Smoothed Analysis in Unsupervised Learning via Decoupling

Aditya Bhaskara*, Aidao Chen†, Aidan Perreault† and Aravindan Vijayaraghavan†
*School of Computing, University of Utah, Salt Lake City, Utah
Email: bhaskaraaditya@gmail.com
†Department of Computer Science, Northwestern University, Evanston, Illinois
Emails: aidao@u.northwestern.edu, aperreault@u.northwestern.edu, aravindv@northwestern.edu

***Abstract*—Smoothed analysis is a powerful paradigm in overcoming worst-case intractability in unsupervised learning and high-dimensional data analysis. While polynomial time smoothed analysis guarantees have been obtained for worst-case intractable problems like tensor decompositions and learning mixtures of Gaussians, such guarantees have been hard to obtain for several other important problems in unsupervised learning. A core technical challenge in analyzing algorithms is obtaining lower bounds on the least singular value for random matrix ensembles with dependent entries, that are given by low-degree polynomials of a few base underlying random variables.**

In this work, we address this challenge by obtaining high-confidence lower bounds on the least singular value of new classes of structured random matrix ensembles of the above kind. We then use these bounds to design algorithms with polynomial time smoothed analysis guarantees for the following three important problems in unsupervised learning:

- ***Robust subspace recovery*, when the fraction of inliers in the d-dimensional subspace T of the n-dimensional Euclidean space is at least $(d/n)^t$ for any positive integer t. This contrasts with the known worst-case intractability when the fraction of inliers is at most d/n, and the previous smoothed analysis result (Hardt and Moitra, 2013).**
- ***Learning overcomplete hidden markov models*, where the size of the state space is any polynomial in the dimension of the observations. This gives the first polynomial time guarantees for learning overcomplete HMMs in the smoothed analysis model.**
- ***Higher order tensor decompositions*, where we generalize and analyze the so-called FOOBI algorithm of Cardoso to find order-t rank-one tensors in a subspace. This gives polynomially robust decomposition algorithms for order-2t tensors with rank n^t.**

***Index Terms*—smoothed analysis; unsupervised learning; tensor decomposition; subspace recovery; hidden markov model; anti-concentration; beyond worst-case analysis**

I. Introduction

Several basic computational problems in unsupervised learning like learning probabilistic models, clustering and representation learning are intractable in the worst-case. Yet practitioners have had remarkable success in designing heuristics that work well on real-world instances. Bridging this disconnect between theory and practice is a major challenge for many problems in unsupervised learning and high-dimensional data analysis.

The paradigm of Smoothed Analysis [1] has proven to be a promising avenue when the algorithm has only a few isolated bad instances. Given *any instance* from the whole problem space (potentially the worst input), smoothed analysis gives good guarantees for most instances in a small neighborhood around it; this is formalized by small random perturbations of worst-case inputs. This powerful beyond worst-case paradigm has been used to analyze the simplex algorithm for solving linear programs [1], linear binary optimization problems like knapsack and bin packing [2], multi-objective optimization [3], local max-cut [4], [5], and supervised learning [6]. Smoothed analysis gives an elegant way of interpolating between traditional average-case analysis and worst-case analysis by varying the size of the random perturbations.

In recent years, smoothed analysis has been particularly useful in unsupervised learning and high-dimensional data analysis, where the hard instances often correspond to adversarial degenerate configurations. For instance, consider the problem of finding a low-rank decomposition of an order-ℓ tensor that can be expressed as $T \approx \sum_{i=1}^{k} a_i \otimes a_i \otimes \cdots \otimes a_i$. It is NP-hard to find a rank-k decomposition in the worst-case when the rank $k \geq 6n$ [7] (this setting where the rank $k \geq n$ is called the *overcomplete* setting). On the other hand, when the factors of the tensor $\{a_i\}_{i\in[k]}$ are perturbed with some small amount of random Gaussian noise, there exist polynomial time algorithms that can successfully find a rank-k decomposition with high probability even when the rank is $k = O(n^{\lfloor(\ell-1)/2\rfloor})$ [8]. Similarly,

2575-8454/19/$31.00 ©2019 IEEE
DOI 10.1109/FOCS.2019.00043

parameter estimation for basic latent variable models like mixtures of spherical Gaussians has exponential sample complexity in the worst case [9]; yet, polynomial time guarantees can be obtained using smoothed analysis, where the parameters (e.g., means for Gaussians) are randomly perturbed in high dimensions [8], [10]–[12].[1] Smoothed analysis results have also been obtained for other problems like overcomplete ICA [13], learning mixtures of general Gaussians [12], fourth-order tensor decompositions [14], and recovering assemblies of neurons [15].

The technical core of many of the above smoothed analysis results involves analyzing the minimum singular value of certain carefully constructed random matrices with dependent entries. Let $\{\widetilde{a}_1, \widetilde{a}_2, \ldots, \widetilde{a}_k\}$ be random (Gaussian) perturbations of the points $\{a_1, \ldots, a_k\} \subset \mathbb{R}^n$ (think of the average length of the perturbation to be $\rho = 1/\mathrm{poly}(n)$). Typically, these correspond to the unknown parameters of the probabilistic model that we are trying to learn. Proving polynomial smoothed complexity bounds often boils down to proving an inverse polynomial lower bound on the least singular value of certain matrices (that depend on the algorithm), where every entry is a multivariate polynomial involving some of the perturbed vectors $\{\widetilde{a}_1, \ldots, \widetilde{a}_k\}$. These bounds need to hold with a sufficiently small failure probability over the randomness in the perturbations.

Let us now consider some examples to give a flavor of the statements that arise in applications.

- In learning mixtures of spherical Gaussians via tensor decomposition, the key matrix that arises is the "product of means" matrix, in which the number of columns is k, the number of components in the mixture, and the ith column is the flattened tensor $\widetilde{a}_i^{\otimes \ell}$, where $\widetilde{a}_i$ is the mean of the ith component.
- In the so-called FOOBI algorithm for tensor decomposition (proposed by [16], which we will study later), the complexity as well as correctness of the algorithm depend on a special matrix Φ being well conditioned. Φ has the following form: each column corresponds to a pair of indices $i, j \in [k]$, and the (i,j)th column is $\widetilde{a}_i^{\otimes 2} \otimes \widetilde{a}_j^{\otimes 2} - (\widetilde{a}_i \otimes \widetilde{a}_j)^{\otimes 2}$.
- In learning hidden Markov models (HMMs), the matrix of interest is one in which each column is a sum of appropriate monomials of the form $\widetilde{a}_{i_1} \otimes \widetilde{a}_{i_2} \otimes \ldots \otimes \widetilde{a}_{i_\ell}$, where $i_1 i_2 \ldots i_\ell$ correspond to *length-ℓ paths* in the graph being learned.

For many of the recent algorithms based on spectral and tensor decomposition methods (e.g., ones in [17], [18]), one can write down matrices whose condition numbers determine the performance of the corresponding algorithms (in terms of running time, error tolerance etc.). While there is a general theory with broadly applicable techniques (sophisticated concentration bounds) to derive high confidence upper bounds on the maximum singular value of such dependent random matrix ensembles, there are comparatively fewer general tools for establishing lower bounds on the minimum singular value (this has more of an "anti-concentration" flavor), except in a few special cases such as tensor decompositions (using ideas like partitioning co-ordinates).

The high level question that motivates this paper is the following: *can we obtain a general characterization of when such matrices have a polynomial condition number with high probability?* For instance, in the first example, we may expect that as long as $k < \binom{n+\ell-1}{\ell}$, the matrix has an inverse polynomial condition number (note that this is $\ll n^\ell$ due to the symmetries).

There are two general approaches to the question above. The first is a characterization that follows from results in algebraic geometry (see [17], [19]). These results state that the matrix of polynomials either has a sub-matrix whose determinant is the identically zero polynomial, or that the matrix is *generically* full rank. This means that the set of $\{\widetilde{a}_i\}$ that result in the matrix having $\sigma_{\min} = 0$ has measure zero. However, note that this characterization is far from being quantitative. For polynomial time algorithms, we typically need $\sigma_{\min} \geq 1/\mathrm{poly}(n)$ with high probability (this is because polynomial sample complexity often requires these algorithms to be robust to inverse polynomial error). A second approach is via known anti-concentration inequalities for polynomials (such as the Carbery-Wright inequality [20]). In certain settings, these can be used to prove that each column must have at least a small non-zero component orthogonal to the span of the other columns (which would imply a lower bound on $\sigma_{\min}$). However, it is difficult to use this approach to obtain strong enough probability guarantees for the condition number.

Our main contributions are twofold. The first is to prove lower bounds on the least singular value for some broad classes of random matrix ensembles where the entries are low-degree multivariate polynomials of the entries of a given set of randomly perturbed vectors.

[1] In many unsupervised learning problems, the random perturbation to the parameters can not be simulated by perturbations to the input (i.e., samples from the mixture). Hence unlike binary linear optimization [2], such smoothed analysis settings in learning are not limited by known NP-hardness and hardness of approximation results.

The technical difficulty arises due to the correlations in the perturbations (as different matrix entries could be polynomials of the same "base" variables). We note that even in the absence of correlations, (i.e., if the entries are perturbed independently), analyzing the least singular value is non-trivial and has been studied extensively in random matrix theory (see [21], [22]).

Our second contribution is to leverage these results and prove new smoothed analysis guarantees for learning overcomplete hidden markov models, and design algorithms with improved bounds for overcomplete tensor decompositions and for robust subspace recovery.

II. Our Results and Techniques

A. Lower bounds on the Least Singular Value.

The first setting we consider is a simple yet natural one. Suppose we have k independently perturbed vectors $\widetilde{a}_1, \ldots, \widetilde{a}_k$, and suppose we have a matrix in which each column is a fixed polynomial function of precisely one of the variables. We give a sufficient condition under which σ_k (kth largest singular value, or the least singular value here since there are only k columns) of this matrix is at least inverse polynomial with high probability.

Theorem II.1. *Let $\ell \in \mathbb{Z}_+$ be a constant and let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a map defined by m homogeneous polynomials $\{f_i\}_{i=1}^m$ of degree ℓ. Suppose that*

$$f_i(x) = \sum_{\substack{J=(j_1,\ldots,j_\ell)\in[n]^\ell \\ j_1\le j_2\le\cdots\le j_\ell}} U_i(j_1,\ldots,j_\ell) x_{j_1}x_{j_2}\ldots x_{j_\ell},$$

and let $U \in \mathbb{R}^{m\times\binom{n+\ell-1}{\ell}}$ denote the matrix of coefficients, with i^{th} row U_i corresponding to f_i. For vectors $a_1, a_2, \ldots, a_k \in \mathbb{R}^n$, let $M_f(a_1, a_2, \ldots, a_k)$ denote the $m \times k$ matrix whose (i,j)th entry is $f_i(a_j)$. Then for any set of vectors $\{a_i\}_{i=1}^k$, with probability at least $1 - k\exp\big(-\Omega_\ell(\delta n)\big)$,

$$\sigma_k\Big(M_f(\widetilde{a}_1,\ldots,\widetilde{a}_k)\Big) \ge \frac{\Omega_\ell(1)}{\sqrt{k}}\Big(\frac{\rho}{n}\Big)^\ell \cdot \sigma_{k+\delta\binom{n+\ell-1}{\ell}}(U), \tag{1}$$

where $\widetilde{a_j}$ represents a random perturbation of a_j with independent Gaussian noise $N(0,\rho^2/n)^n$.

To obtain a non-trivial bound, note that we need $\sigma_{k+\delta\binom{n+\ell-1}{\ell}}(U) > 0$. Qualitatively, $\sigma_k(U)$ being > 0 is unavoidable. But more interestingly, we will see that the second term is also necessary. In particular, we demonstrate that $\Omega(\delta n^\ell)$ non-trivial singular values are necessary for the required concentration bounds even when $k = 1$ (see Proposition IV.13 for details). In this sense, Theorem II.1 gives an almost tight condition for the least singular value of the above random matrix ensemble to be non-negligible.

For an illustration of Theorem II.1, consider the simple vector-valued polynomial function $f(x) = x^{\otimes\ell} \in \mathbb{R}^{n^\ell}$ (the associated matrix U essentially just corresponds to the identity matrix, with some repeated rows). If $\widetilde{a}_1, \ldots, \widetilde{a}_k \in \mathbb{R}^n$ are randomly perturbed vectors, the above theorem shows that the least singular value of the matrix $M_f(\widetilde{a}_1, \ldots, \widetilde{a}_k)$ is inverse polynomial with exponentially small failure probability, as long as $k \le (1 - o(1))\binom{n+\ell-1}{\ell}$ (earlier results only establish this when k is smaller by a $\exp(\ell)$ factor, because of partitioning co-ordinates). In fact, the above example will be crucial to derive improved smoothed polynomial time guarantees for robust subspace recovery even in the presence of errors (Theorem VI.1).

The next setting we consider is one where the jth column of M does not depend solely on $\widetilde{a}_j$, but on a small subset of the columns in $\{\widetilde{a}_1, \ldots, \widetilde{a}_k\}$ in a structured form. Specifically, in the random matrix ensembles that we consider, each of the R columns of the matrix depends on a few of the vectors in $a_1, \ldots, a_k$ as a "monomial" in terms of tensor products i.e., each column is of the form $u_1 \otimes u_2 \otimes \cdots \otimes u_\ell$ where $u_1, u_2, \ldots, u_\ell \in \{\widetilde{a}_1, \ldots, \widetilde{a}_k\}$. To describe our result here, we need some notation. For two monomials $u_1 \otimes \cdots \otimes u_\ell$ and $v_1 \otimes \cdots \otimes v_\ell$, we say that they disagree in s positions if $u_i \ne v_i$ for exactly s different $i \in [\ell]$. For a fixed column $j \in [R], s \in \{0, 1, \ldots, \ell\}$, let $\Delta_s(j)$ represent the number of other columns whose monomial disagrees with that of column j in exactly s positions, and let $\Delta_s = \max_{j\in[R]} \Delta_s(j)$. (Note that $\Delta_0 = 0$ and $\Delta_\ell \le R$ by default).

Theorem II.2. *Let $\{\widetilde{a}_1, \ldots, \widetilde{a}_k\} \subseteq \mathbb{R}^n$ be a set of ρ-perturbed vectors, let $\ell \in \mathbb{Z}_+$ be a constant, and let $M \in \mathbb{R}^{n^\ell \times R}$ be a matrix whose columns $M_1, \ldots, M_R$ are tensor monomials in $\{\widetilde{a}_i\}_{i\in[k]}$. Let Δ_s be as above for $s = 1, \ldots, \ell$. If*

$$\sum_{s=1}^{\ell} \Delta_s \cdot \Big(\frac{n}{\ell}\Big)^{\ell-s} \le c\Big(\frac{n}{\ell}\Big)^\ell \tag{2}$$

for some $c \in (0,1)$, then $\sigma_R(M) > \Omega_\ell(1) \cdot (\rho/n)^\ell/\sqrt{R}$ with probability at least $1 - \exp\big(-\Omega_\ell(1-c)n + \log R\big)$.

The above statement will be useful in obtaining smoothed polynomial time guarantees for learning overcomplete hidden markov models (Theorem II.6), and for higher order generalizations of the FOOBI algorithm of [16] that gives improved tensor decomposition algorithms up to rank $k = n^{\lfloor \ell/2 \rfloor}$ for order ℓ tensors

(Theorem II.7). In both these applications, the matrix of interest (call it M') is not a monomial matrix per se, but we express its columns as linear combinations of columns of an appropriate monomial matrix M. Specifically, it turns out that $M' = MP$, and P has full column rank (in a robust sense). For example, in the case of overcomplete HMMs, each column of M' is a sum of monomial terms of the form $\widetilde{a}_{i_1} \otimes \widetilde{a}_{i_2} \otimes \ldots \otimes \widetilde{a}_{i_\ell}$, where $i_1 i_2 \ldots i_\ell$ correspond to *length-ℓ paths* in the graph being learned. Each term corresponding to a length-ℓ path only shares dependencies with other paths that share a vertex.

a) Failure probability.: The theorems above emphasize the dependence on the failure probability. We ensure that the claimed lower bounds on $\sigma_{\min}$ hold with a sufficiently small failure probability, say $n^{-\omega(1)}$ or even exponentially small (over the randomness in the perturbations). This is important because in smoothed analysis applications, the failure probability essentially describes the fraction of points around any given point that are *bad* for the algorithm. In many of these applications, the time/sample complexity, or the amount of error tolerance (as in the robust subspace recovery application we will see) has an inverse polynomial dependence on the minimum singular value. Hence, if we have a guarantee that $\sigma_{\min} \geq \gamma$ with probability $\geq 1 - \gamma^{1/2}$ (as is common if we apply methods such as the Carbery-Wright inequality), we have that the probability of the running time exceeds T (upon perturbation) is $\leq 1/\sqrt{T}$. Such a guarantee does not suffice to show that the expected running time is polynomial (also called polynomial smoothed complexity).

1) Techniques: Theorem II.1 crucially relies on the following theorem, which may also be of independent interest.

Informal Theorem II.3. *Let V_ℓ be the space of all symmetric order ℓ tensors in $\mathbb{R}^{n \times n \times \cdots \times n}$, and let $S \subset V_\ell$ be an arbitrary subspace of dimension $(1-\delta)\binom{n+1-\ell}{\ell}$, for some $0 < \delta < 1$. Let $\Pi_S^\perp$ represents the projection matrix onto the subspace of V_ℓ orthogonal to S. Then for any vector x and its ρ-perturbation $\widetilde{x}$, we have that $\|\Pi_S^\perp \widetilde{x}^{\otimes \ell}\|_2 \geq 1/\mathrm{poly}_\ell(n, 1/\rho)$ with probability at least $1 - \exp\big(-\Omega_\ell(\delta n)\big)$.*

The proofs of the theorems above use as a black-box the smoothed analysis result of Bhaskara et al. [8] and the improvements in Anari et al. [15] which shows minimum singular value bounds (with exponentially small failure probability) for tensor products of vectors that have been independently perturbed. Given $\ell \times k$ randomly perturbed vectors $\{\widetilde{a}_i^{(j)} : j \in [\ell], i \in [k]\}$, existing results [8], [15] analyze the minimum singular value of a matrix M where the ith column ($i \in [k]$) is given by $\widetilde{a}_i^{(1)} \otimes \widetilde{a}_i^{(2)} \otimes \cdots \otimes \widetilde{a}_i^{(\ell)}$. However this setting does not suffice for proving Theorem II.1, Theorem II.2, or the different applications presented here because existing results assume the following two conditions:

1) The perturbations to the ℓ factors of the ith column i.e., $a_i^{(1)}, \ldots, a_i^{(\ell)}$ are independent. For proving Theorem II.1 (and for Theorem II.5) we need to analyze symmetric tensor products of the form $\widetilde{x}_i^{\otimes \ell}$, where the perturbations across the factors are the same.
2) Each column of M depends on a disjoint set of vectors $\widetilde{a}_i^{(1)}, \ldots, \widetilde{a}_i^{(\ell)}$, i.e., any vector $\widetilde{a}_i^{(j)}$ is involved in only one column. For proving Theorem II.2 (and later in Theorems II.6 and II.7) however, the same perturbed vector may appear in several columns of M.

Our main tool for proving Theorem II.1, and Theorem II.2 are various decoupling techniques to overcome the dependencies that exists in the randomness for different terms. Decoupling inequalities [23] are often used to prove concentration bounds (bounds on the upper tail) for polynomials of random variables. However, in our case they will be used to establish lower bounds on the minimum singular values. This has an anti-concentration flavor, since we are giving an upper bound on the "small ball probability" i.e., the probability that the minimum singular value is close to a small ball around 0. For Theorem II.1 (and Theorem II.3) which handles symmetric tensor products, we use a combination of asymmetric decoupling along with a positive correlation inequality for polynomials that is inspired by the work of Lovett [24].

We remark that one approach towards proving lower bounds on the least singular value for the random matrix ensembles that we are interested in, is through a direct application of anti-concentration inequalities for low-degree polynomials like the Carbery-Wright inequality (see [11] for smoothed analysis bounds using this approach). Typically this yields an $\varepsilon = 1/\mathrm{poly}(n)$ lower bound on $\sigma_{\min}$ with probability $\varepsilon^{1/\ell}$ (where ℓ is the degree). As we observed above, this cannot lead to polynomial smoothed complexity for many problems.

Interestingly we prove along the way, a vector-valued version of the Carbery-Wright anti-concentration inequality [20], [25] (this essentially corresponds to the special case of Theorem II.1 when $k = 1$). In what follows, we will represent a homogenous degree ℓ multivariate polynomial $g_j : \mathbb{R}^n \to \mathbb{R}$ using the symmetric

tensor M_j of order ℓ such that $g_j(x) = \langle M_j, x^{\otimes \ell} \rangle$ (please see Section III for the formal notation).

Informal Theorem II.4. *Let $\varepsilon, \delta \in (0,1)$, $\eta > 0$, and let $g : \mathbb{R}^n \to \mathbb{R}^m$ be a vector-valued degree ℓ homogenous polynomial of n variables given by $g(x) = (g_1(x), \ldots, g_m(x))$ such that the matrix $M \in \mathbb{R}^{m \times n^\ell}$, with the ith row being formed by the co-efficients of the polynomial g_i, has $\sigma_{\delta n^\ell}(M) \geq \eta$. Then for any fixed $u \in \mathbb{R}^n, t \in \mathbb{R}^m$, and $x \sim N(0, \rho^2/n)^n$ we have*

$$\mathbb{P}\left[\|g(u+x) - t\|_2 < \Omega_\ell(\varepsilon\eta) \cdot \left(\frac{\rho^\ell}{n^\ell}\right)\right] < \varepsilon^{\Omega_\ell(\delta n)}. \quad (3)$$

See Theorem IV.2 for a more formal statement. The main feature of the above result is that while we lose in the "small ball" probability with the degree ℓ, we gain an $m^{\Omega(1)}$ factor in the exponent on account of having a vector valued function. The interesting setting of parameters is when $\ell = O(1), \rho = 1/\mathrm{poly}(n), \varepsilon = \mathrm{poly}_\ell(\rho/n)$ and $\delta = n^{-o(1)}$. We remark that the requirement of δn^ℓ non-trivial singular values is necessary, as described in Proposition IV.13.

