} \cdot k \log(1/h)).$$

A key ingredient in the proof is the following "linearization" based lemma.

Lemma 4.8. *Suppose x is a point at which the Jacobian $J(x)$ of a polynomial P has at least k singular values of magnitude $\ge \tau$. Also suppose that the norm of the Hessian of P is bounded by M in the*

domain of interest. Then, for "small" perturbations, $0 < \rho < \frac{\tau}{4Mnk}$*, we have that for any* $\varepsilon > 0$*,*

$$\forall y,\ \mathbb{P}[\|P(\tilde{x}) - y\| < \varepsilon] < \left(\frac{2\varepsilon}{\tau\rho}\right)^k + \left(\frac{2M\rho nk}{\tau}\right)^{k/2}.$$

We remark that the lemma does not imply Theorem 4.7 directly because it only applies to the case where the perturbation ρ is much smaller than the singular value threshold τ.

4.3 Jacobian Rank Property for Khatri-Rao Products

As the first application, let us use the machinery from the previous sections to prove the following.

Theorem 4.9. *Suppose* $U, V \in \mathbb{R}^{n\times m}$ *and suppose their entries are independently perturbed (by Gaussians* $\mathcal{N}(0,\rho^2)$*) to obtain* $\tilde{U}$ *and* $\tilde{V}$*. Then whenever* $m \le n^2/C$ *for some absolute constant* C*, we have*

$$\sigma_{\min}(\tilde{U} \odot \tilde{V}) \ge \text{poly}\left(\rho, \frac{1}{n}\right),$$

with probability $1 - \exp(-\Omega(n))$*.*

Note that the result is stronger in terms of the success probability than the main result of [9] and matches the result of [4]. The following lemma is the main ingredient of the proof, as it proves the CAA property for $\tilde{U} \odot \tilde{V}$. Theorem 4.9 then follows immediately from Theorem 4.2.

Lemma 4.10. *Suppose* $\alpha \in \mathbb{R}^m$ *be a unit vector at least* k *of whose coordinates have magnitude* $\ge \delta$*. Let* U, V *be arbitrary (as above), and let* $\tilde{U}$ *and* $\tilde{V}$ *be* ρ *perturbations. Define* $P(\tilde{U},\tilde{V}) = \sum_i \alpha_i \tilde{u}_i \otimes \tilde{v}_i$*. Then for* $M = (m+n)^2$ *and all* $h > 0$*, we have*

$$\mathbb{P}\left[\|P(\tilde{U},\tilde{V})\| < \delta h \cdot \frac{\rho^2}{64Mnk}\right] < \exp\left(-\frac{1}{16}kn\log(1/h)\right).$$

Remark. To see why this satisfies the CAA property (hypothesis of Theorem 4.2), note that as long as $m < n^2/C$ for a sufficiently large (absolute) constant C, the term $\frac{kn}{16} \ge 16\min(m, km^{1/2})$, thus it satisfies the condition with $\beta = 1/2$.

The Jacobian property used to show Lemma 4.10 can be extended to higher order Khatri-Rao products. We give details in Section B.3.

5 HIGHER ORDER LIFTS AND STRUCTURED MATRICES FROM KRONECKER PRODUCTS

A complete version of this section, including all deferred proofs can be found in Appendix C. We provide the following theorem.

Theorem 5.1. *Suppose* $d \in \mathbb{N}$*, and let* $\Phi : \mathsf{Sym}^d(\mathbb{R}^n) \to \mathbb{R}^D$ *be an orthogonal projection of rank* $R = \delta\binom{n+d-1}{d}$ *for some constant* $\delta > 0$*, and let* $\mathsf{Sym}_d : (\mathbb{R}^n)^{\otimes d} \to \mathsf{Sym}^d(\mathbb{R}^n)$ *be the orthogonal projection on to the symmetric subspace of* $(\mathbb{R}^n)^{\otimes d}$*. Let* $U = (u_i : i \in [m]) \in \mathbb{R}^{n\times m}$ *be an arbitrary matrix, and let* $\tilde{U}$ *be a random* ρ*-perturbation of* U*. Then there exists a constant* $c_d > 0$ *such that for* $m \le c_d\delta n$*, with probability at least* $1 - \exp\big(-\Omega_{d,\delta}(n)\big)$*, we have the least singular value*

$$\sigma_{\binom{m+d-1}{d}}(\Phi\tilde{U}^{\otimes d}) \ge \frac{\rho^d}{n^{O(d)}},\ \textit{where}$$
$$\tilde{U}^{\otimes d} := \Big(\mathsf{Sym}_d(\tilde{u}_{i_1} \otimes \tilde{u}_{i_2} \cdots \otimes \tilde{u}_{i_d}) : 1 \le i_1 \le i_2 \le \cdots \le i_d \le n\Big). \tag{2}$$