The second issue mentioned earlier about [8], [15] is that in many applications each column depends on many of the same underlying few "base" vectors. Theorem II.2 identifies a simple condition in terms of the amount of overlap between different columns that allows us to prove robust linear independence for very different settings like learning overcomplete HMMs and higher order versions of the FOOBI algorithm. Here the decoupling is achieved by building on the ideas in [14], by carefully defining appropriate subspaces where we can apply the existing results on decoupled tensor products [8], [15].

We now describe how the above results give new smoothed analysis results for three different problems in unsupervised learning.

B. Robust Subspace Recovery

Robust subspace recovery is a basic problem in unsupervised learning where we are given m points $x_1, \ldots, x_m \in \mathbb{R}^n$, an $\alpha \in (0,1)$ fraction of which lie on (or close to) a d-dimensional subspace T. When can we find the subspace T, and hence the "inliers", that belong to this subspace? This problem is closely related to designing a robust estimator for subspace recovery: a β-robust estimator for subspace recovery approximately recovers the subspace even when a β fraction of the points are corrupted arbitrarily (think of $\beta = 1 - \alpha$). The largest value of β that an estimator tolerates is called the breakdown point of the estimator. This problem has attracted significant attention in the robust statistics community [26]–[28], yet many of these estimators are not computationally efficient in high dimensions. On the other hand, the singular value decomposition is not robust to outliers. Hardt and Moitra [29] gave the first algorithm for this problem that is both computationally efficient and robust. Their algorithm successfully estimates the subspace T when $\alpha > d/n$, assuming a certain non-degeneracy condition about both the inliers and outliers.[2] This algorithm is also robust to some small amount of noise in each point i.e., the inliers need not lie exactly on the subspace T. They complemented their result with a computational hardness in the worst-case (based on the Small Set Expansion hypothesis) for finding the subspace when $\alpha < d/n$.

We give a simple algorithm that for any constants $\ell \geq 1, \delta > 0$ runs in $\mathrm{poly}(mn^\ell)$ time and in a smoothed analysis setting, provably recovers the subspace T with high probability, when $\alpha \geq (1+\delta)(d/n)^\ell$. Note that this is significantly smaller than the bound of (d/n) from [29] when $\ell > 1$. For instance in the setting when $d = (1-\eta)n$ for some constant $\eta > 0$ (say $\eta = 1/2$), our algorithms recovers the subspace when the fraction of inliers is *any constant* $\alpha > 0$ by choosing $\ell = O(\log(\alpha)/\log(1-\eta))$, while the previous result requires that at least $\alpha > 1 - \eta$ of the points are inliers. On the other hand, when $d/n = n^{-\Omega(1)}$ the algorithm can tolerate any inverse polynomially small α, in polynomial time. In our smoothed analysis setting, each point is given a small random perturbation – each outlier is perturbed with a n-variate Gaussian $N(0, \rho^2)^n$ (think of $\rho = 1/\mathrm{poly}(n)$), and each inlier is perturbed with a projection of a n-variate Gaussian $N(0, \rho^2)^n$ onto the subspace T. Finally, there can be some adversarial noise added to each point (this adversarial noise can in fact depend on the random perturbations).

Informal Theorem II.5. *For any $\delta \in (0,1), \ell \in \mathbb{Z}_+$ and $\rho > 0$. Suppose there are $m = \Omega(n^\ell + d/(\delta\alpha))$ points $x_1, \ldots, x_m \in \mathbb{R}^n$ which are randomly ρ-perturbed according to the smoothed analysis model described above, with an $\alpha \geq (1+\delta)\binom{d+\ell-1}{\ell}/\binom{n+\ell-1}{\ell}$ fraction of the points being inliers, and total adversarial noise $\varepsilon_0 \leq \mathrm{poly}_\ell(\rho/m)$. Then there is an efficient algorithm that returns a subspace T' with $\|sin\Theta(T, T')\|_F \leq \mathrm{poly}_\ell(\varepsilon_0, \rho, 1/m)$ with probability at least $1 - \exp\big(-\Omega_\ell(\delta n) + 2\log m\big) - \exp(-\Omega(d \log m))$.*

See Section VI for a formal statement, algorithm and proof. While the above result gives smoothed analysis

[2]This general position condition holds in a smoothed analysis setting.

guarantees when α is at least $(d/n)^\ell < d/n$, the hardness result of [29] shows that finding a d-dimensional subspace that contains an $\alpha < d/n$ fraction of the points is computationally hard assuming the Small Set Expansion conjecture. Hence our result presents a striking contrast between the intractability result in the worst-case and a computationally efficient algorithm in a smoothed analysis setting when $\alpha > (d/n)^\ell$ for some constant $\ell \geq 1$. Further, we remark that the error tolerance of the algorithm (amount of adversarial error ε_0) does not depend on the failure probability.

a) Techniques and comparisons.: The algorithm for robust subspace recovery at a high level follows the same approach as Hardt and Moitra [29]. Their main insight was that if we sample a set of size slightly less than n from the input, and if the fraction of inliers is $> (1+\delta)d/n$, then there is a good probability of obtaining $> d$ inliers, and thus there exist points that are in the linear span of the others. Further, since we sampled fewer than n points and the outliers are also in general position, one can conclude that the only points that are in the linear span of the other points are the inliers.

Our algorithm for handling smaller α is simple and is also tolerant to an inverse polynomial amount of adversarial noise in the points. Our first observation is that we can use a similar idea of looking for linear dependencies, but with tensored vectors! Let us illustrate in the case $\ell = 2$. Suppose that the fraction of inliers is $> (1+\delta)\binom{d+1}{2}/\binom{n+1}{2}$. Suppose we take a sample of size slightly less than $\binom{n+1}{2}$ points from the input, and consider the flattened vectors $x \otimes x$ of these points. As long as we have more than $\binom{d+1}{2}$ inliers, we expect to find linear dependencies among the tensored inlier vectors. However, we need to account for the adversarial error in the points (this error could depend on the random perturbations as well). For each point, we will look for "bounded" linear combinations that are close to the given point. Using Theorem II.3, we can show that such dependencies cannot involve the outliers. This in turn allows us to recover the subspace even when $\alpha > (d/n)^\ell$ for any constant ℓ in a smoothed analysis sense.

We remark that the earlier least singular value bounds of [8] can be used to show a weaker guarantee about robust linear independence of the matrix formed by columns $\tilde{x}_i^{\otimes \ell}$ with a c^ℓ factor loss in the number of columns (for a constant $c \approx e$). This translates to an improvement over [29] only in the regime when $d < n/c$. Our tight characterization in Theorem II.3 is crucial for our algorithm to beat the d/n threshold of [29] for any dimension $d < n$.

Secondly, if there is no adversarial noise added to the points, it is possible to use weaker concentration bounds (e.g., Carbery-Wright inequality). In this case, our machinery is not required (although to the best of our knowledge, even an algorithm for this noise-free regime with a breakdown point $< d/n$ was not known earlier). In the presence of noise, the weaker concentration inequalities require a noise bound that is tied to the intended failure probability of the algorithm in a strong way. Using Theorem II.3 allows us to achieve a large enough adversarial noise tolerance ε_0, that does not affect the failure probability of the algorithm.

C. Learning Overcomplete Hidden Markov Models

Hidden Markov Models (HMMs) are latent variable models that are extensively used for data with a sequential structure, like reinforcement learning, speech recognition, image classification, bioinformatics etc [30], [31]. In an HMM, there is a hidden state sequence $Z_1, Z_2, \ldots, Z_m$ taking values in $[k]$, that forms a stationary Markov chain $Z_1 \to Z_2 \to \cdots \to Z_m$ with transition matrix P and initial distribution $w = \{w_j\}_{j\in[r]}$ (assumed to be the stationary distribution). The observation X_t is represented by a vector in $x^{(t)} \in \mathbb{R}^n$. Given the state Z_t at time t, X_t (and hence $x^{(t)}$) is conditionally independent of all other observations and states. The matrix $\mathcal{O}$ (of size $n \times r$) represents the probability distribution for the observations: the ith column $\mathcal{O}_i \in \mathbb{R}^n$ represents the expectation of X_t conditioned on the state $Z_t = i$ i.e.

$$\forall i \in [r], t \in [m], \quad \mathbb{E}[X_t | Z_t = i] = \mathcal{O}_i \in \mathbb{R}^n.$$

In an HMM with continuous observations, the distribution of the observation conditioned on state being i can be a Gaussian and ith column of $\mathcal{O}$ would correspond to its mean. In the discrete setting, each column of $\mathcal{O}$ can correspond to the parameters of a discrete distribution over an alphabet of size n.

An important regime for HMMs in the context of many settings in image classification and speech is the *overcomplete setting* where the dimension of the observations n is much smaller than state space r. Many existing algorithms for HMMs are based on tensor decompositions, and work in the regime when $n \leq r$ [18], [32]. In the overcomplete regime, there have been several works [17], [33], [34] that establish identifiability (and identifiability with polynomial samples) under some non-degeneracy assumptions, but obtaining polynomial time algorithms has been particularly challenging in the overcomplete regime. Very recently

Sharan et al. [35] gave a polynomial time algorithm for learning the parameters of an overcomplete discrete HMMs when the observation matrix M is random (and sparse), and the transition matrix P is well-conditioned, under some additional sparsity assumptions on both the transition matrix and observation matrix (e.g., the degree of each node in the transition matrix P is at most $n^{1/c}$ for some large enough constant $c > 1$). Using Theorem II.2, we give a polynomial time algorithm in the more challenging smoothed analysis setting where entries of M are randomly perturbed with small random Gaussian perturbations [3].

Informal Theorem II.6. *Let $\eta, \delta \in (0,1)$ be constants. Suppose we are given a Hidden Markov Model with r states and with $n \geq r^\eta$ dimensional observations with hidden parameters $\widetilde{O}, P$. Suppose the transition matrix P is $d \leq n^{1-\delta}$ sparse (both row and column) and $\sigma_{min}(P) \geq \gamma_1 > 0$, and the each entry of the observation matrix is ρ-randomly perturbed (in a smoothed analysis sense), and the stationary distribution $w \in [0,1]^r$ has $\min_{i \in [r]} w_i \geq \gamma_2 > 0$, then there is a polynomial time algorithm that uses samples of time window $\ell \leq 1/(\eta\delta)$ and recovers the parameters up to ε accuracy (in Frobenius norm) in time $(n/(\rho\gamma_1\gamma_2\varepsilon))^{O(\ell)}$, with probability at least $1 - \exp\big(-\Omega_\ell(n)\big)$.*

For comparison, the result of Sharan et al. [35] applies to discrete HMMs, and gives an algorithm that uses time windows of size $\ell = O(\log_n r)$ in time $\mathrm{poly}(n, r, 1/\varepsilon, 1/\gamma_1, 1/\gamma_2)^\ell$ (there is no extra explicit lower bound on n). But it assumes that the observation matrix $\mathcal{O}$ is fully random, and has other assumptions about sparsity about both $\mathcal{O}$ and P, and about non-existence of short cycles. On the other hand, we can handle the more general smoothed analysis setting for the observation matrix $\mathcal{O}$ for $n = r^\eta$ (for any constant $\eta > 0$), and assume no additional conditions about non-existence of short cycles. To the best of our knowledge, this gives the first polynomial time guarantees in the smoothed analysis setting for learning overcomplete HMMs.

Our results complement the surprising sample complexity lower bound in Sharan et al. [35] who showed that it is statistically impossible to recover the parameters with polynomial samples when $n = \mathrm{polylog}(r)$, even when the observation matrix is random. The algorithm is based on an existing approach using tensor decompositions [8], [17], [18], [35]. The robust analysis of the above algorithm (Theorem II.6) follows by a simple application of Theorem II.2.

[3]While small Gaussian perturbations makes most sense in a continuous observation setting, we believe that these ideas should also imply similar results in the discrete setting for an appropriate smoothed analysis model.

D. Overcomplete Tensor Decompositions

Tensor decomposition has been a crucial tool in many of the recent developments in showing learning guarantees for unsupervised learning problems. The problem here is the following. Suppose $A_1, \ldots, A_R$ are vectors in $\mathbb{R}^n$. Consider the s'th order moment tensor

$$M_s = \sum_{i=1}^{R} A_i^{\otimes s}.$$

The question is if the decomposition $\{A_i\}$ can be recovered given access only to the tensor M_s. This is impossible in general. For instance, with $s = 2$, the A_i can only be recovered up to a rotation. The remarkable result of Kruskal [36] shows that for $s > 2$, the decomposition in "typically" unique, as long as R is a small enough. Several works [14], [16], [18], [37] have designed efficient recovery algorithms in different regimes of R, and assumptions on $\{A_i\}$. The other important question is if the $\{A_i\}$ can be recovered assuming that we only have access to $M_s + \mathsf{Err}$, for some noise tensor Err.

Works inspired by the sum-of-squares hierarchy achieve the best dependence on R (i.e., handle the largest values of R), and also have the best noise tolerance, but require strong incoherence (or even Gaussian) assumptions on the $\{A_i\}$ [38], [39]. Meanwhile, spectral algorithms (such as [8], [13]) achieve a weaker dependence on R and can tolerate a significantly smaller amount of noise, but they allow recoverability for smoothed vectors $\{A_i\}$, which is considerably more general than recoverability for random vectors. The recent work of [14] bridges the two approaches in the case $s = 4$.

Our result here is a decomposition algorithm for 2ℓ'th order tensors that achieves efficient recovery guarantees in the smoothed analysis model, as long as $R \leq cn^\ell$ for a constant c. Our result is based on a generalization of the "FOOBI algorithm" of Cardoso [16], [40], who consider the case $\ell = 2$. We also give a robust analysis of this algorithm (both the FOOBI algorithm for $\ell = 2$, and our generalization to higher ℓ): we show that the algorithm can recover the decomposition to an arbitrary precision ε (up to a permutation), as long as $\|\mathsf{Err}\| \leq \mathrm{poly}_\ell(\varepsilon, 1/n, \rho)$, where ρ is the perturbation parameter in the smoothed analysis model.

Informal Theorem II.7. *Let $\ell \geq 2$ be an integer. Suppose we are given a 2ℓ'th order tensor $T =$*

$\sum_{i=1}^{R} A_i^{\otimes 2\ell} + \mathsf{Err}$*, where* A_i *are* ρ*-perturbations of vectors with polynomially bounded length. Then with probability at least* $1 - \exp(-\Omega_\ell(n))$*, we can find the* A_i *up to any desired accuracy* ε *(up to a permutation), assuming that* $R < cn^\ell$ *for a constant* $c = c(\ell)$*, and* $\|\mathsf{Err}\|_F$ *is a sufficiently small polynomial in* $\varepsilon, \rho, 1/n$*.*

See Theorem VIII.1 and Section VIII for a formal statement and details. We remark that there exists different generalizations of the FOOBI algorithm of Cardoso to higher $\ell > 2$ [41]. However, to the best of our knowledge, there is no analysis known for these algorithms that is robust to inverse polynomial error. Further our new algorithm is a very simple generalization of Cardoso's algorithm to higher ℓ.

This yields an improvement in the best-known dependence on the rank in such a smoothed analysis setting — from $n^{\ell-1}$ (from [8]) to n^ℓ. Previously such results were only known for $\ell = 2$ in [14], who analyzed an SoS-based algorithm that was inspired by the FOOBI algorithm (to the best of our knowledge, their results do not imply a robust analysis of FOOBI). Apart from this quantitative improvement, our result also has a more qualitative contribution: it yields an algorithm for the problem of finding symmetric rank-1 tensors in a linear subspace.

Informal Theorem II.8. *Suppose we are given a basis for an* R *dimensional subspace* S *of* $\mathbb{R}^{n^\ell}$ *that is equal to the span of the flattenings of* $A_1^{\otimes \ell}, A_2^{\otimes \ell}, \ldots A_R^{\otimes \ell}$*, where the* A_i *are unknown* ρ*-perturbed vectors. Then the* A_i *can be recovered in time* $\mathrm{poly}_\ell(n, 1/\rho)$ *with probability at least* $1 - \exp(-\Omega_\ell(n))$*. Further, this is also true if the original basis for* S *is known up to an inverse-polynomial perturbation.*

a) Techniques.: At a technical level, the FOOBI algorithm of [16], [40] for decomposing fourth-order tensors rests on a rank-1 detecting "device" Φ that evaluates to zero if the inputs are a symmetric product vector, and is non-zero otherwise. We construct such a device for general ℓ, and further analyze the condition number of an appropriate matrix that results using Theorem II.2.

We also give an analysis of the robustness of the FOOBI algorithm of [16] and our extension to higher ℓ. While such robustness analyses are often straightforward, and show that each of the terms estimated in the proofs will be approximately preserved. In the case of the FOOBI algorithm, this turns out to be impossible to do (this is perhaps one reason why proving robust guarantees for the FOOBI algorithm even for $\ell = 2$ has been challenging) . The reason is that the algorithm involves finding the top R eigenvectors of the flattened moment matrix, and setting up a linear system of equations in which the coefficients are *non-linear* functions of the entries of the eigenvectors. Now, unless each of the eigenvectors is preserved up to a small error, we cannot conclude that the system of equations that results is close to the one in the noise-free case. Note that for eigenvectors to be preserved approximately after perturbation, we will need to have *sufficient gaps* in the spectrum to begin with. This turns out to be impossible to guarantee using current smoothed analysis techniques. We thus need to develop a better understanding of the solution to the linear system, and eventually argue that even if the system produced is quite different, the solution obtained in the end is close to the original.

III. Preliminaries

In this section, we introduce notation and preliminary results that will be used throughout the rest of the paper.

Given a vector $a \in \mathbb{R}^n$ and a ρ (typically a small inverse polynomial in n), a *ρ-perturbation* of a is obtained by adding independent Gaussian random variables $x_i \sim N(0, \rho^2/n)$ to each coordinate of a. The result of this perturbation is denoted by $\widetilde{a}$.

We will denote the singular values of a matrix M by $\sigma_1(M), \sigma_2(M), \ldots$, in decreasing order. We will usually use k or R to represent the number of columns of the matrix. The maximum and minimum (nonzero) singular values are also sometimes written $\sigma_{max}(M)$ and $\sigma_{min}(M)$.

While estimating the minimum singular value of a matrix can be difficult to do directly, it is closely related to the *leave-one-out distance* of a matrix, which is often much easier to calculate.

Definition III.1. Given a matrix $M \in \mathbb{R}^{n \times k}$ with columns $M_1, \ldots, M_k$, the leave-one-out distance of M is

$$\ell(M) = \min_i dist(M_i, \mathrm{Span}\{M_j : j \neq i\}). \quad (4)$$

The leave-one-out distance is closely related to the minimum singular value, up to a factor polynomial in the number of columns of M [22].

Lemma III.2. *For any matrix* $M \in \mathbb{R}^{n \times k}$*, we have*

$$\frac{\ell(M)}{\sqrt{k}} \leq \sigma_{min}(M) \leq \ell(M). \quad (5)$$

a) Tensors and multivariate polynomials.: An order-ℓ tensor $T \in \mathbb{R}^{n \times n \times \cdots \times n}$ has ℓ modes each of dimension n. Given vectors $u, v \in \mathbb{R}^n$ we will denote by $u \otimes v \in \mathbb{R}^{n \times n}$ the outer product between the vectors

u, v, and by $u^{\otimes \ell}$ the outer product of u with itself ℓ times i.e., $u \otimes u \otimes \cdots \otimes u$.

We will often identify an ℓth order tensor T (with dimension n in each mode) with the vector in $\mathbb{R}^{n^\ell}$ obtained by flattening the tensor into a vector. For sake of convenience, we will sometimes abuse notation (when the context is clear) and use T to represent both the tensor and flattened vector interchangeably. Given two ℓth order tensors T_1, T_2 the inner product $\langle T_1, T_2 \rangle$ denotes the inner product of the corresponding flattened vectors in $\mathbb{R}^{n^\ell}$.

A symmetric tensor T of order ℓ satisfies $T(i_1, i_2, \ldots, i_\ell) = T(i_{\pi(1)}, \ldots, i_{\pi(\ell)})$ for any $i_1, \ldots, i_\ell \in [n]$ and any permutation π of the elements in $[\ell]$. It is easy to see that the set of symmetric tensors is a linear subspace of $\mathbb{R}^{n^{\otimes \ell}}$, and has a dimension equal to $\binom{n+\ell-1}{\ell}$. Given any n-variate degree ℓ homogenous polynomial $g \in \mathbb{R}^n \to \mathbb{R}$, we can associate with g the unique symmetric tensor T of order ℓ such that $g(x) = \langle T, x^{\otimes \ell} \rangle$.

b) Minimum singular value lower bounds for decoupled tensor products.: We will use as a black box high confidence lower bounds on the minimum singular value bounds for decoupled tensor products. The first statement of this form was shown in [8], but this had a worse polynomial dependence on n in both the condition number and the exponent in the failure probability. The following result in [15] gives a more elegant proof, while also achieving much better bounds in both the failure probability and the minimum singular value.