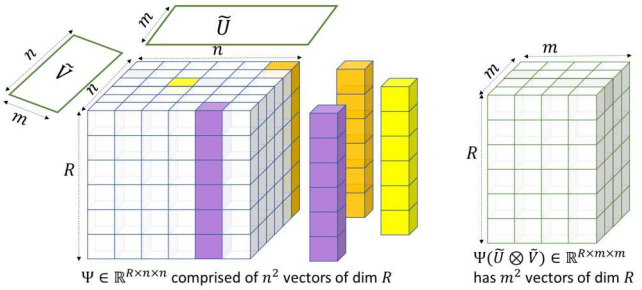

Figure 3: *Left*: The linear operator $\Psi : \mathbb{R}^{n\times n} \to \mathbb{R}^R$ interpreted as a tensor consisting of a $n \times n$ array of R-dimensional vectors. There are *smoothed* or random contractions applied using matrices $\tilde{U}, \tilde{V} \in \mathbb{R}^{n\times m}$. *Right*: The operator $\Psi(\tilde{U}\otimes\tilde{V}) : \mathbb{R}^{m\times m} \to \mathbb{R}^R$ interpreted as an m^2 array of R-dimensional vectors. Theorem 5.2 shows that under the conditions of the theorem, with high probability the robust rank of this operator is m^2 i.e, the least singular value of $R \times m^2$ matrix is inverse polynomial.

In the above statement, one can also consider an arbitrary linear operator Φ and suffer an extra factor of $\sigma_R(\Phi)$ in the least singular value bound (by considering the projector onto the span of the top R singular vectors). In the rest of the section, we assume that Φ is an orthogonal projector of rank R without loss of generality.

Theorem 5.1 follows from the following theorem (Theorem 5.2) which gives a non-symmetric analog of the same statement. The proof of Theorem 5.1 follows from a reduction to Theorem 5.2 that is given by Lemma C.4. In what follows, $\Psi \in \mathbb{R}^{R\times n^d}$ denotes the natural matrix representation of Φ such that $\Psi x^{\otimes d} = \Phi(x^{\otimes d})$ for all $x \in \mathbb{R}^n$.

Theorem 5.2. *Suppose* $\ell \in \mathbb{N}$*,* $R = \delta\binom{n+d-1}{d}$ *for some constant* $\delta > 0$ *and let* $\Psi : (\mathbb{R}^n)^{\otimes\ell} \to \mathbb{R}^D$ *be a linear operator with* $\sigma_R(\Psi) \ge 1$*. Suppose random matrices* $\tilde{U}^{(1)}, \ldots, \tilde{U}^{(d)} \in \mathbb{R}^{n\times m}$ *are generated as follows:*

$$\forall j \in [d],\ \tilde{U}^{(j)} = U^{(j)} + Z^{(j)},\ \textit{where } Z^{(j)} \sim_{i.i.d} \mathcal{N}(0,\rho^2)^{n\times m} \textit{ and is independent of } U^{(j)}, \tag{3}$$

while $U^{(j)} \in \mathbb{R}^{n\times m}$ *is arbitrary and can also depend on* $\tilde{U}^{(j+1)}, \ldots, \tilde{U}^{(d)}$*. Then there exists constants* $c_d, c'_d > 0$ *and an absolute constant* $c_0 \ge 1$ *such that for* $m \le c_d\delta n$*, with probability at least* $1 - \exp\big(-\Omega_{d,\delta}(n)\big)$*, we have*

$$\sigma_{m^d}\Big(\Psi\big(\tilde{U}^{(1)} \otimes \cdots \otimes \tilde{U}^{(d)}\big)\Big) \ge \frac{c'_d\rho^d}{n^{c_0 d}}. \tag{4}$$

While Ψ is specified as a matrix of dimension $R \times n^d$ in Theorem 5.2, one can alternately view Ψ as a $(d+1)$-order tensor of

dimensions $R \times n \times n \times \cdots \times n$ as shown in Figure 4. Theorem 5.2 then gives a lower bound for the multilinear rank (in fact, for its robust analog) under smoothed modal contractions along the d modes of dimension n each.