Lemma III.3 ([15], Lemma 6)**.** *Let $p \in (0,1], \delta \in (0,1)$ be constants, and let $W \subseteq \mathbb{R}^{n^{\otimes \ell}}$ be an arbitrary subspace of dimension at least δn^ℓ. Given any $x_1, \cdots, x_\ell \in \mathbb{R}^n$, then for their random perturbations $\tilde{x}_1, \cdots, \tilde{x}_\ell$ where for each $i \in [\ell]$, $\tilde{x}_i = x_i + N(0, \rho_i^2/(2n\ell))^n$ with $\rho_i^2 \geq \rho^2$, we have*

$$\mathbb{P}\Big[\|\Pi_W(\tilde{x}_1 \otimes \cdots \otimes \tilde{x}_\ell)\|_2 < \frac{c_1(\ell)\rho^\ell}{n^\ell} \cdot p^\ell\Big] \leq p^{c_2(\ell)\delta n}$$

where $c_1(\ell), c_2(\ell)$ are constants that depend only on ℓ.

We remark that the statement of Anari et al. [15] is stated in terms of the distance to the orthogonal subspace $W^\perp$, as long as $dim(W^\perp) \leq c^\ell n^\ell$ for some $c < 1$; this holds above for $c = 1 - \delta/\ell$.

IV. Decoupling and Symmetric Tensor Products

In this section we prove Theorem II.1 and related theorems about the least singular value of random matrices in which each column is a function of a single random vector. The proof of Theorem II.1 relies on the following theorem which forms the main technical theorem of this section.

Theorem IV.1 (Same as Theorem II.3)**.** *Let $\delta \in (0,1)$, and let V_ℓ be space of all symmetric order ℓ tensors in $\mathbb{R}^{n \times n \times \cdots \times n}$ (dimension is $D = \binom{n+\ell-1}{\ell}$), and let $W \subset V_\ell$ be an arbitrary subspace of dimension δD. Then we have for any $x \in \mathbb{R}^n$ and $\tilde{x} = x + z$ where $z \sim N(0, \rho^2/n)^n$*

$$\mathop{\mathbb{P}}_z\Big[\|\Pi_W \tilde{x}^{\otimes \ell}\|_2 \geq \frac{c_1(\ell)\rho^\ell}{n^\ell}\Big] \geq 1 - \exp\Big(-c_2(\ell)\delta n\Big),$$

where $c_1(\ell), c_2(\ell)$ are constants that depend only on ℓ.

Theorem II.1 follows by combining the above theorem with an additional lemma that uses a robust version of Sylvester's inequality for products of matrices (see Section IV-C). Our main tool will be the idea of decoupling, along with Lemma III.3 that handles tensor products of vectors that have been perturbed independently. While decoupling inequalities [23] are often used to prove concentration bounds for polynomials of random variables, here this will be used to establish lower bounds on projections and minimum singular values, which have more of an anti-concentration flavor.

In fact we can use the same ideas to prove the following anti-concentration statement that can be seen as a variant of the well-known inequality of Carbery and Wright [20], [25]. In what follows, we will represent a degree ℓ multivariate polynomial $g_j : \mathbb{R}^n \to \mathbb{R}$ using the symmetric tensor M_j of order ℓ such that $g_j(x) = \langle M_j, x^{\otimes \ell} \rangle$.

Theorem IV.2. *Let $\varepsilon, \delta \in (0,1)$, $\eta > 0$ and let $\ell \geq 2$ be an integer. Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be a vector-valued degree ℓ homogenous polynomial of n variables given by $g(x) = (g_1(x), \ldots, g_m(x))$ where for each $j \in [m]$, $g_j(x) = \langle M_j, x^{\otimes \ell} \rangle$ for some symmetric order ℓ tensor $M_j \in \mathbb{R}^{n^\ell}$. Suppose the matrix $M \in \mathbb{R}^{m \times n^\ell}$ formed with the $(M_j : j \in [m])$ as rows has $\sigma_{\delta n^\ell}(M) \geq \eta$, then for any fixed $u \in \mathbb{R}^n, t \in \mathbb{R}^m$, and $z \sim N(0, \rho^2/n)^n$ we have*

$$\mathbb{P}\left[\|g(u+z) - t\|_2 < c(\ell)\varepsilon\eta \cdot \left(\frac{\rho^\ell}{n^\ell}\right)\right] < \varepsilon^{c'(\ell)\delta n} \quad (6)$$

where $c(\ell), c'(\ell) > 0$ are constants that depend only on ℓ.

Remark IV.3. *Comparison to Carbery-Wright inequality:* Anti-concentration inequalities for polynomials are often stated for a single polynomial. They take the following form: if $g : \mathbb{R}^n \to \mathbb{R}$ is a degree-ℓ polynomial with

$\|g\|_2 \geq \eta$, and $x \sim N(0,1)^n$ (or other distributions like the uniform measure on a convex body), then the probability that

$$\mathop{\mathbb{P}}_{x \sim N(0,1)^n} \Big[|g(x) - t| < \varepsilon\eta \Big] \leq O(\ell) \cdot \varepsilon^{1/\ell}.$$

Our statement in Theorem IV.2 applies to *vector valued* polynomials g. Here, if the g_j are "different enough", one can hope that the dependence above becomes $O(\varepsilon^{m/\ell})$, where m is the number of polynomials. Our statement may be viewed as showing a bound that is qualitatively of this kind (albeit with a much weaker dependence on ℓ, when $\ell \geq 2$), when $m \geq \delta n^\ell$. We capture the notion of g_j being different using the condition on the singular value of the matrix M. We also note that the paper of Carbery and Wright [20] does indeed consider vector-valued polynomials, but their focus is on obtaining $\varepsilon^{1/\ell}$ type bounds with a better constant for ε. To the best of our knowledge, none of the known results try to get an *advantage* due to having multiple g_j.

Remark IV.4. While the condition of δn^ℓ non-negligible singular values seems strong, this in fact turns out to be necessary. Proposition IV.13 shows that the relation between the failure probability and the number of non-negligible singular values is tight up to constants that depend only on ℓ. In fact, $m \geq n^{\ell-1}$ is necessary to get any non-trivial bounds. Getting a tight dependence in the exponent in terms of ℓ is an interesting open question.

The main ingredient in the proof of the above theorems is the following decoupling inequality.

Proposition IV.5. *[Anticoncentration through Decoupling] Let $\varepsilon > 0$ and let $\ell \geq 2$ be an integer, and let $\|\cdot\|$ represent any norm over $\mathbb{R}^n$. Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be given by $g(x) = (g_1(x), \ldots, g_m(x))$ where for each $j \in [m]$, $g_j(x) := \langle M_j, x^{\otimes \ell} \rangle$ is a multivariate homogeneous polynomial of degree ℓ, and M_j is a symmetric tensor of order ℓ. For any fixed $u \in \mathbb{R}^n, t \in \mathbb{R}^m$, and $z \sim N(0,\rho^2)^n$ we have*

$$\mathop{\mathbb{P}}_{z} \Big[\|g(u+z) - t\| \leq \varepsilon \Big] \leq$$

$$\leq \mathbb{P} \Big[\|\widehat{g}(u + z_0, z_1, z_2, \ldots, z_{\ell-1})\| \leq \varepsilon/\ell! \Big]^{1/2^{\ell-1}}, \quad (7)$$

where $\forall j \in [m]$, $\widehat{g}_j(u + z_0, z_1, \ldots, z_{\ell-1}) = \langle M_j, (u + z_0) \otimes z_1 \otimes \cdots \otimes z_{\ell-1} \rangle$, $z_0 \sim N(0, \rho^2(\ell+1)/(2\ell))^n$, and $z_1, z_2, \ldots, z_{\ell-1} \sim N(0, \rho^2/(2\ell))^n$.

Note that in the above proposition, the polynomials $\widehat{g}_j(z_0, z_1, \ldots, z_{\ell-1}) = \langle M_j, (u + z_0) \otimes z_1 \otimes \cdots \otimes z_{\ell-1} \rangle$ correspond to decoupled multilinear polynomials of degree ℓ. Unlike standard decoupling statements, here the different components $u + z_0, z_1, \ldots, z_{\ell-1}$ are not identically distributed. We also note that the proposition itself is inspired by a similar lemma in the work of Lovett [24] on an alternate proof of the Carbery-Wright inequality. Indeed the basic inductive structure of our argument is similar (going via Lemma IV.8 below), but the details of the argument turn out to be quite different. In particular we want to consider a random perturbation around an arbitrary point u, and moreover the proposition above deals with vector-valued polyomials g, as opposed to real valued polynomials in [24].

Theorem IV.1 follows by combining Proposition IV.5 and the theorem for decoupled tensor products (Lemma III.3). This will be described in Section IV-B. Later in Section A, we also give an alternate simple proof of Theorem IV.1 for $\ell = 2$ that is more combinatorial. First we introduce the slightly more general setting for decoupling that also captures the required smoothed analysis statement.

A. Proof of Proposition IV.5

We will start with a simple fact involving signed combinations.

Lemma IV.6. *Let $\alpha_0, \alpha_1, \ldots, \alpha_m$ be real numbers, and let $\zeta_1, \zeta_2, \ldots, \zeta_m \in \{\pm 1\}$ be independent Rademacher random variables. Then*

$$\mathop{\mathbb{E}}_{\zeta} \Big[\big(\alpha_0 + \alpha_1\zeta_1 + \cdots + \alpha_m\zeta_m\big)^{m+1} \prod_{i \in [m]} \zeta_i \Big] = \\ = (m+1)! \cdot \alpha_0\alpha_1 \ldots \alpha_m.$$

Proof. For a subset $S \subseteq [m]$, let $\xi(S) = \prod_{i \in S} \zeta_i$. Then it is easy to check that $\mathbb{E}\left[\xi(S) \prod_{i \in [m]} \zeta_i\right] = 0$ if $S \neq \{1, 2, \ldots, m\}$, and 1 if $S = [m]$. Applying this along with the multinomial expansion for $\big(\alpha_0 + \alpha_1\zeta_1 + \cdots + \alpha_m\zeta_m\big)^m$ gives the lemma. □

Lemma IV.7. *Consider any symmetric order ℓ tensor T, a fixed vector $x \in \mathbb{R}^n$, and let $z_1 \sim N(0, \rho_1^2)^n, \ldots, z_\ell \sim N(0, \rho_\ell^2)^n$ be independent random Gaussians. Then we have*

$$\sum_{\zeta_2, \ldots, \zeta_\ell \in \pm 1} \Big(\prod_{i=2}^{\ell} \zeta_i \Big) \langle T, (x + z_1 + \zeta_2 z_2 + \cdots + \zeta_\ell z_\ell)^{\otimes \ell} \rangle = \\ = 2^{\ell-1} \ell! \cdot \langle T, (x + z_1) \otimes z_2 \otimes \cdots \otimes z_\ell \rangle. \quad (8)$$

Note that the right side corresponds to the evaluation of the tensor T at a random perturbation of $(x, 0, 0, \ldots, 0)$.

Proof. First, we observe that since T is symmetric, it follows that $\langle T, u_1 \otimes u_2 \otimes \cdots \otimes u_\ell \rangle = \langle T, u_{\pi(1)} \otimes u_{\pi(2)} \otimes$

$\cdots \otimes u_{\pi(\ell)}\rangle$ for any permutation π on $(1,2,\ldots,\ell)$. Let $u = x+z_1$, and let $\zeta_2,\zeta_3,\ldots,\zeta_\ell \in \{\pm 1\}$ be independent Rademacher random variables. For any symmetric decomposition into rank-one tensors $T = \sum_j \lambda_j v_j^{\otimes \ell}$ (note that such a decomposition always exists for a symmetric tensor; see [42] for example), we have for every $x \in \mathbb{R}^n$, $\langle T, x^{\otimes \ell}\rangle = \sum_j \lambda_j \langle v_j, x\rangle^\ell$. Applying Lemma IV.6 (with $m = \ell - 1$) to each term separately

$$\forall j,\ \mathop{\mathbb{E}}_{\zeta_2,\ldots,\zeta_\ell}\Big[\Big(\prod_{i=2}^{\ell}\zeta_i\Big)\langle v_j^{\otimes \ell}, (u+\zeta_2 z_2+\cdots+\zeta_\ell z_\ell)^{\otimes \ell}\rangle\Big] = \\ = \ell! \cdot \langle v_j^{\otimes \ell}, u\otimes\zeta_2\otimes\cdots\otimes\zeta_\ell\rangle.$$

Combining them, we get

$$\mathop{\mathbb{E}}_{\zeta_2,\zeta_3,\ldots,\zeta_\ell}\Big[\Big(\prod_{i=2}^{\ell}\zeta_i\Big)\langle T, (u+\zeta_2 z_2+\cdots+\zeta_\ell z_\ell)^{\otimes \ell}\rangle\Big] = \\ = \ell! \cdot \langle T, u\otimes z_2\otimes z_3\otimes\cdots\otimes z_\ell\rangle \\ = \ell! \cdot \langle T, (x+z_1)\otimes z_2\otimes z_3\otimes\cdots\otimes z_\ell\rangle.$$

□

Our proof of the anti-concentration statement (Proposition IV.5) will rely on the signed combination of vectors given in Lemma IV.7 and on a positive correlation inequality that is given below.

Lemma IV.8. *Let $z \sim N(0,\rho^2)^n$ be an n-variate Gaussian random variable, and let $z_0 \sim N(0,\rho^2(\ell+1)/(2\ell))^n$ and $z_1,z_2,\ldots,z_{\ell-1} \sim N(0,\rho^2/(2\ell))^n$ be a collection of independent n-variate Gaussian random variables. Then for any measurable set $S \subset \mathbb{R}^n$ we have*

$$\mathop{\mathbb{P}}_{z}\Big[z\in S\Big] \le \mathop{\mathbb{P}}_{z}\Big[\bigwedge_{\zeta_j\in\{\pm1\}}\big(z_0+\textstyle\sum_{j=1}^{\ell-1}\zeta_j z_j\big)\in S\Big]^{1/(2^{\ell-1})} \tag{9}$$

This inequality and its proof are inspired by the work of Lovett [24] mentioned earlier. The main advantage in our inequality is that the right side here involves the particular signed combinations of the function values at $2^{\ell-1}$ points from ℓ independent copies that directly yields the asymmetric decoupled product (using Lemma IV.7).

Proof. Let $x_0,x_1,\ldots,x_{\ell-1} \sim N(0,\rho^2/\ell)^n$, and for each $k \in [\ell-1]$, let $\widehat{y}_k \sim N(0,\rho^2(k+2)/(2\ell))^n$. Clearly $\mathbb{P}[z\in S] = \mathbb{P}[x_0+\cdots+x_{\ell-1}\in S]$. Let $f(z) = \mathbf{1}_{z\in S}$ represent the indicator function of S. For $0 \le k \le \ell-1$, let

$$E_k = \mathop{\mathbb{E}}_{\substack{\widehat{y}_k,z_1,\ldots,z_k,\\ x_{k+1},\ldots,x_{\ell-1}}}\Bigg[\prod_{\zeta_1,\ldots,\zeta_k\in\{\pm1\}} f\Big(\widehat{y}_k + \\ + \sum_{j=1}^{k}\zeta_j z_j + \sum_{j=k+1}^{\ell-1} x_j\Big)\Bigg]$$

We will prove that for each $k \in [\ell-1]$, $E_{k-1}^2 \le E_k$. Using Cauchy-Schwartz inequality, we have

$$E_{k-1}^2 = \Bigg(\mathop{\mathbb{E}}_{\substack{\widehat{y}_{k-1},z\\ x_{k+1},\ldots,x_{\ell-1}}}\mathop{\mathbb{E}}_{x_k}\Big[\prod_{\zeta_1,\ldots,\zeta_{k-1}\in\{\pm1\}} f\Big(\widehat{y}_{k-1}+ \\ + \textstyle\sum_{j=1}^{k-1}\zeta_j z_j + \sum_{j=k}^{\ell-1} x_j\Big)\Big]\Bigg)^2 \\ \le \mathop{\mathbb{E}}_{\substack{\widehat{y}_{k-1},x_{k+1},\ldots,x_{\ell-1}\\ z_1,\ldots,z_{k-1}}}\Bigg(\mathop{\mathbb{E}}_{x_k}\Big[\prod_{\zeta_1,\ldots,\zeta_{k-1}\in\{\pm1\}} f\Big(\widehat{y}_{k-1}+ \\ + \textstyle\sum_{j=1}^{k-1}\zeta_j z_j + \sum_{j=k}^{\ell-1} x_j\Big)\Big]\Bigg)^2.$$

Now if y_k, z_k are i.i.d variables distributed as $N(0,\rho^2/(2\ell))^n$, then x_k, y_k+z_k, y_k-z_k are identically distributed. More crucially, y_k+z_k and y_k-z_k are independent! Hence

$$E_{k-1}^2 \le \mathop{\mathbb{E}}_{\substack{\widehat{y}_{k-1},x_{k+1},\ldots,x_{\ell-1}\\ z_1,\ldots,z_{k-1}}}\Bigg(\mathop{\mathbb{E}}_{y_k,z_k\sim N(0,\frac{\rho^2}{2\ell}))^n}\Bigg[\\ \prod_{\zeta_1,\ldots,\zeta_{k-1}\in\{\pm1\}} f\Big(\widehat{y}_{k-1}+\textstyle\sum_{j=1}^{k-1}\zeta_j z_j + \\ + (y_k+z_k) + \sum_{j=k+1}^{\ell-1} x_j\Big)\Bigg] \\ \times \mathop{\mathbb{E}}_{y_k,z_k\sim N(0,\frac{\rho^2}{2\ell}))^n}\Big[\prod_{\zeta_1,\ldots,\zeta_{k-1}\in\{\pm1\}} f\Big(\widehat{y}_{k-1}+ \\ + \textstyle\sum_{j=1}^{k-1}\zeta_j z_j + (y_k-z_k) + \sum_{j=k+1}^{\ell-1} x_j\Big)\Big]\Bigg) \\ = \mathop{\mathbb{E}}_{\substack{\widehat{y}_{k-1},x_{k+1},\ldots,x_{\ell-1}\\ z_1,\ldots,z_{k-1}}}\mathop{\mathbb{E}}_{y_k,z_k\sim N(0,\frac{\rho^2}{2\ell}))^n}\Bigg[\\ \prod_{\zeta_1,\ldots,\zeta_{k-1}\in\{\pm1\}} f\Big(\widehat{y}_{k-1}+\textstyle\sum_{j=1}^{k-1}\zeta_j z_j + \\ + (y_k+z_k) + \sum_{j=k+1}^{\ell-1} x_j\Big) \\ \times f\Big(\widehat{y}_{k-1}+\textstyle\sum_{j=1}^{k-1}\zeta_j z_j + (y_k-z_k) + \\ + \sum_{j=k+1}^{\ell-1} x_j\Big)\Bigg]$$

$$= \mathop{\mathbb{E}}_{\substack{\widehat{y}_k, x_{k+1},\ldots,x_{\ell-1} \\ z_1,\ldots,z_k}} \Big[\prod_{\zeta_1,\ldots,\zeta_k \in \{\pm 1\}} f\Big(\widehat{y}_k + \\ + \textstyle\sum_{j=1}^{k} \zeta_j z_j + \sum_{j=k+1}^{\ell-1} x_j\Big)\Big],$$

where the last step follows by identifying $\widehat{y}_{k-1} + y_k$ with $\widehat{y}_k$. The proof of the lemma is completed by observing that $E_0 = \mathbb{P}[\widehat{y}_0 + x_1 + \cdots + x_\ell \in S] = \mathbb{P}[z \in S]$. □

We now proceed to the proof of the main decoupling statement.