Applying Theorem 5.1 along with the block leave-one-out approach (see Lemma A.2) we arrive at the following corollary.

Corollary 5.3. *Suppose $d, t \in \mathbb{N}$ and let $1 \geq \delta_1 > \delta_2 > 0$ be given. Also let $\Phi : \mathrm{Sym}^d(\mathbb{R}^n) \to \mathbb{R}^D$ be an orthogonal projection of rank $R \geq \delta_1 \binom{n+d-1}{d}$. Let $\{U_j\}_{j=1}^t \subset \mathbb{R}^{n\times m}$ be an arbitrary collection of $n \times m$ matrices, and for each j, let $\tilde{U}_j$ be a random ρ-perturbation of U_j. Then there exists a constant $c_d > 0$ such that if $t\binom{m+d-1}{d} \leq \delta_2 \binom{n+d-1}{d}$ and $m \leq c_d(\delta_1 - \delta_2)n$, then with probability at least $1 - \exp\left(-\Omega_{d,\delta_1,\delta_2}(n)\right)$, we have the least singular value*

$$\sigma_{t\binom{m+d-1}{d}}\left(\Phi\left[\tilde{U}_1^{\otimes d} \quad \tilde{U}_2^{\otimes d} \quad \ldots \quad \tilde{U}_t^{\otimes d}\right]\right) \geq \frac{\rho^d}{\sqrt{t}n^{O(d)}}. \tag{5}$$

5.1 Proof of Theorem 5.2

We will prove Theorem 5.2 for general d by induction on d. The following crucial lemma considers a linear operator Ψ acting on the space $\mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2}$, and shows that if Ψ has large rank $\Omega(n_1 n_2)$, then it has many "blocks" of large relative rank as described in Section 2.3.

Lemma 5.4. *Let $\Psi \in \mathbb{R}^{R\times(n_1 n_2)}$ be a projection matrix of rank $R = \delta n_1 n_2$ for some constant $\delta > 0$, and let $\Psi = [\Psi_1\ \Psi_2\ \ldots\ \Psi_{n_1}]$ where the blocks $\Psi_i \in \mathbb{R}^{R\times n_2}$ $\forall i \in [n_1]$. Then there exists constants $c_1, c_2, c_3 > 0$ and a subset $S_1 \subset [n_1]$ with $|S_1| \geq c_1\delta n_1$ such that*

$$\forall i \in S_1,\ \sigma_{c_2\delta n_2}\left(\Pi^{\perp}_{S_1\setminus\{i\}}\Psi_i\right) \geq \frac{1}{(nk)^{c_3}}, \tag{6}$$

where $\Pi^{\perp}_S$ is the projection orthogonal to $\mathrm{span}\left(\cup_{i\in S}\mathrm{colspan}(\Psi_i)\right)$.

We note that while the statement of Lemma 5.4 is quite intuitive, the proof is non-trivial because we require that in any selected block, there must be many vectors with a large component orthogonal to the *entire span* of the other selected blocks. We prove this lemma in Section C.2 by restricting to randomly chosen columns as described in the overview (Section 2.3).

The following lemma will be important in the inductive proof of the theorem. It reasons about the robust rank (also called multilinear rank) after the modal contraction by a smoothed matrix along a specific mode. The lemma is proved in slightly more generality; we will use it for the theorem with $\varepsilon = 1$.