Proof of Proposition IV.5. Let $S := \{z \in \mathbb{R}^n : \|g(z+u) - t\| \le \varepsilon\}$. Let $z_0 \sim N(0, \rho^2(\ell+1)/(2\ell))^n$ and $z_1, \ldots, z_{\ell-1} \sim N(0, \rho^2/(2\ell))^n$ be independent n-variate Gaussian random variables. From Lemma IV.8 we have for $z \sim N(0, \rho^2)^n$,

$$\begin{aligned}
&\mathop{\mathbb{P}}_{z}\Big[\|g(z+u) - t\| \le \varepsilon\Big] \le \\
&\quad \le \mathop{\mathbb{P}}_{z_0,\ldots,z_{\ell-1}} \Big[\bigwedge_{\zeta_1,\ldots,\zeta_{\ell-1} \in \{\pm 1\}} \Big(\|g(u + z_0 + \\
&\qquad + \textstyle\sum_{j=1}^{\ell-1} \zeta_j z_j) - t\| \le \varepsilon\Big)\Big]^{1/(2^{\ell-1})} \\
&\quad \le \mathop{\mathbb{P}}_{z_0 \ldots z_{\ell-1}} \Big[\sum_{\zeta_1,\ldots,\zeta_{\ell-1} \in \{\pm 1\}} \|g(u + z_0 + \\
&\qquad + \textstyle\sum_{j=1}^{\ell-1} \zeta_j z_j) - t\| \le 2^{\ell-1}\varepsilon\Big]^{1/(2^{\ell-1})} \\
&\quad \le \mathop{\mathbb{P}}_{z_0 \ldots z_{\ell-1}} \Big[\Big\| \sum_{\zeta_1,\ldots,\zeta_{\ell-1} \in \{\pm 1\}} (\prod_{j=1}^{\ell-1} \zeta_j) g\big(u + z_0 + \\
&\qquad + \textstyle\sum_{j=1}^{\ell-1} \zeta_j z_j\big)\Big\| \le 2^{\ell-1}\varepsilon\Big]^{1/(2^{\ell-1})},
\end{aligned}$$

where the last inequality follows from triangle inequality, and observing that the signed combinations of t cancel out when $\ell \ge 2$. Now applying Lemma IV.7 for each $i \in [m]$, we get

$$\begin{aligned}
&\mathop{\mathbb{P}}_{z \sim N(0,\rho^2)^n}\Big[\|g(z+u) - t\| \le \varepsilon\Big] \le \\
&\le \mathop{\mathbb{P}}_{z_0,\ldots,z_{\ell-1}}\Big[\|\widehat{g}(u + z_0, z_1, \ldots, z_{\ell-1})\| \le \varepsilon/\ell!\Big]^{1/(2^{\ell-1})}.
\end{aligned}$$

□

B. Proofs of Theorem IV.1 and Theorem IV.2

Proof of Theorem IV.1. Let $m = \delta D$, and let $M_1, M_2, \ldots, M_m$ be an orthonormal basis of symmetric tensors in $W \subset \mathbb{R}^{n^{\otimes \ell}}$. We will also denote by M the $m \times n^\ell$ matrix formed by flattening $M_1, \ldots, M_m$ respectively. For each $j \in [m]$, let $g_j(x) = \langle M_j, x^{\otimes \ell}\rangle$. Let $\tilde{x} = x + z$ where $z \sim N(0, \rho^2/n)^n$. We would like to lower bound $\|\Pi_W \tilde{x}^{\otimes \ell}\|_2 = \|g(x+z)\|_2$. Using Proposition IV.5 with $t = 0$, for all $\varepsilon > 0$, we have

$$\begin{aligned}
&\mathbb{P}\Big[\|g(x+z)\|_2 < \varepsilon\Big] \le \\
&\le \mathbb{P}\Big[\|\Pi_W(x + z_0) \otimes z_1 \otimes \cdots \otimes z_{\ell-1}\|_2 < \varepsilon/\ell!\Big]^{1/(2^{\ell-1})},
\end{aligned} \tag{10}$$

where $z_0 \sim N(0, \frac{\rho^2(\ell+1)}{2\ell n})^n$, $z_1, z_2, \ldots, z_{\ell-1} \sim N(0, \frac{\rho^2}{2\ell n})^n$. Then

$$\mathbb{P}\Big[\|\Pi_W \tilde{x}^{\otimes \ell}\|_2 < \frac{c(\ell)\rho^\ell}{n^\ell}\Big] \le \exp\Big(-c'(\ell)\delta n\Big), \tag{11}$$

with $c(\ell), c'(\ell) > 0$ being constants that depend only on ℓ. The last inequality follows from (10) and Lemma III.3 applied with $p = 1/e, x_1 = x, x_2 = x_3 = \cdots = x_\ell = 0$, and $\delta' = \delta/\ell^\ell$. This concludes the proof of Theorem IV.1. □

Please see Appendix A for an alternate combinatorial proof when $\ell = 2$. Note that we can also obtain a similar statement for general lower bound of $\varepsilon\eta$ with $\varepsilon \in (0, 1/poly(n))$ (as in Theorem IV.2), where the failure probability becomes $\varepsilon^{\Omega_\ell(\delta n)}$. The proof is exactly the same, except that we can apply Lemma III.3 with $p = \varepsilon^{1/\ell}$ instead. Finally, the proof of Theorem IV.2 is almost identical to Theorem IV.1. In fact Theorem IV.2 essentially corresponds to the special case of Theorem II.1 when $k = 1$. We include a proof of Theorem IV.2 in Appendix A.

C. Condition Number Lower Bounds for Arbitrary Polynomials

We are now ready to complete the proof of Theorem II.1. We start by re-stating the theorem.

Theorem IV.9 (Same as Theorem II.1)**.** *Let $\ell \in \mathbb{Z}_+$ be a constant and let $a_1, a_2, \ldots, a_k \in \mathbb{R}^n$ be any arbitrary collection of vectors, let $f_1, f_2, \ldots, f_m$ be a collection of arbitrary homogeneous polynomials $f_i : \mathbb{R}^n \to \mathbb{R}$ of degree ℓ given by*

$$f_i(x) = \sum_{\substack{J=(j_1,\ldots,j_\ell) \in \binom{[n]}{\ell} \\ j_1 \le j_2 \le \cdots \le j_\ell}} U_i(j_1, \ldots, j_\ell) x_{j_1} x_{j_2} \ldots x_{j_\ell},$$

and let $M_f(a_1, \ldots, a_k) = \big(f_i(a_j)\big)_{i \in [m], j \in [k]}$ be the $m \times k$ matrix formed by applying each of these polynomials with the k vectors $a_1, \ldots, a_k$. Denote by $U \in \mathbb{R}^{m \times D}$ with $D = \binom{n+\ell-1}{\ell}$, with row $i \in [m]$ representing

coefficients of f_i. We have that with probability at least $1 - \exp\big(-\Omega_\ell(\delta n) + \log k\big)$ that

$$\sigma_k\Big(M_f(\tilde{a}_1, \ldots, \tilde{a}_k)\Big) \geq \frac{\Omega_\ell(1)}{\sqrt{k}} \cdot \frac{\rho^\ell}{n^\ell} \cdot \sigma_{k+\delta D}(U), \quad (12)$$

where $\tilde{a_j}$ represents a random perturbation of a_j with independent Gaussian noise $N(0, \rho^2/n)^n$.

Remark IV.10. We note that the condition on U is almost tight, since $\sigma_k(U)$ being non-negligible is a necessary condition (irrespective of A). Proposition IV.13 shows that the additive δn^ℓ term in number of non-negligble singular values is necessary even when $k = 1$. Also note that by choosing a projection matrix U for a subspace of dimension δD, we recover Theorem IV.1. Finally as before, we can obtain an analogous statement for $\varepsilon \in (0, 1/poly_\ell(n))$ as in Theorem IV.2 (see Section IV-B).

Definition IV.11. Let $D = \binom{n+\ell-1}{\ell}$. For $x_1, \cdots, x_n \in \mathbb{R}$, $P_\ell(x_1, \cdots, x_n) \in \mathbb{R}^D$ is a vector whose entries corresponding to D different degree-ℓ monomials of $x_1, \cdots, x_n$.

The idea behind the proof is to view $M_f(a_1, \ldots, a_k)$ as the product of a coefficient matrix and the matrix whose ith column is $P_\ell(a_i)$. Call the latter matrix Y. The following lemma show how to use the property that Theorem IV.2 gives about Y to show Theorem II.1.

Lemma IV.12. *Let $\delta \in (0, 1)$, and let U be a $D' \times D$ matrix, and let $Y \in \mathbb{R}^{D \times R}$ be a random matrix with independent columns $\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_R$ satisfying the following condition: for each $j \in [R]$, and any fixed subspace $\mathcal{V}$ of dimension at least δD, $\|\Pi_{\mathcal{V}} \tilde{Y}_j\|_2 \geq \kappa_1$ with probability at least $1 - \gamma/R$ over the randomness in $\tilde{Y}_j$. Then assuming $\sigma_{R+\delta D}(U) \geq \kappa_2$, we have that $\sigma_R(UY) \geq \kappa_1\kappa_2/\sqrt{R}$ with probability at least $1 - \gamma$.*

Proof. For convenience let $r := R + \delta D$. We will lower bound the minimum singular value of $M = UY$ using the leave-one-out-distance. Fix an $j \in [R]$; we want column $M_j = U\tilde{Y}_j$ to have a non-negligible component orthogonal to $\mathcal{W} = span\big(\{U\tilde{Y}_i : i \in [R], j \neq i\}\big)$ w.h.p.

Let $\Pi_{\mathcal{W}}, \Pi_{\mathcal{W}^\perp}$ be the projectors onto the space $\mathcal{W}, \mathcal{W}^\perp$ respectively. Note that $\sigma_r(U) = \sigma_{R+\delta D} \geq \kappa_2$, and $\sigma_{D'-R+1}(\Pi_{\mathcal{W}^\perp}) \geq 1$. We can use the following robust version of Sylvester's inequality for products of matrices using the variational characterization of singular values to conclude

$$\begin{aligned}\sigma_{r-R+1}\big(\Pi_{\mathcal{W}^\perp} U\big) &\geq \sigma_{D'-R+1}\big(\Pi_{\mathcal{W}^\perp}\big)\sigma_r(U) \\ &\geq \kappa_2.\end{aligned}$$

Let $\mathcal{V}$ be the subspace spanned by the top $r - R + 1$ right singular vectors of $\Pi_{\mathcal{W}^\perp} U$. Since the dimension of $\mathcal{V}$ is at least $r - R + 1 \geq \delta D$, we can then use the condition of the lemma to conclude that with probability at least $1 - \gamma/R$, $\|\Pi_{\mathcal{V}} \tilde{Y}_j\|_2 \geq \kappa_2\kappa_1$. Hence, by using a union bound over all $j \in [R]$ and using the leave-one-out distance the lemma follows. □

We can now complete the proof of the main result of the section.

Proof of Theorem II.1. The idea is to apply Lemma IV.12 with $D' = m, D = \binom{n+\ell-1}{\ell}, R = k$, where U is the corresponding coefficient matrix, and Y is the matrix whose jth column is $\tilde{a}_j^{\otimes \ell}$. Note that the naive representation of $\tilde{a}_j^{\otimes \ell} \in \mathbb{R}^{n^{\otimes \ell}}$ is in n^ℓ dimensions, whereas the rows of the co-efficient matrix U is in $\mathbb{R}^D$. However $\tilde{a}_j^{\otimes \ell}$ are elements of the D-dimensional space of symmetric tensors of order ℓ (alternately each row of U can be seen as a n^ℓ dimensional vector constructed by flattening the corresponding symmetric order ℓ tensor for that row of U). Hence, Theorem IV.1 implies that Y satisfies the conditions of Lemma IV.12, and this completes the proof. □

D. Tight Example for Theorem II.1 and IV.2

We now give a simple example that demonstrates that the condition on many non-trivial singular values for the matrix M that encodes g is necessary.

Proposition IV.13. *In the notation of Theorem IV.2, for any $r \geq 1$, there exists a matrix $M \in \mathbb{R}^{m \times n^\ell}$ (where $m = rn^{\ell-1}$), with the jth row corresponding to a symmetric order ℓ tensor M_j, such that $\sigma_{rn^{\ell-1}}(M) = \Omega_\ell(1)$, but*

$$\mathop{\mathbb{P}}_{z \sim N(0,1/n)^n}\Big[\|g(z)\|_2 = \|Mz^{\otimes \ell}\| \leq \varepsilon\Big] \geq (c\varepsilon)^{O_\ell(r)},$$

for some absolute constant $c > 0$.

Considering the subspace of symmetric tensors spanned by the rows of M also gives a similar tight example for Theorem IV.1. Moreover, the above example also gives a tight example for Theorem II.1 even when $k = 1$, by considering the function $f(x) := g(x)$, and $a_1 = 0$ (so $\widetilde{a}_1 = z$).

Proof. Let $e_1, \ldots, e_n$ constitute the standard basis for $\mathbb{R}^n$. Let $\mathcal{U}$ be the space $\mathbb{R}^{n^{\ell-1}}$, and let $\mathcal{V} \subset \mathbb{R}^n$ be the subspace spanned by $e_1, e_2, \ldots, e_r$. Let $E_1, E_2, \ldots, E_{n^{\ell-1}} \in \mathbb{R}^{n^{\ell-1}}$ constitute the standard basis of $\mathcal{U}$ given by all the $\ell - 1$ wise tensor products of $e_1, \ldots, e_n$. Consider the product space $\mathcal{W} = \mathcal{U} \otimes \mathcal{V}$,

and let B be the matrix whose $m = rn^{\ell-1}$ rows correspond to the orthonormal basis of $\mathcal{W}$ given by $\{E_I \otimes e_j : I \in [n]^{\ell-1},\ j \in [r]\}$. Note that each of these vectors are 1-sparse. Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be given by $\forall j \in [m],\ g_j(x) = \langle B_j, x^{\otimes \ell}\rangle$. First note that by definition, $\|g(x)\|_2 = \|\Pi_{\mathcal{U}\otimes\mathcal{V}} x^{\otimes\ell}\|_2$. Hence, if $z \sim N(0, 1/n)^n$, we have

$$\begin{aligned}
\mathop{\mathbb{P}}_{z}\left[\|g(z)\|_2 \le \varepsilon\right] &= \mathop{\mathbb{P}}_{z}\left[\|\Pi_{\mathcal{U}\otimes\mathcal{V}} z^{\otimes\ell}\|_2 \le \varepsilon\right] \\
&= \mathbb{P}\left[\|\Pi_{\mathcal{U}} z^{\otimes(\ell-1)}\|\|\Pi_{\mathcal{V}} z\| \le \varepsilon\right] \\
&\ge \mathbb{P}\left[\|\Pi_{\mathcal{V}} z\| \le \varepsilon/2, \|z\| \le 2^{1/(\ell-1)}\right] \\
&= \mathbb{P}\left[\|\Pi_{\mathcal{V}} z\| \le \varepsilon/2 \mid \|z\| \le 2^{1/(\ell-1)}\right] \cdot (1 - o(1)) \\
&\ge \mathbb{P}\left[\|\Pi_{\mathcal{V}} z\| \le \varepsilon/2\right] \cdot (1 - o(1)) \\
&\ge (c\varepsilon)^r,
\end{aligned}$$

for some absolute constant $c > 0$, using standard properties of Gaussians. The second-to-last step follows by Lemma A.1 in the Appendix.

We now just need to give a lower bound of $\Omega(rn^{\ell-1})$ for the number of non-trivial singular values of the matrix M, where M_j is the symmetric order ℓ tensor representing g_j i.e., $\langle M_j, x^{\otimes\ell}\rangle = \langle B_j, x^{\otimes\ell}\rangle$ for every $x \in \mathbb{R}^n$. In other words M_j is just the symmetrization (projection onto the space of all symmetric tensors) of B_j. Note that each M_j is $\ell!$ sparse (since B_j were 1-sparse). Hence there are at least $rn^{\ell-1}/\ell!$ vectors M_j which have disjoint support. Hence at least $rn^{\ell-1}/\ell!$ singular values of M are at least $1/\sqrt{\ell!}$, as required. □

V. Polynomials of Few Random Vectors

In this section, we consider random matrix ensembles, where each column is a constant degree "monomial" involving a few of the columns. We will first consider a matrix M whose columns are degree ℓ monomials in the input vectors $\widetilde{a}_1, \ldots, \widetilde{a}_k$ (that is, tensors of the form $\widetilde{a}_{f(1)} \otimes \ldots \otimes \widetilde{a}_{f(\ell)}$ with $f(i) \in [k]$ for $i = 1, \ldots, \ell$). Since the same vector may appear in many columns or multiple times within the same column, there are now dependencies in the perturbations between columns as well as within a column, so we cannot apply [8] directly. We deal with these dependencies by extending an idea of Ma, Shi and Steurer [14], carefully defining appropriate subspaces that will allow us to decouple the randomness.

Since one type of dependence comes from the same input vector appearing in many different columns, it is natural to require that the number of these overlaps be small. Because of the decoupling technique used to avoid dependencies within a column, the troublesome overlaps are only those in which the same input vector appears in two different columns of M in the same position within the tensor product. This motivates the following definition.

Definition V.1. Let M be a matrix whose columns $M_1, \ldots, M_R$ consist of order-ℓ tensor products of $\{\widetilde{a}_1, \ldots, \widetilde{a}_k\}$. For $s \in [\ell]$ and a fixed column M_i, let $\Delta_s(i)$ be the number of other columns that differ from M_i in exactly s spots. (If $M_i = \widetilde{a}_{f(1)} \otimes \ldots \otimes \widetilde{a}_{f(\ell)}$ and $M_j = \widetilde{a}_{f'(1)} \otimes \ldots \otimes \widetilde{a}_{f'(\ell)}$, then the number of spots in which M_i and M_j differ is $|\{i : f(i) \ne f'(i)\}|$.) Finally, let $\Delta_s = \max_i \Delta_s(i)$.

Theorem V.2 (Same as Theorem II.2)**.** *Let $\{\widetilde{a}_1, \ldots, \widetilde{a}_k\} \subseteq \mathbb{R}^n$ be a set of ρ-perturbed vectors, let $\ell \in \mathbb{Z}_+$ be a constant, and let $M \in \mathbb{R}^{n^\ell \times R}$ be a matrix whose columns $M_1, \ldots, M_R$ are tensor monomials in $\{\widetilde{a}_i\}$. Let Δ_s be as above for $s = 1, \ldots, \ell$. If*

$$\sum_{s=1}^{\ell} \Delta_s \cdot \left(\frac{n}{\ell}\right)^{\ell-s} \le c\left(\frac{n}{\ell}\right)^{\ell} \tag{13}$$

for some $c \in (0, 1)$, then $\sigma_R(M) > \Omega_\ell(1) \cdot (\rho/n)^\ell/\sqrt{R}$ with probability at least $1 - \exp(-\Omega_\ell(1)(1-c)n + \log R)$.

Remark V.3. The condition (13) is tight up to a multiplicative constant depending only on ℓ. We give a simple upper bound on Δ_s. Assume $\sigma_R(M) > 0$, and fix a column M_i of M. There are $\binom{\ell}{s}$ ways to choose a set of s spots in which to differ from M_i, and once we make this choice, the dimension of the available space is n^s since each of the s spots contributes n dimensions. Therefore the subspace of $\mathbb{R}^{n^\ell}$ consisting of all tensors that differ from M_i in exactly s spots has dimension at most $\binom{\ell}{s} n^s$. Since all subsets of columns of M must be linearly independent, we must have $\Delta_s \le \binom{\ell}{s} n^s$. Therefore our condition is tight up to a factor of at most $\ell^{2\ell+1}$.

In the above theorem, as stated, the columns of M are "monomials" involving the underlying vectors $\widetilde{a}_1, \ldots, \widetilde{a}_k$. However in our applications (e.g., Sections VII and VIII) the matrix of interest M' will have columns that are more general polynomials of the underlying vectors. Such matrices are expressible as $M' = MP$ where $P \in \mathbb{R}^{R \times R'}$ is a coefficient matrix with $\sigma_{R'}(P) > 1/\mathrm{poly}(n, 1/\rho)$. Hence, our theorem implies that $\sigma_{R'}(M') > 1/\mathrm{poly}(n, 1/\rho)$ in these cases w.h.p.

As in [8], we will use leave-one-out distance, denoted $\ell(M)$, as a surrogate for the smallest singular value. The proof will make use of Lemma III.3, which we will use

to bound leave-one-out distances. Our goal will be to find a suitable subspace W that is both large enough and independent of the column of M we are projecting.

Proof of Theorem V.2. Let $L_1, \ldots, L_\ell$ be an equipartition of $[n]$. Define a new matrix $M' \in \mathbb{R}^{(\frac{n}{\ell})^\ell \times R}$ by restricting the columns of M to the indices $L_1 \times L_2 \times \ldots \times L_\ell$. In other words, if M_i is a column of M with $M_i = \widetilde{a}_{f(1)} \otimes \ldots \otimes \widetilde{a}_{f(\ell)}$, then $M'_i = \widetilde{a}_{f(1),L_1} \otimes \ldots \otimes \widetilde{a}_{f(\ell),L_\ell}$, where a_L denotes the restriction of the vector a to the coordinates in the set L. This ensures that for every column M_i, the perturbations of each factor of this tensor product are independent.

Fix a column M'_i of M', and let W be the subspace spanned by all other columns of M'. We want to find a subspace V satisfying:

1) $W \subseteq V$.
2) V is independent of M'_i.
3) $\dim V^\perp = c'(\frac{n}{\ell})^\ell$ for some $c' \in (0,1)$.

Given such a V, properties 2 and 3 allow us to apply Lemma III.3 to obtain that $\|\mathrm{Proj}_{V^\perp} M'_i\| \geq \Omega_\ell((\rho/n)^\ell)$ with probability at least $1 - \exp(-\Omega(c'n))$. Since $W \subseteq V$, we have

$$\|\mathrm{Proj}_{W^\perp} M'_i\| \geq \|\mathrm{Proj}_{V^\perp} M'_i\| \geq \Omega_\ell((\rho/n)^\ell)$$

with high probability. Taking a union bound over all columns of M' gives that $\ell(M') \geq \Omega_\ell((\rho/n)^\ell)$ with probability at least $1 - \exp(-\Omega_\ell(1) \cdot c'n + \log R)$. Since adding more rows to M' can only increase the magnitude of the projection of any column onto some subspace, $\ell(M) \geq \ell(M')$. Now using properties of the leave-one-out distance (Lemma III.2), we have

$$\sigma_{min}(M) \geq \frac{\ell(M)}{\sqrt{R}} \geq \Omega_\ell(1) \cdot \frac{\rho^\ell}{n^\ell \sqrt{R}}.$$

Next we construct the subspace V. Let $M'_{i'}$, $i' \neq i$ be some other column of M'. Let $S \subseteq [\ell]$ be the set of indices at which M'_i and $M'_{i'}$ share a factor, and let $s = |S|$. In order to ensure V is independent of M'_i, we must avoid touching any factors of $M'_{i'}$ shared by M'_i. Therefore we include in V all vectors of the form $\widetilde{u}_1 \otimes \ldots \otimes \widetilde{u}_\ell$, where $\widetilde{u}_j$ agrees with the jth factor of $M'_{i'}$ if $j \notin S$ and $\widetilde{u}_j$ is any vector in $\mathbb{R}^{n/\ell}$ otherwise. As desired, V now includes $M'_{i'}$ and is independent of M'_i, at a cost of adding $(\frac{n}{\ell})^s$ dimensions to V.