Lemma 5.5 (Robust rank under random contractions)**.** *Suppose $\varepsilon \in (0, 1]$ is a constant. For every constant $\gamma, C > 0$, there is a constant $c \in (0, 1)$ such that the following holds for all $s = 2^{o(k)}$. Consider matrices $A_1, A_2, \ldots, A_s \in \mathbb{R}^{R\times k}$, $C_1, \ldots, C_s \in \mathbb{R}^{R\times m}$ and $\forall j \in [s]$ let $\Pi^{\perp}_{-j}$ denote the projector orthogonal to the span of the column spaces of $\{A_{j'} : j' \neq j, j' \in [s]\}$. Suppose the following conditions are satisfied:*

$$\forall j \in [s],\ \sigma_{\varepsilon k}(\Pi^{\perp}_{-j}A_j) \geq k^{-\gamma} \tag{7}$$

and $\sigma_1(A_j), \sigma_1(C_j) \leq k^C$. For a random ρ-perturbed matrix $\tilde{U} \in \mathbb{R}^{k\times m}$ with $m \leq c\varepsilon k$, we have with probability at least $1-\exp(-\Omega(\varepsilon k))$ that

$$\text{if } \forall j \in [s],\ M_j = C_j + A_j\tilde{U},\ \text{then } \sigma_{sm}\Big(M_1 \mid \cdots \mid M_s\Big) \geq \frac{\rho}{2k^{\gamma+1}\sqrt{s}}.$$

Finally, we reduce the the setting of symmetric products to that of non-symmetric products. We provide details in Section C.3.

ACKNOWLEDGMENTS

Vaidehi Srinivas and Aravindan Vijayaraghavan were supported by the National Science Foundation under Grant Nos. CCF-1652491, ECCS-2216970. Vaidehi Srinivas was also supported by the Northwestern Presidential Fellowship. Aditya Bhaskara was supported by the National Science Foundation under Grant Nos. CCF-2008688 and CCF-2047288. We thank the (anonymous) reviewers for their detailed feedback which helped improve the presentation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Kwangjun Ahn, Dhruv Medarametla, and Aaron Potechin. 2016. Graph Matrices: Norm Bounds and Applications. *arXiv: Combinatorics* (2016). https://api.semanticscholar.org/CorpusID:211252816

[2] Anima Anandkumar, Rong Ge, Daniel J. Hsu, Sham M. Kakade, and Matus Telgarsky. 2015. Tensor Decompositions for Learning Latent Variable Models (A Survey for ALT). In *Algorithmic Learning Theory - 26th International Conference, ALT 2015, Banff, AB, Canada, October 4-6, 2015, Proceedings (Lecture Notes in Computer Science, Vol. 9355)*, Kamalika Chaudhuri, Claudio Gentile, and Sandra Zilles (Eds.). Springer, 19–38. https://doi.org/10.1007/978-3-319-24486-0_2

[3] Animashree Anandkumar, Daniel J. Hsu, and Sham M. Kakade. 2012. A Method of Moments for Mixture Models and Hidden Markov Models. In *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland (JMLR Proceedings, Vol. 23)*, Shie Mannor, Nathan Srebro, and Robert C. Williamson (Eds.). JMLR.org, 33.1–33.34. http://proceedings.mlr.press/v23/anandkumar12/anandkumar12.pdf

[4] Nima Anari, Constantinos Daskalakis, Wolfgang Maass, Christos Papadimitriou, Amin Saberi, and Santosh Vempala. 2018. Smoothed Analysis of Discrete Tensor Decomposition and Assemblies of Neurons. In *Advances in Neural Information Processing Systems*.

[5] Pranjal Awasthi, Alex Tang, and Aravindan Vijayaraghavan. 2021. Efficient Algorithms for Learning Depth-2 Neural Networks with General ReLU Activations. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

[6] Baruch Awerbuch and Robert Kleinberg. 2008. Online linear optimization and adaptive routing. *J. Comput. System Sci.* 74, 1 (2008), 97–114.

[7] Mitali Bafna, Jun-Ting Hsieh, Pravesh K. Kothari, and Jeff Xu. 2022. Polynomial-Time Power-Sum Decomposition of Polynomials. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. 956–967. https://doi.org/10.1109/FOCS54457.2022.00094

[8] Boaz Barak, Pravesh K. Kothari, and David Steurer. 2017. Quantum entanglement, sum of squares, and the log rank conjecture. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*.

[9] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. 2014. Smoothed Analysis of Tensor Decompositions. In *Symposium on the Theory of Computing (STOC)*.

[10] Aditya Bhaskara, Moses Charikar, and Aravindan Vijayaraghavan. 2014. Uniqueness of Tensor Decompositions with Applications to Polynomial Identifiability. *Conference on Learning Theory* (2014).

[11] Aditya Bhaskara, Aidao Chen, Aidan Perreault, and Aravindan Vijayaraghavan. 2019. Smoothed Analysis in Unsupervised Learning via Decoupling. In *Proceedings of the 60th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*.