Repeat this process for each $i' \neq i$, and let V be the span of all vectors included at each step. Since the number of overlaps with M'_i can be s at most $\Delta_{\ell-s}$ times, the total dimension of V is at most $\sum_{s=1}^{\ell} \Delta_s (\frac{n}{\ell})^{\ell-s}$. By our assumption on the Δ_ss, we get $\dim V^\perp = c'(\frac{n}{\ell})^\ell$ as desired, with $c' = 1 - c$. □

VI. Robust Subspace Recovery

We introduce the following smoothed analysis framework for studying robust subspace recovery. The following model also tolerates some small amount of error in each point i.e., inliers need not lie exactly on the subspace, but just close to it.

A. Input model

In what follows, $\alpha, \varepsilon_0, \rho \in (0,1)$ are parameters.

1) An adversary chooses a hidden subspace T of dimension d in $\mathbb{R}^n$, and then chooses αm points from T and $(1-\alpha)m$ points from $\mathbb{R}^n$. We denote these points inliers and outliers respectively. Then the adversary mixes them in arbitrary order. Denote these points $a_1, a_2, \ldots, a_m$. Let $A = (a_1, a_2, \ldots, a_m)$, and I_{in}, I_{out} be the set of indices of inliers and outliers respectively. For convenience, we assume that all the points have lengths in the range $[1/2, 1]$.[4]
2) Each inlier is ρ-perturbed with respect to T. (Formally, this means considering an orthonormal basis B_T for T and adding $B_T v$, where $v \sim \mathcal{N}(0, \rho^2/d)^d$.) Each outlier is ρ-perturbed with respect to $\mathbb{R}^n$. Let G denote the perturbations, and let us write $\widetilde{A} = A + G$.
3) With the constraint $\|E\|_F \leq \varepsilon_0$, the adversary adds noise $E \in \mathbb{R}^{n \times m}$ to A, yielding $\widetilde{A}' = \widetilde{A} + E = (\widetilde{a}'_1, \widetilde{a}'_2, \cdots)$. Note that this adversarial noise can depend on the random perturbations in step 2.
4) We are given $\widetilde{A}'$.

The goal in the subspace recovery problem is to return a subspace T' close to T.

a) Notation.: As introduced above, $\tilde{A} = A + G$ denotes the perturbed vectors. $\tilde{a}_i$ denotes the i'th column of $\tilde{A}$. We also use the notation A_I to denote the submatrix of A corresponding to columns in a set I.

B. Our result

We show the following theorem about the recoverability of T.

Theorem VI.1. *Let $\delta \in (0,1)$, $\ell \in \mathbb{Z}_+$ and $\rho > 0$. Suppose we are given $m \geq n^\ell + 8d/(\delta\alpha)$ points $x_1, x_2, \cdots, x_m \in \mathbb{R}^n$ generated as described above, where the fraction of inliers α satisfies $\alpha \geq (1+\delta)\binom{d+\ell-1}{\ell}/\binom{n+\ell-1}{\ell}$. Then there exists $\varepsilon_0 = \mathrm{poly}_\ell(\rho/m)$ such that whenever $\|E\|_F \leq \varepsilon_0$, there is an efficient*

[4] If the perturbations in step (2) are done proportional to the norm, this assumption can be made without loss of generality. (Since the algorithm can scale the lengths of each of the points.)

deterministic algorithm that returns a subspace T' that satisfies

$$\|sin\Theta(T,T')\|_F \leq \|E\|_F \cdot \mathrm{poly}_\ell(m, 1/\rho),$$

$$\textit{w.p.} \geq 1 - 2m^2[\exp(-\Omega_\ell(\delta n)) + \exp(-\Omega(d\log m))]. \tag{14}$$

When $d/n < 1$, the above theorem gives recovery guarantees even when the fraction of inliers is approximately $(d/n)^\ell$. This can be significantly smaller than d/n (shown in [29]) for any constant $\ell > 1$.

a) Algorithm overview.: We start by recalling the approach of [29]. The main insight there is that if we sample a set of size slightly less than n from the input, and if the fraction of inliers is $> (1+\delta)d/n$, then there is a good probability of obtaining $> d$ inliers, and thus there exist points that are in the linear span of the others. Further, since we sampled fewer than n points and the outliers are also in general position, one can conclude that the only points that are in the linear span of the other points are the inliers! In our algorithm, the key idea is to use the same overall structure, but with tensored vectors. Let us illustrate in the case $\ell = 2$. Suppose that the fraction of inliers is $> (1+\delta)\binom{d+1}{2}/\binom{n+1}{2}$. Suppose we take a sample of size slightly less than $\binom{n+1}{2}$ points from the input, and consider the flattened vectors $x \otimes x$ of these points. As long as we have more than $\binom{d+1}{2}$ inliers, we expect to find linear dependencies among the tensored inlier vectors. Further, using Theorem IV.1 (with some modifications, as we will discuss), we can show that such dependencies cannot involve the outliers. This allows us to find sufficiently many inliers, which in turn allows us to recover the subspace T up to a small error.

Given m points, the algorithm (Algorithm 1) considers several batches of points each of size $b = (1 - \frac{\delta}{3})\binom{n+\ell-1}{\ell}$. Suppose for now that m is a multiple of b, and that the m/b batches form an arbitrary partition of the m points. (See the note in Section VI-C for handling the general case.) In every batch, the algorithm does the following: for each point u in the batch, it attempts to represent $u^{\otimes \ell}$ as a "small-coefficient" linear combination (defined formally below) of the tensor products of the other points in the batch. If the error in this representation is small enough, the point is identified as an inlier.

Definition VI.2 (c-bounded linear combination)**.** Let $v_1, v_2, \ldots, v_m$ be a set of vectors. A vector u is said to be expressible as a c-bounded linear combination of the $\{v_i\}$ if there exist $\{\alpha_i\}_{i=1}^m$ such that $|\alpha_i| \leq c$ for all i, and $u = \sum_i \alpha_i v_i$. Further, u is said to be expressible as a c-bounded combination of the $\{v_i\}$ with error δ if there exist $\{\alpha_i\}_{i=1}^m$ as above with $|\alpha_i| \leq c$ for all i, and $\|u - \sum_i \alpha_i v_i\|_1 \leq \delta$.

Notice that in the above definition, the error is measured by ℓ_1 norm. In the algorithm, we will need a subprocedure to check whether a vector is expressible as a 1-bounded combination of some other vectors with some fixed error. By the choice of ℓ_1 norm, this subprocedure can be formulated as a Linear Programming problem, hence we can solve it efficiently.

Algorithm 1 Robust subspace recovery

1: Set threshold $\tau = \Omega_\ell(\rho^\ell/n^\ell)$(which is the threshold from Theorem IV.1). Set batchsize $b = (1 - \delta/3)\binom{n+\ell-1}{\ell}$.
2: Let $V_1, V_2, \cdots, V_r$ be the $r \leq m$ batches each of size b as defined above.
3: Initialize $C = \emptyset$.
4: **for** $i = 1, 2, \cdots, r$ **do**
5: Let S be the set of all $u \in V_i$ such that $\tilde{a}_u'^{\otimes \ell}$ can be expressed as 1-bounded combinations of $\{\tilde{a}_v'^{\otimes \ell} : v \in V_i \setminus \{u\}\}$, with error $\leq \tau/2$.
6: $C = C \cup S$
7: Return the subspace T' corresponding to the top d singular values of $\tilde{A}'_{\overline{C}}$, for any $2d$-sized subset $\overline{C}$ of C

b) Proof outline.: The analysis involves two key steps. The first is to prove that none of the outliers are included in S in step 5 of the algorithm. This is where we use 1-bounded linear combinations. If the coefficients were to be unrestricted, then because the error matrix E is arbitrary, it is possible to have a tensored outlier being expressible as a linear combination of the other tensored vectors in the batch. The second step is to prove that we find enough inliers overall. On average, we expect to find at least $\frac{\delta}{3}\binom{d+\ell-1}{\ell}$ inlier columns in each batch. We "collect" these inliers until we get a total of $2d$ inliers. Finally, we prove that these can be used to obtain T up to a small error.

For convenience, let us write $g(n) := \Omega_\ell(\delta n)$ (which is the exponent in the failure probability from Theorem IV.1). Thus the failure probabilities can be written as $\exp(-g(n))$.

Lemma VI.3. *With probability at least* $1 - \exp(-g(n) + 2\log m)$*, none of the outliers are chosen. I.e.,* $C \cap I_{out} = \emptyset$.

Proof. The proof relies crucially on the choice of the batch size. Let us fix some batch V_j. Note that by the

way the points are generated, each point in V_j is $\widetilde{a}_i'$, for some a_i that is either an inlier or an outlier.

Let us first consider only the perturbations (i.e., without the noise addition step). Recall that we denoted these vectors by $\widetilde{a}_i$. Let us additionally denote by $B^{(j)}$ the matrix whose columns are $\widetilde{a}_i^{\otimes \ell}$ for all i in the phase j. Consider any i corresponding to an outlier. Now, because the batch size is only $(1-\frac{\delta}{3})\binom{n+\ell-1}{\ell}$, we have (using Theorem IV.1) that the projection of the column $B_i^{(j)}$ orthogonal to the span of the remaining columns (which we denote by $B_{-i}^{(j)}$) is large enough, with very high probability. Formally,

$$\mathbb{P}[\text{dist}(B_i^{(j)}, \text{span}(B_{-i}^{(j)})) \geq \tau] \geq 1 - \exp(-g(n)). \quad (15)$$

Indeed, taking a union bound, we have that the inequality $\text{dist}(B_i^{(j)}, \text{span}(B_{-i}^{(j)})) \geq \tau$ holds for all outliers i (and their corresponding batch j) with probability $\geq 1 - m^2 \exp(-g(n))$.

We need to show that moving from the vectors $\widetilde{a}_i$ to $\widetilde{a}_i'$ maintains the distance. For this, the following simple observation will be useful.

Observation VI.4. If a_i is an outlier, then

$$\mathbb{P}[\|\widetilde{a}_i\| \geq 1 + 2\rho] \leq \exp(-n/2).$$

On the other hand if a_i is an inlier,

$$\mathbb{P}[\|\widetilde{a}_i\| \geq 1 + 4\rho\sqrt{\log m}] \leq \exp(-4d \log m).$$

Both the inequalities are simple consequences of the fact that the vectors a_i were unit length to start with, and are perturbed by $\mathcal{N}(0, \rho^2/n)$ and $\mathcal{N}(0, \rho^2/d)$ respectively.

Now let us consider the vectors with noise added, $\widetilde{a}_i'$. Note that $\|\widetilde{a}_i - \widetilde{a}_i'\| \leq \varepsilon_0$. Since $\|a_i\| \leq 1$ and since i is an outlier, we have (using Observation VI.4), $\|\widetilde{a}_i'\| \leq 1 + 2\rho + \varepsilon_0$, with probability $\geq 1 - \exp(-n/2)$. Thus for the flattened vectors $\widetilde{a}_i^{\otimes \ell}$, with the same probability,

$$\begin{aligned}
\|\widetilde{a}_i^{\otimes \ell} - (\widetilde{a}_i')^{\otimes \ell}\| &= \|\Big(\widetilde{a}_i^{\otimes \ell} - \widetilde{a}_i^{\otimes (\ell-1)} \otimes \widetilde{a}_i'\Big) \\
&\quad + \Big(\widetilde{a}_i^{\otimes (\ell-1)} \otimes \widetilde{a}_i' - \widetilde{a}_i^{\otimes (\ell-2)} \otimes \widetilde{a}_i'^{\otimes 2}\Big) \\
&\quad + \ldots\| \\
&\leq \ell(\max\{\|\widetilde{a}_i\|, \|\widetilde{a}_i'\|\})^{\ell-1}\varepsilon_0 \\
&\leq \ell(1 + 2\rho + \varepsilon_0)^{\ell}\varepsilon_0. \quad (16)
\end{aligned}$$

Thus, for any 1-bounded linear combination of the b vectors in the batch (which may contain both inliers and outliers), $\widetilde{a}_i'^{\otimes \ell}$ is at a Euclidean distance $\leq b\ell(1 + \varepsilon_0 + 4\rho\sqrt{\log m})^{\ell}\varepsilon_0$ to the corresponding linear combination of the ℓth powers of the vectors in the batch prior to the addition of noise (i.e., the columns of $B_{-i}^{(j)}$). Thus if $b\ell(1 + \varepsilon_0 + 4\rho\sqrt{\log m})^{\ell}\varepsilon_0 < \tau/2$, then $\widetilde{a}_i'^{\otimes \ell}$ cannot be expressed as a 1-bounded combination of the other lifted vectors in the batch with Euclidean error $< \tau/2$, let alone ℓ_1 error.

This means that none of the outliers are added to the set S, with probability at least $1 - m[\exp(-g(n)) - \exp(-4d \log m)]$. □

Next, we turn to proving that sufficiently many inliers are added to S. The following simple lemma will help us show that restricting to 1-bounded combinations does not hurt us.

Lemma VI.5. *Let $u_1, u_2, \ldots, u_{d+c}$ be vectors that all lie in a d-dimensional subspace of $\mathbb{R}^n$. Then at least c of the u_i can be expressed as 1-bounded linear combinations of $\{u_j\}_{j \neq i}$.*

Proof. As the vectors lie in a d-dimensional subspace, there exists a non-zero linear combination of the vectors that adds up to zero. Suppose $\sum_i \alpha_i u_i = 0$. Choose the i with the largest value of $|\alpha_i|$. This u_i can clearly be expressed as a 1-bounded linear combination of $\{u_j\}_{j \neq i}$.

Now, remove the u_i from the set of vectors. We are left with $d + c - 1$ vectors, and we can use the same argument inductively to show that we can find $c - 1$ other vectors with the desired property. This completes the proof. □

The next lemma now proves that the set C at the end of the algorithm is large enough.

Lemma VI.6. *For the values of the parameters chosen above, we have that at the end of the algorithm,*

$$|C| \geq \frac{\delta/3}{1 + \delta/3}\alpha m,$$

with probability at least $1 - \exp(-4d \log m)$.

Proof. We start with the following corollary to Lemma VI.5. Let us consider the jth batch.

Observation. Let n_j be the number of inliers in the jth batch. If $n_j \geq \binom{d+\ell-1}{\ell} + k$, then the size of S found in Step 5 of the algorithm is at least k.

Proof of Observation. Define $B^{(j)}$ as in the proof of Lemma VI.3. Now, since the inliers are all perturbed within the target subspace, we have that the vectors $\widetilde{a}_i^{\otimes \ell}$ corresponding to the inliers all live in a space of dimension $\binom{d+\ell-1}{\ell}$. Thus by Lemma VI.5, at least k of the vectors $B_i^{(j)}$ can be written as 1-bounded linear combinations of the vectors $B_{-i}^{(j)}$.

For inliers i, using the fact that a_i are perturbed by $\mathcal{N}(0, \rho^2/d)$, we have

$$\mathbb{P}[\|\widetilde{a}_i\| \geq (1 + 4\rho\sqrt{\log m})] \leq \exp(-4d \log m).$$

Using (16) again, we have that $\widetilde{a}_i'^{\otimes \ell}$ can be expressed as a 1-bounded linear combination of the other vectors in the batch, with Euclidean error bounded by $b\ell \cdot (1 + 5\rho\sqrt{\log m})^\ell \varepsilon_0$. We know ℓ_1 norm is a $\sqrt{n^\ell}$-approximation of ℓ_2 norm. By assumption, the ℓ_1 norm of the error is $< \tau/2$, thereby completing the proof of the observation. □

Now, note that we have $\sum_j n_j \geq \alpha m$, by assumption. This implies that

$$\begin{aligned}\sum_{j=1}^{m/b} \max\left\{0, n_j - \binom{d+\ell-1}{\ell}\right\} &\geq \\ &\geq \alpha m - \frac{m}{b}\binom{d+\ell-1}{\ell} \\ &\geq \frac{\delta/3}{1+\delta/3}\alpha m.\end{aligned}$$

The last inequality follows from our choice of α and the batch size b. Thus the size of S in the end satisfies the desired lower bound. □

Finally, we prove that using any set of $2d$ inliers, we can obtain a good enough approximation of the space T, with high probability (over the choice of the perturbations). The probability will be high enough that we can take a union bound over all $2d$-sized subsets of $[m]$.

Lemma VI.7. *Let $I \subseteq I_{in}$ be any (fixed) set of size $2d$. Then if $\|E\|_F \leq \mathrm{poly}(\rho/m)$, the subspace U corresponding to the top d singular value of $\tilde{A}_I'$ will satisfy*

$$\|\sin\Theta(U, T)\|_F \leq \mathrm{poly}(m, 1/\rho) \cdot \|E\|_F$$

with probability at least $1 - e^{-4d \log m}$.

Proof. We start by considering the matrix $\tilde{A}_I$ (the matrix without addition of error). This matrix has rank $\leq d$ (as all the columns lie in the subspace T). The first step is to argue that $\sigma_d(\tilde{A}_I)$ is large enough. This implies that the space of the top d SVD directions is precisely T. Then by using Wedin's theorem [43], the top d SVD space U of $\tilde{A}_I'$ satisfies

$$\|\sin\Theta(U, T)\|_F \leq \frac{2\sqrt{d}\|E\|_F}{\sigma_d(\tilde{A}_I) - \|E\|_F}. \tag{17}$$

Hence it suffices to show $\sigma_d(\tilde{A})$ is at least inverse-polynomial with high probability.

Recall that $\tilde{A}_I = A_I + G_I$, where G_I is a random matrix. Without loss of generality we can assume that T is spanned by the first d co-ordinate basis; in this case every non-zero entry of G_I is independently sampled from $\mathcal{N}(0, \frac{\rho^2}{d})$. We can thus regard A_I, G_I as being $d \times 2d$ matrices. Recall that leave-one-out distance is a good approximation of least singular value, it suffices to show $\ell((A_I + G_I)^T)$ is at least inverse-polynomial with high probability. Let A_j, G_j denote the jth row of A_I, G_I correspondingly. Consider $j \in [d]$, fix all other rows except jth. Let W be the subspace of $\mathbb{R}^{2d}$ that is orthogonal to $\mathrm{span}(\{A_k + G_k : k \in [d], k \neq j\})$, and let $w_1, w_2, \ldots, w_{d+1}$ be an orthonormal basis for W. Then for any $t > 0$, if the projection of $(A_j + G_j)$ to W is $< t$ (equivalent to the leave-one-out distance $< t$), then for all $1 \leq i \leq d+1$, we must have $|\langle w_i, A_j + G_j\rangle| \leq t$. Using the anti-concentration of a Gaussian and the orthogonality of the w_i, this probability can be bounded by $(t/\rho)^{d+1}$. Choosing $t = \rho/m^4$, this can be made $< (1/m^4)^{d+1}$, and thus after taking a union bound over the m choices of j, we have that the leave-one-out distance is $> \rho/m^4$ (and thus $\sigma_d > \rho/m^5$) with probability $\geq 1 - \exp(-4d \log m)$ □

We can now complete the proof of the theorem.

Proof of Theorem VI.1. Suppose that $\|E\|_F \leq \varepsilon_0$ is small enough. Now by Lemma VI.3, we have that $C \subseteq I_{in}$ with probability at least $1 - \exp(-g(n) + \log m$. By Lemma VI.6 and our assumption that m is at least $\Omega(d/(\delta\alpha))$, we know $|C| \geq 2d$ with probability $1 - e^{-4d \log m}$. Finally, by Lemma VI.7 and a union bound over all $2d$ sized subsets of $[m]$, we have that with probability at least $1 - \exp(-\Omega(d \log m))$, for any subset of inliers with size $2d$, the subspace T' corresponding to the top-d singular value will satisfy $\|\sin\Theta(T, T')\|_F \leq \|E\|_F/\mathrm{poly}(m)$. □

C. Batches when m is not a multiple of b

he case of m not being a multiple of b needs some care because we cannot simply ignore say the last few points (most of the inliers may be in that portion). But we can handle it as follows: let m' be the largest multiple of b that is $< m$. Clearly $m' > m/2$. Now for $1 \leq j \leq n$, define $\mathcal{D}_j = \{x_j, x_{j+1}, \ldots, x_{j+m'-1}\}$ (with the understanding that $x_{n+t} = x_t$). This is a set of m' points for every choice of j. Each $\mathcal{D}_j$ is a possible input to the algorithm, and it has at least $m' > m/2$ points, and additionally the property that $b|m'$.

At least one of the $\mathcal{D}_j$ has $\geq \alpha$ fraction of its points being inliers (by averaging). Thus the procedure above (and the guarantees) can be applied to recover the space. To ensure that no outlier is chosen in step 5 of the algorithm (Lemma VI.3), we take an additional union bound to ensure that Lemma VI.3 holds for all $\mathcal{D}_j$.

VII. Learning Hidden Markov Models

We consider the setup of Hidden Markov Models considered in [17], [32]. A hidden state sequence $Z_1, Z_2, \ldots, Z_m \in [r]$ forms a stationary Markov chain with transition matrix P and initial distribution $w = \{w_k\}_{k\in[r]}$, assumed to be the stationary distribution. The observations $\{X_t\}_{t\in[m]}$ are vectors in $\mathbb{R}^n$. The observation matrix of the HMM is denoted by $\mathcal{O} \in \mathbb{R}^{n\times r}$; the columns of $\mathcal{O}$ represent the conditional means of the observation $X_t \in \mathbb{R}^n$ conditioned on the hidden state Z_t i.e., $\mathbb{E}[X_t|Z_t = i] = \mathcal{O}_i$, where $\mathcal{O}_i$ represents the ith column of $\mathcal{O}$. We also assume that X_t has a subgaussian distribution about its mean (e.g., X_t is distributed as a multivariate Gaussian with mean $\mathcal{O}_i$ when the hidden state $Z_t = i$). In the smoothed analysis setting, the model is generated using a randomly perturbed observation matrix $\widetilde{\mathcal{O}}$, obtained by adding independent Gaussian random vectors drawn from $N(0, \rho^2/n)^n$ to each column of $\mathcal{O}$. We remark that some prior works [17], [35] consider the more restrictive discrete setting where the observations are discrete over an alphabet of size n.[5] While our smoothed analysis model with small Gaussian perturbations is natural for the more general continuous setting, it may not be an appropriate smoothed analysis model for the discrete setting (for example, the perturbed vector $\mathcal{O}_i$ could have negative entries).