[12] Ankit Garg, Neeraj Kayal, and Chandan Saha. 2020. Learning sums of powers of low-degree polynomials in the non-degenerate case. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*. 889–899. https://doi.org/10.1109/FOCS46700.2020.00087

[13] Rong Ge, Qingqing Huang, and Sham M. Kakade. 2015. Learning Mixtures of Gaussians in High Dimensions. In *Symposium on Theory of Computing*.

[14] Navin Goyal, Santosh Vempala, and Ying Xiao. 2014. Fourier PCA and Robust Tensor Decomposition. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing* (New York, New York) *(STOC '14)*. Association for

Computing Machinery, New York, NY, USA, 584–593. https://doi.org/10.1145/2591796.2591875

[15] Navin Goyal, Santosh Vempala, and Ying Xiao. 2014. Fourier PCA and Robust Tensor Decomposition. In *Symposium on the Theory of Computing (STOC)* (New York, New York) *(STOC '14)*. Association for Computing Machinery, New York, NY, USA, 584–593. https://doi.org/10.1145/2591796.2591875

[16] Venkatesan Guruswami and Ali Kemal Sinop. 2012. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, Yuval Rabani (Ed.). SIAM, 1207–1214. https://doi.org/10.1137/1.9781611973099.95

[17] Aram W. Harrow and Ashley Montanaro. 2010. An Efficient Test for Product States with Applications to Quantum Merlin-Arthur Games. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 633–642. https://doi.org/10.1109/FOCS.2010.66

[18] Elad Hazan and Zohar Karnin. 2016. Volumetric spanners: an efficient exploration basis for learning. *Journal of Machine Learning Research* (2016).

[19] Nathaniel Johnston, Benjamin Lovitz, and Aravindan Vijayaraghavan. 2023. Computing linear sections of varieties: quantum entanglement, tensor decompositions and beyond. In *Proceedings of the IEEE conference on the Foundations of Computer Science (FOCS)*.

[20] J. Lindenstrauss and L. Tzafriri. 2013. *Classical Banach Spaces I: Sequence Spaces*. Springer Berlin Heidelberg.

[21] Ankur Moitra and Alexander S. Wein. 2019. Spectral methods from tensor networks. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, Moses Charikar and Edith Cohen (Eds.). ACM, 926–937. https://doi.org/10.1145/3313276.3316357

[22] Goutham Rajendran and Madhur Tulsiani. [n. d.]. *Concentration of polynomial random matrices via Efron-Stein inequalities*. 3614–3653. https://doi.org/10.1137/1.9781611977554.ch138 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611977554.ch138

[23] Tim Roughgarden. 2020. *Beyond the Worst-Case Analysis of Algorithms*. Cambridge University Press.

[24] Mark Rudelson and Roman Vershynin. 2008. The Littlewood–Offord problem and invertibility of random matrices. *Advances in Mathematics* 218, 2 (2008), 600 – 633. https://doi.org/10.1016/j.aim.2008.01.010

[25] Hanie Sedghi and Anima Anandkumar. 2016. Training Input-Output Recurrent Neural Networks through Spectral Methods. *CoRR* abs/1603.00954 (2016). arXiv:1603.00954 http://arxiv.org/abs/1603.00954

[26] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. 2017. Tensor Decomposition for Signal Processing and Machine Learning. 65, 13 (July 2017), 3551–3582.

[27] Daniel A. Spielman and Shang-Hua Teng. 2004. Smoothed Analysis of Algorithms: Why the Simplex Algorithm Usually Takes Polynomial Time. *J. ACM* 51, 3 (may 2004), 385–463. https://doi.org/10.1145/990308.990310

[28] Shang-Hua Teng. 2023. "Intelligent Heuristics Are the Future of Computing". *ACM Trans. Intell. Syst. Technol.* (oct 2023). https://doi.org/10.1145/3627708 Just Accepted.

[29] Roman Vershynin. 2020. Concentration inequalities for random tensors. *Bernoulli* 26, 4 (2020), 3139 – 3162. https://doi.org/10.3150/20-BEJ1218

[30] Van Vu. 2017. *Anti-concentration Inequalities for Polynomials*. 801–810. https://doi.org/10.1007/978-3-319-44479-6_32