Using a trick from [17], [32], we will translate the problem into the setting of multi-view models. Let $m = 2\ell+1$ for some ℓ to be chosen later, and use the hidden state $Z_{\ell+1}$ as the latent variable. In what follows, we will abuse notation and also represent the states using the standard basis vectors $e_1, e_2, \ldots, e_r \in \mathbb{R}^r$: for each $j \in [r], \ell \in [m]$, $Z_\ell = e_j \in \mathbb{R}^r$ iff the state at time ℓ is j. Our three views are obtained by looking at the past, present, and future observations: the first view is $X_\ell \otimes X_{\ell-1} \otimes \ldots \otimes X_1$, the second is $X_{\ell+1}$ and the third is $X_{\ell+2} \otimes X_{\ell+3} \otimes \ldots X_{2\ell+1}$. We can access these views by viewing the moment tensor $X_1 \otimes \ldots \otimes X_{2\ell+1}$ as a 3-tensor of shape $n^\ell \times n \times n^\ell$. The conditional expectations of these three views are given by matrices A, B, and C of dimensions $n^\ell \times r$, $n \times r$, and $n^\ell \times r$ respectively. Explicitly, these matrices satisfy

$$\begin{aligned}\mathbb{E}[X_\ell \otimes \ldots \otimes X_1|Z_{\ell+1}] &= AZ_{\ell+1},\\ \mathbb{E}[X_{\ell+1}|Z_{\ell+1}] &= BZ_{\ell+1},\\ \mathbb{E}[X_{\ell+2} \otimes \ldots \otimes X_{2\ell+1}|Z_{\ell+1}] &= CZ_{\ell+1}.\end{aligned}$$

Let $P' = \mathrm{diag}(w)P^T\mathrm{diag}(w)^{-1}$, which is the reverse transition matrix $Z_i \to Z_{i-1}$, and let $X \odot Y$ denote the Khatri-Rao product of X and Y, given in terms of its columns by $(X \odot Y)_i = X_i \otimes Y_i$. Then we can write down A, B, and C in terms of the transition and observation matrices as follows. This fact is straightforward to check, so we leave the details to [17].

$$A = ((\ldots(\widetilde{\mathcal{O}}P') \odot \widetilde{\mathcal{O}})P') \odot \widetilde{\mathcal{O}}) \ldots P') \odot \widetilde{\mathcal{O}})P' \quad (18)$$

$$B = \widetilde{\mathcal{O}} \quad (19)$$

$$C = ((\ldots(\widetilde{\mathcal{O}}P) \odot \widetilde{\mathcal{O}})P) \odot \widetilde{\mathcal{O}}) \ldots P) \odot \widetilde{\mathcal{O}})P, \quad (20)$$

where $\widetilde{\mathcal{O}}$ and P or P' appear ℓ times each in A and C. Our goal is to upper bound the condition numbers of A and C. Once we do this, we will be able to use a argument similar to that in [33] to obtain P and $\widetilde{\mathcal{O}}$ up to an inverse polynomial error.

The proof of this theorem will use a simple lemma relating the minimum singular value of a matrix A to that of a matrix obtained by adding together rows of A.

Lemma VII.1. *Let n_1, n_2, n_3 be positive integers with $n_2 \geq n_3$. Let $A = (A_{(i_1,i_2),j}) \in \mathbb{R}^{n_1 n_2 \times n_3}$ be a matrix, and let $B \in \mathbb{R}^{n_2 \times n_3}$ be the matrix whose i_2th row is $\sum_{i_1} A_{[(i_1,i_2),:]}$. Then $\sigma_{n_3}(A) \geq \frac{1}{\sqrt{n_1}}\sigma_{n_3}(B)$.*

Proof. We can write $B = MA$, where $M \in \mathbb{R}^{n_2 \times n_1 n_2}$ is a matrix whose ith row consists of $n_1(i-1)$ zeros, then n_1 ones, then $n_1(n_2 - i)$ zeros. For any $v = (v_{ij}) \in \mathbb{R}^{n_1 n_2}$, applying the Cauchy-Schwarz inequality gives

$$\|Mv\|^2 = \sum_{i=1}^{n_2}(M_{[i,:]} \cdot v)^2 = \sum_{i=1}^{n_2}\left(\sum_{j=1}^{n_1} v_{ij}\right)^2 \leq n_1\|v\|^2.$$

Therefore $\sigma_{max}(M) \leq \sqrt{n_1}$. Since $\sigma_{min}(B) \leq \sigma_{max}(M)\sigma_{min}(A)$, we have

$$\sigma_{min}(A) \geq \frac{1}{\sqrt{n_1}}\sigma_{min}(B).$$

□

Theorem VII.2. *Let $\ell \in \mathbb{Z}_+$ be a constant. Suppose we are given a Hidden Markov Model in the setting described above, satisfying the following conditions:*

[5] These observations can be represented using the n standard basis vectors for the n alphabets and column $\mathcal{O}_i$ gives the probability distribution conditioned on the current state being $i \in [r]$.

1) *$P \in \mathbb{R}^{r\times r}$ is d-sparse, where $d < O(\min\{n/\ell^2, n/r^{1/\ell}\})$ and $n = \Omega(r^{1/\ell})$. In addition, we assume $\sigma_{min}(P) \geq \gamma_1$.*
2) *The columns of $\mathcal{O} \in \mathbb{R}^{n\times r}$ are polynomially bounded (i.e. the lengths are bounded by some polynomial in n) and are perturbed by independent Gaussian noise $N(0, \rho^2/n)^n$ to obtain $\widetilde{\mathcal{O}}$, with columns $\{\widetilde{\mathcal{O}}_i\}$.*
3) *The stationary distribution w of P has $w_i > \gamma_2$ for all $i \in [r]$.*

Then there is an algorithm that recovers P and $\widetilde{\mathcal{O}}$ up to ε error (in the Frobenius norm) with probability at least $1-\exp(-\Omega_\ell(n))$, using samples of $m = 2\ell+1$ consecutive observations of the Markov chain. The algorithm runs in time $(n/(\rho\gamma_1\gamma_2\varepsilon))^{O(\ell)}$.

Proof. We will show that C is well-conditioned. First note that since the columns of $\widetilde{\mathcal{O}}$ (and therefore of C) are polynomially bounded, $\sigma_{max}(C)$ is also bounded by some polynomial in n and r. Therefore we only need to give a lower bound on $\sigma_{min}(C)$. Since $\sigma_{min}(P') \geq \gamma_2 \cdot \sigma_{min}(P)$, the proof for A is identical. We can write $C = M(\widetilde{\mathcal{O}}, P) \cdot F(P)$, where $M \in \mathbb{R}^{n^\ell \times R}$ is a matrix whose columns are order-ℓ tensor products of $\{\widetilde{\mathcal{O}}_i\}$ and $F(P) \in \mathbb{R}^{R\times r}$ is a matrix of coefficients. We will show that each of these factors is well-conditioned, which will give us a bound on the condition number of C.

First we work with M. The columns of M are all of the tensor products of $\{\widetilde{\mathcal{O}}_i\}$ that appear in C. Specifically, if the columns of $\widetilde{\mathcal{O}}$ are $\{\widetilde{\mathcal{O}}_i\}_{i\in[r]}$, then the columns of M are all tensor products of the form

$$\widetilde{\mathcal{O}}_{i_1} \otimes \ldots \otimes \widetilde{\mathcal{O}}_{i_\ell}, \tag{21}$$

where $P_{i_s,i_{s+1}} \neq 0$ for all $s = 1, \ldots, \ell-1$. The key here is that while the noise coming from the ρ-perturbations of $\{\mathcal{O}_i\}$ is not independent column to column, any column of M has noise that is highly correlated with only a few other columns.

In order to apply Theorem II.2, we need to find $\Delta_1, \ldots, \Delta_\ell$. Fix a column M_i of M. For $s < \ell$, we have

$$\Delta_s(M_i) \leq \binom{\ell}{s} d^s. \tag{22}$$

To show why, we describe a way of generating all columns of M that differ from M_i in s factors. First, choose a set $S \subseteq [\ell]$ with $|S| = s$, which will specify the places at which the new column will differ from M_i. Begin at one place at which the new column will not differ, which is possible because $s < \ell$. Fill in the remaining factors by progressing by step forwards and backwards until each factor is chosen. Each time a place in S is encountered, we have at most d choices due to the sparsity of P.

Remark VII.3. Note that not all of these choices may correspond to a path through the state space of the Markov chain. Thus additional conditions limiting the number of short cycles in the graph of the Markov chain could lead to smaller upper bounds on $\{\Delta_s\}$.

For $s = \ell$, we have $\Delta_\ell(M_i) \leq R \leq r \cdot d^\ell$ since all of the ℓ factors are arbitrary as long as they determine a path in the Markov chain.

Now the condition of Theorem II.2 becomes

$$rd^\ell + \sum_{s=1}^{\ell-1} \binom{\ell}{s} d^s \left(\frac{n}{\ell}\right)^{\ell-s} \leq c\left(\frac{n}{\ell}\right)^\ell \qquad \text{for } c \in (0,1), \tag{23}$$

which holds by the restrictions on d and r. Therefore we conclude that $\sigma_{min}(M) \geq \Omega_\ell(1) \cdot (\rho/n)^\ell/\sqrt{R}$ with probability at least $1 - \exp(-\Omega_\ell(n) + \log R) \geq 1 - \exp(-\Omega_\ell(n) + \log n^\ell) = 1 - \exp(-\Omega_\ell(n))$.

Next, we show that F is well-conditioned. To simplify notation, we write as if $R = r^\ell$ (in which case M would have many unused columns and F would have many zero rows and columns). Index the rows of F by a tuple $(i_1, \ldots, i_\ell)$. We have

$$F_{(i_1,\ldots,i_\ell),j} = P_{ji_1}P_{i_1i_2}\cdots P_{i_{\ell-1}i_\ell}. \tag{24}$$

In other words, the coefficient of $\widetilde{\mathcal{O}}_{i_1} \otimes \ldots \otimes \widetilde{\mathcal{O}}_{i_\ell}$ in column j of C is the probability, given that you begin at state j, of traveling through states $i_1, \ldots, i_\ell$.

We want to give a lower bound for the least singular value of F. Lemma VII.1 shows that it is enough to bound the least singular value of a matrix obtained by adding together rows of F. Using this idea, we sum over all rows with the same i_ℓ to obtain a matrix $F' \in \mathbb{R}^{r\times r}$ with entries

$$F'_{i,j} = \sum_{i_1,\ldots,i_{\ell-1}} P_{ji_1}P_{i_1i_2}\cdots P_{i_{\ell-1}i}. \tag{25}$$

Thus we have $F' = (P^\ell)^T$, which has $\sigma_{min}(F') \geq \gamma_1^\ell$. Therefore Lemma VII.1 gives $\sigma_{min}(F) \geq \gamma_1^\ell/r^{\ell/2}$.

These two results show that

$$\begin{aligned}\sigma_{min}(C) &\geq \Omega_\ell(1) \cdot (\rho\gamma_1)^\ell/(n\sqrt{rd})^\ell r^{1/2} \\ &\geq \Omega_\ell(1) \cdot \left(\frac{\rho\gamma_1}{\sqrt{n^3 r}}\right)^\ell\end{aligned}$$

with probability at least $1 - \exp(-\Omega_\ell(n))$.

As mentioned above, we also get $\sigma_{min}(A) \geq \Omega_\ell(1) \cdot (\rho\gamma_1\gamma_2/\sqrt{n^3 r})^\ell$ with the same probability. In order to recover P and $\widetilde{\mathcal{O}}$, we use an algorithm similar to Algorithm 1 from Sharan et al. [35]. First, we can

estimate the moment tensor $X_1 \otimes \ldots \otimes X_{2\ell+1}$ to sufficient accuracy using $\mathrm{poly}_\ell(n, 1/\varepsilon)$ many samples since each observation vector has a conditional distribution which is subgaussian. This follows from standard large deviation bounds, for example see Lemma C.1 in [33]. Next, we can obtain A, B, and C up to an error $\delta = \mathrm{poly}(\varepsilon, n, \rho)$ using a tensor decomposition algorithm such as in [8]. Since $B = \widetilde{\mathcal{O}}$, it only remains to find P. To do this, we use a similar trick to [17]. We will use the fact that C and P are both well-conditioned. First, let $D = (C \odot \widetilde{\mathcal{O}})P$. Note that we can obtain D by following the entire procedure again but increasing ℓ by one. Since we already have $\widetilde{\mathcal{O}}$ up to a small error, we can also find $C \odot \widetilde{\mathcal{O}}$. Now $\sigma_{min}(C \odot \widetilde{\mathcal{O}}) \geq \sigma_{min}(D)/\sigma_{max}(P)$, and $\sigma_{max}(P) \leq \sqrt{r}$. Therefore we can recover P from D and $C \odot \widetilde{\mathcal{O}}$ up to the required inverse polynomial error. □

VIII. Higher Order Tensor Decompositions

In this section, we describe an algorithm to decompose 2ℓ'th order tensors of rank up to n^ℓ. Let us start by recalling the problem: suppose $A_1, \ldots, A_R$ are vectors in $\mathbb{R}^n$. Consider the 2ℓ'th order moment tensor

$$M_{2\ell} = \sum_{i=1}^{R} A_i^{\otimes 2\ell}.$$

The tensor decomposition problem asks to find the vectors A_i to a desired precision (up to a re-ordering), given only the tensor $M_{2\ell}$. The question of *robust recovery* asks to find the vectors A_i to a desired precision given access to a *noisy* version of $M_{2\ell}$, specifically, given only the tensor $T = M_{2\ell} + \mathsf{Err}$. The aim is to show that recovery is possible, assuming that $\|\mathsf{Err}\|$ is bounded by some polynomial in n and the desired precision for recovering the A_i. We give an algorithm for robust recovery, under certain condition number assumptions on the A_i. Then using the methods developed earlier in the paper, we show that these assumptions hold in a smoothed analysis model.

A. Robust decomposition assuming non-degeneracy

We will now consider a generalization of the algorithm of Cardoso [16], and prove robust recovery guarantees under certain non-degeneracy assumptions. As stated in the introduction, our contribution is along two directions: the first is to extend the algorithms of [16] and [44] to the case of 2ℓ'th order tensors. Second (and more importantly), we give a robustness analysis.

We now define an operator, and then a matrix whose condition number is important for our argument. Given ℓ'th order tensors X, Y, we define the operator Φ as $\Phi(X, Y) = \Psi(X, Y) + \Psi(Y, X)$, where $\Psi : \mathbb{R}^\ell \times \mathbb{R}^\ell \mapsto \mathbb{R}^{2\ell}$ is defined by:

$$\Psi(X, Y)(i_1, i_2, \ldots, i_\ell, j_1, j_2, \ldots, j_\ell) = \\ = X_{i_1 \ldots i_{\ell-1} i_\ell} Y_{j_1 \ldots j_{\ell-1} j_\ell} - X_{i_1 \ldots i_{\ell-1} j_\ell} Y_{j_1 \ldots j_{\ell-1} i_\ell} \quad (26)$$

One of the nice properties of Φ above is that $\Phi(X, X) = 0$ for a *symmetric* tensor[6] X iff $X = \mathbf{u}^{\otimes \ell}$, for some $\mathbf{u} \in \mathbb{R}^n$ (and for this reason, [16] who introduced such an operator for $\ell = 2$ and subsequent works refer to this as a rank-1 "detector"). The algorithm and its analysis only use the easy direction of the above statement, namely $\Phi(\mathbf{u}^{\otimes \ell}, \mathbf{u}^{\otimes \ell}) = 0$ for any $\mathbf{u} \in \mathbb{R}^n$, and thus we do not prove the property above.

The following matrix plays a crucial role in the analysis: consider the $\binom{R}{2}$ vectors of the form $\Phi(A_i^{\otimes \ell}, A_j^{\otimes \ell})$, for $i < j$. Let M_Φ be the matrix with all of these vectors as columns. Thus M_Φ is of dimensions $n^{2\ell} \times \binom{R}{2}$.

a) Relevant condition numbers.: Our robustness analysis will depend on (a) the condition number of the matrix $U := A^{\odot \ell}$, which we will denote by κ_U, and (b) the condition number of the matrix M_Φ described above, which we will denote by κ_M. For convenience, let us also define $\mathbf{u}_i = A_i^{\otimes \ell}$, flattened. From the definition of U above, we also have $M_{2\ell}$ equal to UU^T, when matricized.

The following is our main result of the section.

Theorem VIII.1. *Given the tensor $T = M_{2\ell} + \mathsf{Err}$, an accuracy parameter ε, and the guarantee that $\|\mathsf{Err}\|_F \leq \varepsilon^c/(\kappa_U \kappa_M)^{c'}$ for some constants c, c', there is an algorithm that outputs, with failure probability $1 - \gamma$, a set of vectors $\{B_i\}_{i=1}^R$ such that*

$$\min_\pi \sum_i \|A_i - B_{\pi(i)}\| \leq \varepsilon.$$

Furthermore, this algorithm runs in time $\mathrm{poly}(n^\ell, \kappa_U, \kappa_M, \log(1/\gamma))$.

b) Remark.: We note that the above statement does not explicitly require a bound on the rank R. However, the finiteness of the condition numbers κ_U and κ_M implies that $R \leq n^\ell/2$. Our theorem VIII.13 shows that when $R \leq n^\ell/2$, the condition numbers are both polynomial in n in a smoothed analysis model. Also, we do not explicitly compute c, c'. From following the proof naïvely, we get them to be around 8, but they can likely be improved.

[6] An ℓ'th order tensor T is said to be symmetric if $T_{i_1 i_2 \ldots i_\ell} = T_{\pi(i_1)\pi(i_2)\ldots\pi(i_\ell)}$ for any permutation π.

1) Outline of the proof and techniques: We will start (section VIII-A2) by presenting the FOOBI procedure for arbitrary ℓ. The algorithm proceeds by considering the top eigenvectors of the matricized version of $M_{2\ell}$, and tries to find *product vectors* (i.e. vectors of the form $x^{\otimes \ell}$) in their span. This is done via writing a linear system involving the basis vectors.

In section VIII-A3, we show that the entire procedure can be carried out even if $M_{2\ell}$ is only known up to a small error. The technical difficulty in the proof arises for the following reason: while a small perturbation in $M_{2\ell}$ does not affect the top-R SVD of the (matricized) $M_{2\ell}$, if we have no guarantees on the gaps between the top R eigenvalues, the eigen*vectors* of the perturbed matrix can be quite different from those of $M_{2\ell}$. Now the FOOBI procedure performs non-trivial operations on these eigenvectors when setting up the linear system we mentioned in the previous paragraph. Showing that the solutions are close despite the systems being different is thus a technical issue we need to overcome.

2) Warm-up: the case of $\mathsf{Err} = 0$*:* Let us start by describing the algorithm in the zero error case. This case illustrates the main ideas behind the algorithm and generalizes the FOOBI procedure to arbitrary ℓ.

The algorithm starts by computing the SVD of the matricized $M_{2\ell}$ (i.e., UU^T). Thus we obtain matrices E and Λ such that $UU^T = E\Lambda E^T$. Let H denote the matrix $E\Lambda^{1/2}$. Then we have $HH^T = UU^T$, and thus there exists an orthogonal $R \times R$ matrix Q such that $U = HQ$. Thus, finding U now reduces to finding the orthogonal matrix Q.

This is done using in a clever manner using the rank-1 detecting device Φ. Intuitively, if we wish to find one of the columns of U, we may hope to find a linear combination $\sum_j \alpha_j H_j$ of the $\{H_j\}$ such that $\Phi(\sum_j \alpha_j H_j, \sum_j \alpha_j H_j) = 0$. Each column of Q would provide a candidate solution α. However, this is a quadratic system of equations in α_i, and it is not clear how to solve the system directly.

The main idea in [16] is to find an alternate way of computing Q. The first observation is that Φ is bi-linear (i.e., linear in its arguments X, Y). Thus, we have $\Phi(\sum_j \alpha_j H_j, \sum_j \alpha_j H_j) = \sum_{i,j\in[R]} \alpha_i \alpha_j \Phi(H_i, H_j)$. Now, consider the *linear* system of equations

$$\sum_{i,j\in[R]} W_{ij}\Phi(H_i, H_j) = 0. \tag{27}$$

This is a system of $n^{2\ell}$ equations in R^2 variables. The reasoning above shows that for every column Q_i of Q, we have that $W = Q_i Q_i^T$ is a solution to (27). Because of linearity, this means that for any diagonal matrix $\mathbf{D}$, $Q\mathbf{D}Q^T$ is a solution to the linear system as well. The main observation of [16] is now that any symmetric solution W (i.e. one that satisfies $W_{ij} = W_{ji}$) is of this form! Thus, the matrix Q can be computed by simply finding a "typical" symmetric solution W and computing its eigen-decomposition. Let us now formalize the above.

Lemma VIII.2. *[16] The space of symmetric solutions to the system of equations* (27) *has dimension precisely R, and any solution is of the form $W = Q\mathbf{D}Q^T$, for some diagonal matrix $\mathbf{D}$.*

Proof. Consider any symmetric solution W. Because of bi-linearity, using the fact that $HQ = U$, or $H = UQ^T$, we have that $H_i = \sum_s U_s (Q^T)_{si} = \sum_s U_s Q_{is}$. Thus for any i, j,

$$\Phi(H_i, H_j) = \sum_{s,t} Q_{is} Q_{jt} \cdot \Phi(U_s, U_t).$$

Thus,

$$\begin{aligned}\sum_{i,j} W_{ij}\Phi(H_i, H_j) &= \sum_{s,t} \Phi(U_s, U_t) \cdot \sum_{i,j} W_{ij} Q_{is} Q_{jt} \\ &= \sum_{s,t} \Phi(U_s, U_t) \langle W, Q_s Q_t^T \rangle. \end{aligned} \tag{28}$$

Since $\kappa_M < \infty$, we have that $\{\Phi(U_s, U_t) : s < t\}$ is linearly independent. Now, since $\Phi(U_s, U_t) = \Phi(U_t, U_s)$, and since $\Phi(U_s, U_t) \neq 0$ for all $s \neq t$ (the latter is a simple computation, using the fact that $A_s \neq A_t$), we must have that

$$\text{for all } s \neq t, \quad \langle W, Q_s Q_t^T \rangle = 0.$$

Now, since Q is an orthogonal matrix, we have that $\{Q_s Q_t^T\}_{s,t\in[R]}$ forms an orthonormal basis for all $R \times R$ matrices. The above equality thus means that W lies only in the span of $\{Q_s Q_s^T\}_{s\in[R]}$. This implies that $W = Q\mathbf{D}Q^T$, for some diagonal matrix $\mathbf{D}$.

Plugging back into (28), we see that any W of this form satisfies the equation. As the $Q_s Q_s^T$ are all orthogonal, we have found a solution space of dimension precisely R. □

To handle the robust case, we also need a slight extension of the lemma above. Let H_Φ denote a matrix that has $R(R+1)/2$ columns, described as follows. The columns correspond to pairs $i, j \in [R]$, for $i \leq j$. For $i = j$, the corresponding column is $\Phi(H_i, H_i)$ and for $i < j$, the corresponding column is $\sqrt{2} \cdot \Phi(H_i, H_j)$. We note that the null space of H_Φ can be mapped in a one-one manner to symmetric $R \times R$ matrices W. For any $z = (z_{ij})_{i\leq j}$, define the symmetric $R \times R$ matrix $\psi(z)$ to have $\psi(z)_{ii} = z_{ii}$ and $\psi(z)_{ij} = \psi(z)_{ji} = \frac{z_{ij}}{\sqrt{2}}$. The

point of this definition is that $\langle z, z' \rangle = \langle \psi(z), \psi(z') \rangle$. Note that ψ^{-1} is well-defined, and that it takes symmetric matrices to $R(R+1)/2$-dimensional vectors (and preserves dot-products).

Further, we have

$$\begin{aligned} H_\Phi z &= \sum_i \Phi(H_i, H_i) z_{ii} + \sum_{i<j} \sqrt{2} \cdot \Phi(H_i, H_j) z_{ij} \\ &= \sum_i \Phi(H_i, H_i) z_{ii} + \sum_{i<j} 2\Phi(H_i, H_j) \psi(z)_{ij} \\ &= \sum_{i,j} \Phi(H_i, H_j) \psi(z)_{ij}. \end{aligned} \tag{29}$$

Using this correspondence, Lemma VIII.2 implies that H_Φ has a null space of dimension precisely R (corresponding to the span of $\psi^{-1}(Q_s Q_s^T)$, for $s \in [R]$). We now claim something slightly stronger.

Lemma VIII.3. *Let λ denote the $(R+1)$th smallest singular value of H_Φ. We have that $\lambda \geq \sigma_{\min}(M_\Phi)$. Recall that M_Φ was defined to be the matrix with columns $\Phi(A_i^{\otimes \ell}, A_j^{\otimes \ell})$, for $i < j$.*

Proof. Consider any z orthogonal to span$\{\psi^{-1}(Q_s Q_s^T) : s \in [R]\}$. Then, $\psi(z)$ is orthogonal to $Q_s Q_s^T$ for all s, as ψ preserves dot-products. Thus, using our earlier observation that $\{Q_s Q_t^T\}$ forms an orthonormal basis for all $R \times R$ matrices, we can write

$$\psi(z) = \sum_{s \neq t} \alpha_{st} Q_s Q_t^T.$$

Since $\psi(z)$ is symmetric, we also have $\alpha_{st} = \alpha_{ts}$. Now, using the expansion (28) with $W_{ij} = \psi(z)_{ij}$, we have

$$\begin{aligned} \sum_{i,j} \psi(z)_{ij} \Phi(H_i, H_j) &= \sum_{s,t} \alpha_{st} \Phi(U_s, U_t) \\ &= 2 \sum_{s<t} \alpha_{st} \Phi(U_s, U_t). \end{aligned}$$

Combining this with (29), and the definition of the smallest singular value, we obtain

$$\|H_\Phi z\|_F^2 \geq 4 \left(\sum_{s<t} \alpha_{st}^2 \right) \sigma_{\min}(M_\Phi)^2.$$

Finally, since $\|z\|^2 = \|\psi(z)\|_F^2 = 2 \sum_{s<t} \alpha_{st}^2$, the desired conclusion follows (indeed with a slack of a factor $\sqrt{2}$). □

The following theorem then gives the algorithm to recover Q, in the case $\mathsf{Err} = 0$.

Theorem VIII.4. *Let S be the subspace (of $\mathbb{R}^{R \times R}$) of all symmetric solutions to the system of equations $\sum_{ij} W_{ij} \Phi(H_i, H_j) = 0$. Let Z be a uniformly random Gaussian vector in this subspace of unit variance in each direction. Then with probability at least $9/10$, we have that*

$$Z = \sum_i \alpha_i Q_i Q_i^T, \text{ where } \min_{i \neq j} |\alpha_i - \alpha_j| \geq \frac{1}{20R^2}.$$

Thus the SVD of Z efficiently recovers the Q_i, with probability $\geq 9/10$.

Proof. From the lemmas above, we have that the space S is precisely the span of $Q_s Q_s^T$, for $s \in [R]$. These are all orthogonal vectors. Thus a random Gaussian vector in this space with unit variance in each direction is of the form $\sum_i \alpha_i Q_i Q_i^T$, where the α_i are independent and distributed as the univariate Gaussian $\mathcal{N}(0,1)$.

Now, for any i, j, we have that $\alpha_i - \alpha_j$ is distributed as $\mathcal{N}(0,2)$, and thus

$$\mathbb{P}\left[|\alpha_i - \alpha_j| \leq \frac{1}{20R^2} \right] \leq \frac{1}{20R^2}.$$

Taking a union bound over all pairs i, j now gives the result. □

This completes the algorithm for the case $\mathsf{Err} = 0$. Let us now see how to extend this analysis to the case in which $\mathsf{Err} \neq 0$.

3) A robust analysis: We will now prove an approximate recovery bound by following the above analysis, when Err is non-zero (but still small enough, as in the statement of Theorem VIII.1). As is common in such analyses, we will use the classic Davis-Kahan Sin-θ theorem. We start by recalling the theorem. To do so, we need some notation.

Suppose V_1 and V_2 are two $n \times d$ matrices with orthonormal columns. Then the matrix of *principal angles* between the column spans of V_1 and V_2 is denoted by $\Theta(V_1, V_2)$, and is defined to be the diagonal matrix whose entries are $\arccos(\lambda_i)$, where λ_i are the singular values of $V_1^T V_2$.

Theorem VIII.5 (Sin-θ theorem, [45])**.** *Let Σ and $\Sigma' \in \mathbb{R}^{n \times n}$ be symmetric, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n$ and $\lambda'_1 \geq \lambda'_2 \geq \cdots \geq \lambda'_n$. Let $1 \leq r \leq s \leq n$, and let $d = s - r + 1$. Let V be a matrix with columns being the eigenvectors corresponding to $\lambda_r \ldots \lambda_s$ in Σ, and suppose V' is similarly defined. Let*

$$\begin{aligned} \delta := \inf\{ &|\lambda' - \lambda| : \lambda \in [\lambda_s, \lambda_r], \\ &\lambda' \in (-\infty, \lambda'_{s+1}] \cup [\lambda'_{r-1}, \infty)\}, \end{aligned}$$

which we assume is > 0. Then we have

$$\|\sin \Theta(V, V')\|_F \leq \frac{\|\Sigma - \Sigma'\|_F}{\delta}.$$

Furthermore, there exists an orthogonal matrix O' such that

$$\|V - V'O'\|_F \leq \frac{\sqrt{2}\|\Sigma - \Sigma'\|_F}{\delta}. \tag{30}$$

We note that the precise statement above is from [**?**]. Our proof will follow the outline of the $\mathsf{Err} = 0$ case. The first step is to symmetrize the matricized version of T, so that we can take the SVD. We have the following simple observation.

Lemma VIII.6. *Let $A \in \mathbb{R}^{n\times n}$, and define $A' = (A + A^T)/2$. Let $B \in \mathbb{R}^n$ be symmetric. Then $\|A' - B\|_F \leq \|A - B\|_F$.*

Proof. The lemma follows from observing that A' is the projection of A onto the linear space of symmetric $n \times n$ matrices, together with the fact that projections to convex sets only reduces the distance. □

Let T' be the symmetric version of the matricized version of T. Then we have $\|T' - UU^T\|_F \leq \|\mathsf{Err}\|_F$. Likewise, let $\widehat{T}$ be the projection of T' onto the PSD cone (we obtain $\widehat{T}$ by computing the SVD and zero'ing out all the negative eigenvalues). By the same reasoning, we have $\|\widehat{T} - UU^T\|_F \leq \|\mathsf{Err}\|_F$. For convenience, in what follows, we denote $\|\mathsf{Err}\|_F$ by η.

Next, we need a simple lemma that relates the error in a square root to the error in a matrix.

Lemma VIII.7. *Let Z and H be $n \times d$ matrices with $d \leq n$, and suppose $\|ZZ^T - HH^T\| \leq \delta$. Then there exists an orthogonal matrix Q such that*

$$\|ZQ - H\|_F \leq (d\delta)^{1/2} + \frac{2\delta d\sigma_1(H)}{\sigma_d(H)^2}.$$

Proof. Let $ZZ^T = M_1\Sigma_1 N_1^T$, and let $HH^T = M_2\Sigma_2 N_2^T$, where M_i, N_i are $n \times d$ matrices with orthonormal columns. Now, the theory of operator-monotone functions acting on PSD matrices (see e.g. [**?**], Theorem X.1.1) implies that

$$\|M_1\Sigma_1^{1/2}N_1^T - M_2\Sigma_2^{1/2}N_2^T\|_F \leq \delta^{1/2}.$$

Now we may apply the Sin-θ theorem (with $r = 1$ and $s = d$ in the statement above) to conclude that there exists an orthogonal matrix Q_1 such that $\|N_1Q_1 - N_2\|_F \leq \frac{\sqrt{2}\ \delta}{\sigma_d(H)^2}$. Thus, writing $N_2 = N_1Q_1 + \Delta$, the LHS above becomes

$$\|M_1\Sigma_1^{1/2}N_1^T - M_2\Sigma_2^{1/2}Q_1^TN_1^T - M_2\Sigma_2^{1/2}\Delta^T\|_F.$$

Now, we have $\|M_2\Sigma_2^{1/2}\Delta^T\|_F \leq \|M_2\Sigma_2^{1/2}\|_F\|\Delta\|_F$. The first term is simply $(\mathrm{tr}(\Sigma_2))^{1/2} \leq d^{1/2}\sigma_1(H)$. Using this, we obtain

$$\begin{aligned}\|(M_1\Sigma_1^{1/2} - M_2\Sigma_2^{1/2}Q_1^T)N_1^T\|_F &= \|M_1\Sigma_1^{1/2}N_1^T - M_2\Sigma_2^{1/2}N_2^T + M_2\Sigma_2^{1/2}\Delta^T\|_F \\ &\leq \delta^{1/2} + \frac{2\delta d^{1/2}\sigma_1(H)}{\sigma_d(H)^2}.\end{aligned}$$

We can now appeal to the simple fact that for a matrix X, for any N_2 with orthonormal columns, we have $\|X\|_F = \|XN_2^TN_2\| \leq \|XN_2^T\|_F d^{1/2}$. This gives us

$$\|M_1\Sigma_1^{1/2} - M_2\Sigma_2^{1/2}Q_1^T\|_F \leq (d\delta)^{1/2} + \frac{2\delta d\sigma_1(H)}{\sigma_d(H)^2}.$$

Thus, since $Z = M_1\Sigma_1^{1/2}Q'$ for an orthogonal matrix Q' and likewise for H, and because the product of orthogonal matrices is orthogonal, we have the desired result. □

In what follows, to simplify the notation, we introduce the following definition.

Definition VIII.8 (Poly-bounded function)**.** We say that a function f of a parameter η is *poly-bounded* if $f(\eta)$ is of the form $\eta^c \cdot \mathrm{poly}(n, R, \kappa_U, \kappa_M)$, where $c > 0$ is a constant.

Intuitively speaking, by choosing η to be "polynomially small" in n, R and the condition numbers κ_U, κ_M, we can make $f(\eta)$ arbitrarily small.

Now, the lemma above gives the following as a corollary.

Corollary VIII.9. *Let $\widehat{E}\widehat{\Lambda}\widehat{E}^T$ be the rank-R SVD of $\widehat{T}$, and let $UU^T = E\Lambda E^T$ be the SVD as before. Define $\widehat{H} = \widehat{E}\widehat{\Lambda}^{1/2}$ and $H = E\Lambda^{1/2}$. Then there exists an orthogonal matrix P such that $\|\widehat{H}P - H\|_F \leq f_1(\eta)$ for some poly-bounded function f_1.*

Proof. The desired conclusion follows from Lemma VIII.7 if we show that $\|\widehat{E}\widehat{\Lambda}\widehat{E}^T - E\Lambda E^T\| \leq 2\eta$. This follows from the fact that $\|\widehat{T} - \widehat{E}\widehat{\Lambda}\widehat{E}^T\| \leq \eta$ (which is true because the SVD gives the closest rank-k matrix to $\widehat{T}$ — and UU^T is at distance at most η), together with the triangle inequality. □

Informally speaking, we have shown that $\widehat{H}P \approx_\eta H$, for an orthogonal matrix P. We wish to now use our machinery from Section VIII-A2 to find the matrix U, which will then allow us to obtain the vectors in the decomposition.

Let us define $H' = HP^T$, where P is as above. Thus we have $H = H'P$ (and thus $H' \approx_\eta \widehat{H}$, informally). Further, if Q is the orthogonal matrix such that $U = HQ$ (as in Section VIII-A2), we have $U = H'PQ$.

a) Outline of the remainder.: We first sketch the rest of the argument. The key idea is the following: suppose we run the whole analysis in Section VIII-A2 using the matrices H' and PQ instead of H and Q, we obtain that the set of symmetric solutions to the system of equations $\sum_{i,j\in[R]} W_{ij}\Phi(H'_i, H'_j)$ is precisely the span of the matrices $(PQ)_s(PQ)_s^T$. Thus, a random matrix in the space of symmetric solutions can be diagonalized to obtain $(PQ)_s$. Using $U = H'(PQ)$, one can reconstruct U. Now, we have access to $\widehat{H}$ and not H'. However, we can relate the space of symmetric *approximate solutions* to the perturbed system to the original one in a clean way. Taking a random matrix in this space, and utilizing the "gap" in Theorem VIII.4, we obtain the matrix PQ approximately. This is then used to find $\widehat{U}$ that approximates U, completing the argument.

Lemma VIII.10. *For any $i, j \in [R]$, we have*

$$\|\Phi(H'_i, H'_j) - \Phi(\widehat{H}_i, \widehat{H}_j)\| \leq \\ \leq O\left(\|H'_i - \widehat{H}_i\|\|H'_j\| + \|H'_j - \widehat{H}_j\|\|\widehat{H}_i\|\right).$$

Proof.

$$\|\Phi(H'_i, H'_j) - \Phi(\widehat{H}_i, \widehat{H}_j)\| \leq \\ \leq \|\Phi(H'_i, H'_j) - \Phi(\widehat{H}_i, H'_j)\| \\ + \|\Phi(\widehat{H}_i, H'_j) - \Phi(\widehat{H}_i, \widehat{H}_j)\|.$$

The first term can be bounded by $2\|H'_j\|\|H'_i - \widehat{H}_i\|$, and so also the second term is bounded by $2\|\widehat{H}_i\|\|H'_i - \widehat{H}_i\|$, which implies the lemma. □

Next, as in Section VIII-A2, define the $R(R+1)/2$ dimensional matrices $\widehat{H}_\Phi$ and H'_Φ. Specifically, these matrices have columns corresponding to pairs $1 \leq i \leq j \leq R$, and for $i = j$, the corresponding column of H'_Φ is $\Phi(H'_i, H'_i)$ and for $i \neq j$, the column is $\sqrt{2}\cdot\Phi(H'_i, H'_j)$. A simple corollary of the lemma above is that

$$\|\widehat{H}_\Phi - H'_\Phi\|_F \leq O\left(\|\widehat{H} - H'\|_F \cdot (\|\widehat{H}\|_F + \|H'\|_F)\right) \\ = f_2(\eta), \tag{31}$$

for some poly-bounded function f_2. This follows from the lemma and corollary above, together with an application of the Cauchy-Schwarz inequality. Next, we show the following.

Lemma VIII.11. *For $1 \leq r \leq R$, we have $\sigma_r(\widehat{H}_\Phi) \leq f_2(\eta)$. Also, we have $\sigma_{R+1}(\widehat{H}_\Phi) \geq \sigma_{\min}(M_\Phi) - f_2(\eta)$.*

Proof. The main idea, as mentioned in the outline, is to apply Lemma VIII.3 to H'. If λ' denotes the $(R+1)$th smallest singular value of H'_Φ, then this lemma implies that H'_Φ has R zero singular values and $\lambda' \geq \sigma_{\min}(M_\Phi)$. Weyl's inequality[7] now immediately implies the lemma. □

From now on, suppose that η is chosen small enough that $f_2(\eta) < \frac{\sigma_{\min}(M_\Phi)}{2}$. Next, let us define the spaces S' and $\widehat{S}$ as in Theorem VIII.4: let S' be the space of all symmetric solutions to the linear system

$$\sum_{i,j} W_{ij}\Phi(H'_i, H'_j) = 0.$$

Likewise, let $\widehat{S}$ be the space of symmetric matrices $\psi(z)$ (see Section VIII-A2 for the definition of ψ), where z is in the span of the R smallest singular values of $\widehat{H}_\Phi$. The analog of Theorem VIII.4 is the following.

Theorem VIII.12. *Let Z be a uniformly random Gaussian vector in $\widehat{S}$, and suppose that $Z = G\Sigma G^T$ is the SVD of Z. Then with probability $\geq 9/10$, we have $\|G - PQ\|_F \leq f_3(\eta)$, for some poly-bounded function f_3.*

Proof. The first step is to show that the spaces S' and $\widehat{S}$ are close. This is done via the Sin-θ theorem, applied to the matrices $(H'_\Phi)^T H'_\Phi$ and $\widehat{H}_\Phi^T \widehat{H}_\Phi$. Let T' and $\widehat{T}$ be the spans of the smallest R singular vectors of the two matrices. By Theorem VIII.5 and the bounds on σ_{R+1}, we have that there exist orthonormal bases Υ and $\widehat{\Upsilon}$ for these spaces such that for some orthonormal matrix Q',

$$\|\Upsilon Q' - \widehat{\Upsilon}\|_F \leq \frac{\|(H'_\Phi)^T H'_\Phi - \widehat{H}_\Phi^T \widehat{H}_\Phi\|_F}{\sigma_{\min}(M_\Phi)^2}.$$

Now, appealing to the simple fact that for any two matrices X, Y, $\|X^T X - Y^T Y\|_F \leq \|X^T(X - Y) + (X^T - Y^T)Y\|_F \leq \|X - Y\|_F(\|X\|_F + \|Y\|_F)$, we can bound the quantity above by $f_4(\eta)$ for some poly-bounded function f_4.

We can now obtain bases for $\widehat{S}$ and S' by simply applying ψ to the columns of the matrices $\widehat{\Upsilon}$ and Υ respectively. Let us abuse notation slightly and call these bases $\widehat{S}$ and S' as well. By properties of ψ, we have that

$$\|S'Q' - \widehat{S}\|_F \leq \|\Upsilon Q' - \widehat{\Upsilon}\|_F \leq f_4(\eta). \tag{32}$$

[7]Recall that the inequality bounds the change in eigenvalues due to a perturbation of a matrix by the spectral norm (and hence also the Frobenius norm) of the perturbation.

Next, note that a random unit Gaussian vector in the space $\widehat{S}$ can be viewed as first picking $v \in \mathcal{N}(0,1)^R$ and taking $\widehat{S}v$. Now, using Theorem VIII.4 if we consider the matrix $S'v$ (which is a random Gaussian vector in the space S'), with probability at least $9/10$, we have an eigenvalue gap of at least $\frac{1}{20R^2}$. Thus, using this and (32), together with the Sin-θ theorem (used this time with precisely one eigenvector, and thus the rotation matrix disappears), we have that $\|G_i - (PQ)_i\| \leq 20R^2 f_4(\eta)$. Summing over all i (after taking the square), the theorem follows. □

We can now complete the proof of the main theorem of this section.

Proof of Theorem VIII.1. Theorem VIII.12 shows that the matrix G gives a good approximation to the rotation matrix (PQ) with probability $9/10$. (Since this probability is over the randomness in the algorithm, we can achieve a probability of $1-\gamma$ by running the algorithm $O(\log 1/\gamma)$ times.) We now show that $\widehat{H}G \approx H'PQ$:

$$\begin{aligned}\|\widehat{H}G - H'PQ\|_F &\leq \|\widehat{H}(G-PQ) + (\widehat{H}-H')PQ\|_F \\ &\leq f_5(\eta). \end{aligned} \tag{33}$$

Note now that $H'PQ$ is precisely U! Thus the matrix $\widehat{U} := \widehat{H}G$ (which we can compute as discussed above) is an approximation up to an error $f_5(\eta)$. Finally, to obtain a column U_i of U, we reshape $\widehat{U}$ into an $n \times n^{\ell-1}$ matrix, apply an SVD, and output the top left-singular-vector. This yields an error $f_6(\eta)$, for some poly-bounded function of η. □

B. Smoothed analysis

Finally, we show that Theorem VIII.1 can be used with our earlier results to show the following.

Theorem VIII.13. *Suppose $T = \sum_{i\in[R]} \tilde{A}_i^{\otimes 2\ell} + E$, where $\{A_i\}$ have polynomially bounded length. Given an accuracy parameter ε and any $0 < \delta < 1/\ell^2$, with probability at least $1 - \exp(-\Omega_\ell(n))$ over the perturbation in $\tilde{A}$, there is an efficient algorithm that outputs a set of vectors $\{B_i\}_{i=1}^R$ such that*

$$\min_\pi \sum_i \|\tilde{A}_i - B_{\pi(i)}\| \leq \varepsilon,$$

as long as $R \leq \delta n^\ell$, and $\|E\|_F \leq \mathrm{poly}(1/n, \rho, \varepsilon)$, for an appropriate polynomial in the arguments.

The proof of this theorem goes via the robust decomposition algorithm presented in Theorem VIII.1. In order to use the theorem, we need to bound the two condition numbers κ_U and κ_M. Since the columns of A are polynomially bounded, the columns of U and M_Φ are as well, so $\sigma_{max}(U)$, $\sigma_{max}(M_\Phi)$ are bounded by some polynomial in n. Therefore we only need to give lower bounds on $\sigma_{min}(U)$ and $\sigma_{min}(M_\Phi)$. We now use Theorem II.2 to prove that these quantities are both polynomially bounded with high probability in a smoothed analysis setting. This would complete the proof of Theorem VIII.13.

Lemma VIII.14. *Let $U = \widetilde{A}^{\odot \ell}$, and M_Φ be the matrix whose columns are indexed by pairs $i, j \leq R$, and whose $\{i,j\}$'th column is $\Phi(\widetilde{A}_i^{\otimes\ell}, \widetilde{A}_j^{\otimes\ell})$. Then for $R \leq n^\ell/\ell^2$, with probability at least $1-\exp(-\Omega_\ell(n))$, we have both $\sigma_R(U)$ and $\sigma_{R(R-1)/2}(M_\Phi)$ to be $\geq \mathrm{poly}(1/n, \rho)$.*

Proof. The desired inequality for the matrix U was already shown in earlier sections. Let us thus consider M_Φ. We can write the $\{i,j\}$'th column as

$$\begin{aligned}(M_\Phi)_{i,j} = &(\widetilde{a}_i^{\otimes\ell} \otimes \widetilde{a}_j^{\otimes\ell}) \\ &- (\widetilde{a}_i^{\otimes(\ell-1)} \otimes \widetilde{a}_j \otimes \widetilde{a}_j^{\otimes(\ell-1)} \otimes \widetilde{a}_i) \\ &+ (\widetilde{a}_j^{\otimes\ell} \otimes \widetilde{a}_i^{\otimes\ell}) \\ &- (\widetilde{a}_j^{\otimes(\ell-1)} \otimes \widetilde{a}_i \otimes \widetilde{a}_i^{\otimes(\ell-1)} \otimes \widetilde{a}_j).\end{aligned}$$

We will show a stronger statement, namely that a matrix with four different columns (corresponding to each term above) for each pair $\{i,j\}$ has $\sigma_{\min} \geq \mathrm{poly}(1/n, \rho)$. In this matrix, which we call M'_Φ, we have two columns for every (ordered) pair (i,j). The first column is $\widetilde{a}_i^{\otimes\ell} \otimes \widetilde{a}_j^{\otimes\ell}$ and the second is $\widetilde{a}_i^{\otimes(\ell-1)} \otimes \widetilde{a}_j \otimes \widetilde{a}_j^{\otimes(\ell-1)} \otimes \widetilde{a}_i$.

For any of the columns, we thus have

$\Delta_2 = 1$ (same i,j, different terms)
$\Delta_\ell = R-1$ (same i, different j)
$\Delta_{\ell+1} = R-1$ (same i, different j, different terms)
$\Delta_{2\ell-2} = 1$ (i and j swapped, different terms)
$\Delta_{2\ell-1} = R-1$ (same i, different j, different terms)
$\Delta_{2\ell} = R^2$.

The rest of the Δ values are zero. Thus, we observe that we can apply Theorem V.2 (with $c = \Omega(1)$), as the dominant terms are the ones corresponding to $\Delta_2, \Delta_\ell, \Delta_{2\ell}$. This completes the proof. □

APPENDIX

Lemma A.1. *Let X, Y be two independent real random variables, for all $a, b \in \mathbb{R}$ such that $\mathbb{P}[X+Y \le b] > 0$, we have*

$$\mathbb{P}[X \le a] \le \mathbb{P}[X \le a | X+Y \le b]$$

Proof. WLOG, assume $\mathbb{P}[X \le a] > 0, \mathbb{P}[X > a] > 0$, then we have

$$\begin{aligned}\mathbb{P}[X+Y \le b | X \le a] &\ge \mathbb{P}[Y \le b-a | X \le a] \\ &= \mathbb{P}[Y \le b-a] \\ &= \mathbb{P}[Y \le b-a | X > a] \\ &\ge \mathbb{P}[X+Y \le b | X > a]\end{aligned}$$

Hence $\mathbb{P}[X+Y \le b | X \le a] \ge \mathbb{P}[X+Y \le b]$, then

$$\begin{aligned}\mathbb{P}[X \le a | X+Y \le b] &= \frac{\mathbb{P}[X \le a]\mathbb{P}[X+Y \le b | X \le a]}{\mathbb{P}[X+Y \le b]} \\ &\ge \mathbb{P}[X \le a]\end{aligned}$$

□

The proof of Theorem IV.2 is almost identical to Theorem IV.1.

Proof of Theorem IV.2. By Proposition IV.5, it suffices to show that

$$\mathbb{P}[\|\hat{g}(u+z_0, z_1, \cdots, z_{\ell-1})\|_2 < c(\ell)\varepsilon\eta \cdot \frac{\rho^\ell}{n^\ell}] < \varepsilon^{c'(\ell)\delta n}$$

where $z_0 \sim N(0, \rho^2(\ell+1)/(2n\ell))^n$ and $z_1, z_2, \cdots, z_{\ell-1} \sim N(0, \rho^2/(2n\ell))$, $c(\ell), c'(\ell) > 0$ are constants depending only on ℓ. Let W be the span of the top δn^ℓ right singular vectors of M. Observe that

$$\begin{aligned}&\|\hat{g}(u+z_0, z_1, \cdots, z_{\ell-1})\|_2 = \\ &\quad = \|M(u+z_0) \otimes z_1 \otimes \cdots \otimes z_{\ell-1}\|_2 \\ &\quad \ge \eta \|\Pi_W\ (u+z_0) \otimes z_1 \otimes \cdots \otimes z_{\ell-1}\|_2.\end{aligned}$$

The theorem then follows by applying Lemma III.3 with $x_1 = u, x_2 = x_3 = \cdots = x_\ell = 0$ and $p = \varepsilon^{1/\ell}$.

□

We now give a self-contained combinatorial proof of Theorem IV.1 for $\ell = 2$, that uses decoupling and Lemma III.3. Let D' be the dimension of the subspace W, and let $M_1, M_2, \ldots, M_{D'}$ be a basis for W. Let $z \sim N(0, \rho^2)^n$ and $z_1, z_2, \ldots, z_r \sim N(0, \rho^2/r)^n$ be independent Gaussian random vectors for $r = O(\sqrt{n})$. Note that $x + z_1 \pm z_2 \pm z_3 \pm \cdots \pm z_r$ are all identically distributed as $\tilde{x}$.

Consider the following process for generating $\tilde{x} = x + z$. We first generate $z_1, z_2, \ldots, z_r$ and random signs $\zeta = (\zeta_2, \zeta_3, \ldots, \zeta_r) \in \{\pm 1\}^{r-1}$ all independently, and return $z = z_1 + \sum_{i=2}^r \zeta_2 z_2$. It is easy to see that $z \sim N(0, \rho^2)$. We will now prove that at most one of the 2^{r-1} signed combinations $z_1 \pm z_2 \pm \cdots \pm z_r$ has a non-negligible projection onto W.

Consider any fixed pair $\zeta, \zeta' \in \{\pm 1\}^{r-1}$, and let $u = z_1 + \sum_{i=2}^r \zeta_i z_i$ and $u' = z_1 + \sum_{i=2}^r \zeta_i' z_i$. We will use the basic decoupling Lemma IV.7 to show w.h.p. at least one of $\|\Pi_W u^{\otimes 2}\|_2$ or $\|\Pi_W (u')^{\otimes 2}\|_2$ is non-negligible. Using decoupling (with $\ell = 2$) in Lemma IV.7 we have for each $j \in [D']$

$$\begin{aligned}\left\langle M_j, (x+u)^{\otimes 2} \right\rangle - \left\langle M_j, (x+u')^{\otimes 2} \right\rangle &= \\ = 4 \left\langle M_j, (x+u+u') \otimes (u-u') \right\rangle &\\ = 4 \left\langle M_j, (x+v_1) \otimes v_2 \right\rangle, &\end{aligned} \tag{34}$$

$$\text{where } v_1 = z_1 + \sum_{2 \le i \le r : \zeta_i = \zeta_i'} \zeta_i z_i,$$

$$v_2 = \sum_{2 \le i \le r : \zeta_i \ne \zeta_i'} \zeta_i z_i.$$

Also from Lemma III.3, we have that the above decoupled product $(x+v_1) \otimes v_2$ has a non-negligible projection onto W; hence with probability at least $1 - \exp\big(-\Omega(\delta n)\big)$,

$$\begin{aligned}\|\Pi_W (x+v_1) \otimes v_2\|_2^2 &= \sum_{j=1}^{D'} \langle M_j, (x+v_1) \otimes v_2 \rangle^2 \\ &\ge \frac{\Omega(\rho^4)}{r^2 n^4}\end{aligned}$$

$$\text{i.e., } \exists j^* \in [D'] \text{ s.t. } |\langle M_{j^*}, (x+v_1) \otimes v_2 \rangle| \ge \frac{\Omega(\rho^2)}{rn^3}.$$

Applying (34) with the above inequality for j^*,

$$|\left\langle M_{j^*}, (x+u)^{\otimes 2} \right\rangle - \left\langle M_{j^*}, (x+u')^{\otimes 2} \right\rangle| \ge \frac{\Omega(\rho^2)}{rn^3}.$$

Hence,

$$\|\Pi_W (x+u)^{\otimes 2}\|_2 + \|\Pi_W (x+u')^{\otimes 2}\|_2 \ge \Omega\Big(\frac{\rho^2}{rn^3}\Big), \tag{35}$$

with probability at least $1 - \exp(-\Omega(\delta n))$.

Since $r = c_1 \delta n$ (for a sufficiently small constant $c_1 > 0$), we can apply (35) for each of the 2^{2r-1} pairs

of $\zeta, \zeta' \in \{\pm 1\}^{r-1}$, and union bound over them to conclude that with probability at least $1-\exp(-\Omega(\delta n))$,

$$\forall \zeta \neq \zeta' \in \{\pm 1\}^{r-1},$$
$$\max\Big\{\big|\langle M_j, (x + z_1 + \sum_{i=2}^{r} \zeta_i z_i)^{\otimes 2}\rangle\big|,$$
$$\big|\langle M_j, (x + z_1 + \sum_{i=2}^{r} \zeta'_i z_i)^{\otimes 2}\rangle\big|\Big\} \geq \frac{\Omega(\rho^2)}{rn^3}.$$

Hence w.h.p. at most one of the 2^{r-1} signed combinations $x + z_1 \pm z_2 \pm \cdots \pm z_r$ has a negligible projection onto W. Hence, with probability at least $1 - 2^{-r+1}$ i.e., with probability at least $1 - 2^{-\Omega(\delta n)}$, $\|\Pi_W \tilde{x}^{\otimes 2}\|_2 \geq \Omega(\rho^2)/n^4$. This establishes Theorem IV.1. An identical proof also works for Theorem IV.2 when $\ell = 2$.

ACKNOWLEDGMENTS

We would like to thank Anindya De for several helpful discussions, particularly those that led to the algorithm for robust subspace recovery. The second, third and last authors were supported by the National Science Foundation (NSF) under Grant No. CCF-1652491 and CCF-1637585. Additionally, the third author was supported by an undergraduate research grant from Northwestern University.

REFERENCES

[1] D. A. Spielman and S.-H. Teng, "Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time," *J. ACM*, vol. 51, no. 3, pp. 385–463, May 2004. [Online]. Available: http://doi.acm.org/10.1145/990308.990310

[2] R. Beier and B. Vöcking, "Typical properties of winners and losers in discrete optimization," *SIAM J. Comput.*, vol. 35, no. 4, pp. 855–881, 2006. [Online]. Available: https://doi.org/10.1137/S0097539705447268

[3] A. Moitra and R. O'Donnell, "Pareto optimal solutions for smoothed analysts," in *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, ser. STOC '11. New York, NY, USA: ACM, 2011, pp. 225–234. [Online]. Available: http://doi.acm.org/10.1145/1993636.1993667

[4] M. Etscheid and H. Röglin, "Smoothed analysis of local search for the maximum-cut problem," *ACM Trans. Algorithms*, vol. 13, no. 2, pp. 25:1–25:12, Mar. 2017. [Online]. Available: http://doi.acm.org/10.1145/3011870

[5] O. Angel, S. Bubeck, Y. Peres, and F. Wei, "Local max-cut in smoothed polynomial time," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2017. New York, NY, USA: ACM, 2017, pp. 429–437. [Online]. Available: http://doi.acm.org/10.1145/3055399.3055402

[6] A. T. Kalai, A. Samorodnitsky, and S. Teng, "Learning and smoothed analysis," in *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, Oct 2009, pp. 395–404.

[7] J. Håstad, "Tensor rank is np-complete," *Journal of Algorithms*, vol. 11, no. 4, pp. 644–654, 1990.

[8] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan, "Smoothed analysis of tensor decompositions," in *Proceedings of the 46th Symposium on Theory of Computing (STOC)*. ACM, 2014.

[9] A. Moitra and G. Valiant, "Settling the polynomial learnability of mixtures of Gaussians," in *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*. IEEE, 2010, pp. 93–102.

[10] D. Hsu and S. M. Kakade, "Learning Gaussian mixture models: Moment methods and spectral decompositions," *arXiv preprint arXiv:1206.5766*, 2012.

[11] J. Anderson, M. Belkin, N. Goyal, L. Rademacher, and J. R. Voss, "The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures," in *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, 2014, pp. 1135–1164. [Online]. Available: http://jmlr.org/proceedings/papers/v35/anderson14.html

[12] R. Ge, Q. Huang, and S. M. Kakade, "Learning mixtures of Gaussians in high dimensions," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, 2015, pp. 761–770. [Online]. Available: http://doi.acm.org/10.1145/2746539.2746616

[13] N. Goyal, S. Vempala, and Y. Xiao, "Fourier PCA and robust tensor decomposition," in *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, 2014, pp. 584–593. [Online]. Available: http://doi.acm.org/10.1145/2591796.2591875

[14] T. Ma, J. Shi, and D. Steurer, "Polynomial-time tensor decompositions with sum-of-squares," in *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*. IEEE, 2016, pp. 438–446.

[15] N. Anari, C. Daskalakis, W. Maass, C. Papadimitriou, A. Saberi, and S. Vempala, "Smoothed analysis of discrete tensor decomposition and assemblies of neurons," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 880–10 890.

[16] J. . Cardoso, "Super-symmetric decomposition of the fourth-order cumulant tensor. blind identification of more sources than sensors," in *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, April 1991, pp. 3109–3112 vol.5.

[17] E. S. Allman, C. Matias, and J. A. Rhodes, "Identifiability of parameters in latent structure models with many observed variables," *The Annals of Statistics*, vol. 37, no. 6A, pp. 3099–3132, 2009.

[18] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *arXiv preprint arXiv:1210.7559*, 2012.

[19] V. Strassen, "Rank and optimal computation of generic tensors," *Linear Algebra and its Applications*, vol. 52, pp. 645 – 685, 1983. [Online]. Available: http://www.sciencedirect.com/science/article/pii/002437958380041X

[20] A. Carbery and J. Wright, "Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$," *Math. Res. Lett.*, vol. 8, no. 3, pp. 233–248, 2001. [Online]. Available: https://doi.org/10.4310/MRL.2001.v8.n3.a1

[21] T. Tao, "Topics in random matrix theory," *Book By Terry Tao*, 2011. [Online]. Available: http://terrytao.files.wordpress.com/2011/02/matrix-book.pdf

[22] M. Rudelson and R. Vershynin, "The littlewood–offord problem and invertibility of random matrices," *Advances in Mathematics*, vol. 218, no. 2, pp. 600 – 633, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001870808000224

[23] V. H. de la Pena and S. J. Montgomery-Smith, "Decoupling inequalities for the tail probabilities of multivariate u-statistics," *Ann. Probab.*, vol. 23, no. 2, pp. 806–816, 04 1995. [Online]. Available: http://dx.doi.org/10.1214/aop/1176988291

[24] S. Lovett, "An elementary proof of anti-concentration of polynomials in gaussian variables," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 17, p. 182, 2010. [Online]. Available: http://eccc.hpi-web.de/report/2010/182

[25] F. Nazarov, M. Sodin, and A. Vol'berg, "The geometric Kannan-Lovász-Simonovits lemma, dimension-free estimates for the distribution of the values of polynomials, and the distribution of the zeros of random analytic functions." *St. Petersbg. Math. J.*, vol. 14, no. 2, pp. 214–234, 2003.

[26] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American statistical association*, vol. 79, no. 388, pp. 871–880, 1984.

[27] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. John wiley & sons, 2005, vol. 589.

[28] D. L. Donoho and P. J. Huber, "The notion of breakdown point," *A festschrift for Erich L. Lehmann*, vol. 157184, 1983.

[29] M. Hardt and A. Moitra, "Algorithms and hardness for robust subspace recovery," in *Conference on Learning Theory*, 2013, pp. 354–375.

[30] S. R. Eddy, "Hidden markov models," *Current opinion in structural biology*, vol. 6, no. 3, pp. 361–365, 1996.

[31] M. Gales and S. Young, "The application of hidden markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[32] A. Anandkumar, D. Hsu, and S. M. Kakade, "A method of moments for mixture models and hidden markov models," *arXiv preprint arXiv:1203.0683*, 2012.

[33] A. Bhaskara, M. Charikar, and A. Vijayaraghavan, "Uniqueness of tensor decompositions with applications to polynomial identifiability," *Proceedings of the Conference on Learning Theory (COLT).*, 2014.

[34] Q. Huang, R. Ge, S. Kakade, and M. Dahleh, "Minimal realization problems for hidden markov models," 2015.

[35] V. Sharan, S. Kakade, P. Liang, and G. Valiant, "Learning overcomplete hmms," *CoRR*, vol. abs/1711.02309, 2017. [Online]. Available: http://arxiv.org/abs/1711.02309

[36] J. B. Kruskal, "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear algebra and its applications*, vol. 18, no. 2, pp. 95–138, 1977.

[37] R. A. Harshman, "Foundations of the parafac procedure: models and conditions for an explanatory multimodal factor analysis," 1970.

[38] B. Barak, J. A. Kelner, and D. Steurer, "Dictionary learning and tensor decomposition via the sum-of-squares method," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, ser. STOC '15. New York, NY, USA: ACM, 2015, pp. 143–151. [Online]. Available: http://doi.acm.org/10.1145/2746539.2746605

[39] S. B. Hopkins, T. Schramm, J. Shi, and D. Steurer, "Fast spectral algorithms from sum-of-squares proofs: Tensor decomposition and planted sparse vectors," in *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, ser. STOC '16. New York, NY, USA: ACM, 2016, pp. 178–191. [Online]. Available: http://doi.acm.org/10.1145/2897518.2897529

[40] L. De Lathauwer, J. Castaing, and J. Cardoso, "Fourth-order cumulant-based blind identification of underdetermined mixtures," *IEEE Trans. on Signal Processing*, vol. 55, no. 6, pp. 2965–2973, 2007.

[41] L. Albera, A. Ferréol, P. Comon, and P. Chevalier, "Blind identification of overcomplete mixtures of sources (biome)," *Linear Algebra and its Applications*, vol. 391, pp. 3 – 30, 2004, special Issue on Linear Algebra in Signal and Image Processing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0024379504002435

[42] P. Comon, G. Golub, L. Lim, and B. Mourrain, "Symmetric tensors and symmetric tensor rank," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1254–1279, 2008. [Online]. Available: https://doi.org/10.1137/060661569

[43] P.-Å. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, no. 1, pp. 99–111, 1972.

[44] W. F. de la Vega, M. Karpinski, R. Kannan, and S. Vempala, "Tensor decomposition and approximation schemes for constraint satisfaction problems," in *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*. ACM, 2005, pp. 747–754.

[45] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. III," *SIAM Journal on Numerical Analysis*, vol. 7, no. 1, pp. 1–46, mar 1970. [Online]. Available: