Spectral Methods from Tensor Networks

Ankur Moitra
Department of Mathematics, and Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, USA
moitra@mit.edu

Alexander S. Wein
Department of Mathematics, Courant Institute of Mathematical Sciences, New York University, USA
awein@cims.nyu.edu

ABSTRACT

A tensor network is a diagram that specifies a way to "multiply" a collection of tensors together to produce another tensor (or matrix). Many existing algorithms for tensor problems (such as tensor decomposition and tensor PCA), although they are not presented this way, can be viewed as spectral methods on matrices built from simple tensor networks. In this work we leverage the full power of this abstraction to design new algorithms for certain continuous tensor decomposition problems.

An important and challenging family of tensor problems comes from orbit recovery, a class of inference problems involving group actions (inspired by applications such as cryo-electron microscopy). Orbit recovery problems over finite groups can often be solved via standard tensor methods. However, for infinite groups, no general algorithms are known. We give a new spectral algorithm based on tensor networks for one such problem: continuous multi-reference alignment over the infinite group SO(2). Our algorithm extends to the more general heterogeneous case.

CCS CONCEPTS

• **Theory of computation → Algorithm design techniques**.

KEYWORDS

Tensor networks, spectral methods, orbit recovery, multi-reference alignment

ACM Reference Format:
Ankur Moitra and Alexander S. Wein. 2019. Spectral Methods from Tensor Networks. In *Proceedings of the 51st Annual ACM SIGACT Symposium on the Theory of Computing (STOC '19), June 23–26, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3313276.3316357

STOC '19, June 23–26, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6705-9/19/06…$15.00
https://doi.org/10.1145/3313276.3316357

1 INTRODUCTION

Algorithms for decomposing low-rank tensors have had a wide range of applications in machine learning and statistics. They can be leveraged to give efficient algorithms for phylogenetic reconstruction [27], topic modeling [4], community detection [5], independent component analysis [25] and learning various mixture models [22, 23]. However there are important families of problems where the low-order moment tensors are known to achieve statistically-optimal rates of estimation but there are no known *efficient* algorithms for finding the parameters from the moments.

The familiar symmetric third-order tensor decomposition problem asks: Given a $p \times p \times p$ low-rank tensor of the form

$$T = \sum_{i=1}^{r} a_i^{\otimes 3}$$

can we recover the vectors $a_1, \ldots, a_r \in \mathbb{R}^p$? When $r \leq p$ it is called the *undercomplete* case and when $r > p$ it is called the *overcomplete* case. In the undercomplete case, Jennrich's algorithm (see [25]) gives a polynomial time algorithm based on generalized eigendecompositions that works provided that the vectors $a_1, \ldots, a_r$ are linearly independent. In the overcomplete case, a line of work has culminated in a polynomial time algorithm that works when the vectors a_i are random (i.i.d. Gaussian) and $r \lesssim p^{3/2}$ [20, 21, 24]. In applications, the vectors $a_1, \ldots, a_r$ represent the parameters of a model we would like to learn and T represents moments of the distribution specified by the model whose entries we can estimate from samples.

However, in some applications the parameters are not uniquely defined, except up to equivalence under some *continuous group action*. This leads to a new sort of problem that we call *orbit tensor decomposition* in which we want to recover a vector $\theta \in \mathbb{R}^p$ given a tensor of the form

$$T = \int_{A \in \mathcal{A}} (A\theta)^{\otimes 3} dA$$

where $\mathcal{A}$ is a known, possibly infinite, set of $p \times p$ matrices (equipped with a measure over which to integrate). We assume furthermore that $\mathcal{A}$ possesses a particular group symmetry which results in nonuniqueness of the solution: θ and $A\theta$ are equally-good solutions for any $A \in \mathcal{A}$. There are important real-world applications such as cryo-electron microscopy (cryo-EM) [3, 28, 36] and multi-reference alignment (MRA) [1, 8, 11, 14, 18, 29, 31, 39] where these sort of tensor decomposition problems arise when using the method of moments. Here A is a random rotation of a two- or three-dimensional signal whose orientation we cannot control when we are measuring it. Despite considerable interest in such problems there are few algorithms with provable guarantees, in large part because working with the symmetries of the group is challenging *algorithmically*.

We will focus on the *continuous multi-reference alignment* (continuous MRA) problem which can be described as follows. The goal is to recover a signal θ which is a real-valued function on the unit circle in $\mathbb{R}^2$. We assume θ is band-limited so that in the Fourier basis we can think of θ as a finite-dimensional vector $\theta \in \mathbb{R}^p$. The compact group $G = \mathrm{SO}(2)$ (rotations in the plane) acts on θ by rotating the signal around the unit circle. For $g \in G$ and $\theta \in \mathbb{R}^p$ we denote the result of the rotation as $g \cdot \theta \in \mathbb{R}^p$. Now we observe many independent samples of the form $y_i = g_i \cdot \theta + \xi_i$ where g_i is a uniformly random element of SO(2) and ξ_i is i.i.d. Gaussian noise. *In other words, we observe many copies of the true signal that are both noisy and randomly-rotated.* It is known that for this problem (and a large class of similar problems), optimal sample complexity in the large-noise limit is achieved by the method of moments [1, 2, 8, 10]. First we use the samples to estimate the third moment $T = \int_{g \in G} (g \cdot \theta)^{\otimes 3} dg$. Recovering θ (up to equivalence under group action) is now an instance of the orbit tensor decomposition problem from above.

Existing tensor methods fail because (i) T is no longer low-rank. In fact T has an infinite number of components and when θ is generic would plausibly have essentially full rank. We can no longer hope to decompose T by finding a rank-one term that we can subtract off and lower the rank. Instead, we need to find a continuous collection of rank-one tensors at once! (ii) We can only hope to recover the *orbit* of θ, i.e. to recover a vector that (approximately) lies in the orbit $\{g \cdot \theta \ : \ g \in G\}$. This symmetry implies that any finite-rank decomposition of the tensor cannot be unique, which seems to rule out many spectral methods such as Jennrich's algorithm (whose analysis relies on having a unique decomposition).

We remark that for *discrete multi-reference alignment* (discrete MRA) where G is a *finite* group of rotations of order p, these issues do not arise. In fact, the samples y_i can be thought of as coming from a mixture of p spherical Gaussians where the centers are related (in that they are rotations of each other). By ignoring these interrelationships and learning the distribution as a mixture of spherical Gaussians via tensor decomposition, it is possible to obtain algorithms with provable guarantees [29]. In contrast, continuous MRA is a continuous mixture model where we crucially must exploit the relationship between the (infinitely-many) centers. The continuous nature of our problem poses a fundamental challenge for applying tensor methods. To overcome this, we will first randomly break the symmetry and then apply a spectral method that resembles a tailor-made variant of the tensor power method.

In this paper, we leverage this methodology to give a polynomial-time algorithm for *list recovery* for the continuous MRA problem and for its so-called *heterogeneous* generalization in which there are multiple true signals $\theta^1, \dots, \theta^K \in \mathbb{R}^p$ and each sample comes from a random one of them. (The homogeneous case $K = 1$ is known to admit a simple "frequency marching" solution [14]; see Section 2.4.1.) Here *list recovery* means that we output a list of polynomially-many candidate vectors such that every true signal is well correlated with at least one candidate. To achieve this, we need to delicately exploit symmetries in the orbit of each θ^k, but cope with the fact that the orbits of different components are unrelated. More broadly, our success gives us hope that our methodology for designing tensor spectral methods can be adapted to a wide variety of problems that have thus far resisted attack. As in work on overcomplete tensor decomposition [20, 21, 24], our analysis assumes that the signals θ^k are drawn at random (i.i.d. Gaussian). To the best of our knowledge, our algorithm provides the first polynomial-time solution to an orbit recovery problem over an infinite group, other than a few special cases that admit *ad hoc* closed-form solutions (see Section 2.4.1). In particular, we give the first polynomial-time solution to a *heterogeneous* orbit recovery problem over an infinite group.

We now motivate and describe our approach for the continuous MRA problem. Many existing methods for overcomplete tensor decomposition are based on the idea of finding a vector $v \in \mathbb{R}^p$ that maximizes the cubic form $\langle T, v^{\otimes 3} \rangle = \langle \sum_{i=1}^r a_i^{\otimes 3}, v \rangle$. If the a_i are random, it can be shown that approximately, the maximizers of $\langle T, v^{\otimes 3} \rangle$ are $a_1, \dots, a_r$ provided $r \lesssim p^{3/2}$ [20]. A popular heuristic for optimizing $\langle T, v^{\otimes 3} \rangle$ over unit vectors is the tensor power method, in which we iteratively update $v \in \mathbb{R}^p$ according to

$$v_i \leftarrow \sum_{jk} T_{ijk} v_j v_k. \tag{1}$$

Similarly to the matrix power method, the intuition here is that by "multiplying" the tensor by itself, we are repeatedly amplifying the signal without having the noise build up too much. There are rigorous guarantees for this non-convex method for random overcomplete tensor decomposition, but require a very warm start [6, 7].

Perhaps fortuitously, unlike the matrix case there are many different ways that one can "multiply" third-order tensors together to create other "power methods." A *tensor network* is a diagram that specifies a recipe for multiplying a collection of tensors together. This concept has been used in areas such as quantum physics [16]. Tensor network notation is illustrated in Figure 1 and will be central to our work. One of our key observations is that, although they were not explained this way, many existing tensor methods in the literature can be re-interpreted as spectral methods on matrices derived from tensor networks. In particular, the spectral method of [21] for random overcomplete tensor decomposition is based on the tensor network shown in Figure 1(c); this method is a starting point for our work. In Appendix B of the full version [26] we catalog related results for the *tensor PCA* problem and how they can also be described as coming from certain tensor networks.

The tensor network abstraction gives us freedom to explore more complicated tensor networks, which helps us cope with the symmetries of continuous MRA. Ultimately we will use the tensor network in Figure 2. We will show that with decent probability over a random tensor u, the top eigenvector of the associated matrix is close to a vector in the orbit of θ. To accomplish this, we will employ the trace moment method which, in our setting, gives us a way to spectrally bound a certain noise term by counting certain *valid* labelings of the edges of a much larger tensor network that is obtained by stringing together many copies of Figure 2. The constraints imposed on a valid labeling are dictated by the SO(2) group structure.

We remark that our tensor T is quite sparse in the Fourier domain. (This is in stark contrast to the situation in random overcomplete tensor decomposition or tensor PCA.) In particular, T is $p \times p \times p$ but only supported on the $\sim p^2$ entries T_{ijk} for which $i + j + k = 0$.

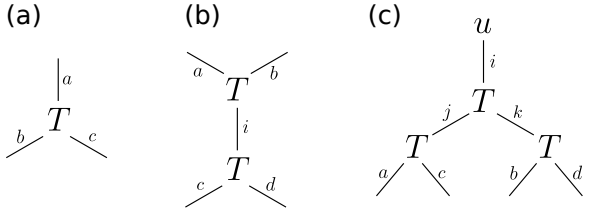

Figure 1: An introduction to tensor network notation. (a) A single copy of the third-order tensor T (with entries T_{abc}) has three *legs*, one for each mode. (b) Two copies of T connected by contracting (summing over) the index i. The result is the fourth-order tensor $B_{abcd} = \sum_i T_{abi} T_{cdi}$. (c) The spectral method in [21] uses the $(\{a,b\},\{c,d\})$-flattening of this tensor network (which is a $p^2 \times p^2$ matrix). We explain this in more detail in Section 3.3. Here u is a random vector.

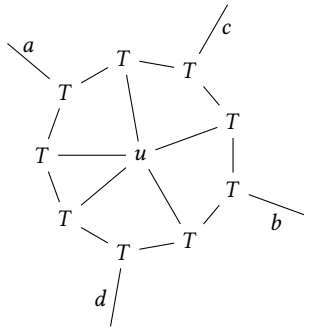

Figure 2: In this paper we will analyze a spectral method on the $p^2 \times p^2$ matrix given by the $(\{a,b\},\{c,d\})$-flattening of the tensor network shown here. Here u is a random order-5 tensor.

This comes from the fact that due to integrating over the group action, T is a projection of $\theta^{\otimes 3}$ onto a particular subspace (namely the span of the degree-three invariant polynomials; see [10]). The above sparsity pattern influences the combinatorics of the trace moment method. In particular, our valid labelings (discussed above) require that the three incoming legs to each copy of the tensor sum to zero. This is a rather different sort of combinatorics problem than typically arises in applications of the trace moment method to random matrix theory, and at a high-level, is why we need such a complex tensor network. In Appendix C of the full version [26], we discuss in more detail the considerations behind choosing the particular tensor network in Figure 2.

2 ORBIT RECOVERY PROBLEMS

2.1 Problem Statement

We now formally define *orbit recovery* problems including continuous MRA. These are a class of problems for which the method of moments gives rise to an *orbit tensor decomposition* problem.

Let G be a *compact group*. We do not formally define the notion of a compact group here, but some examples of interest include: (i) any finite group, such as the symmetric group S_L (permutations of $\{1,\dots,L\}$) and the cyclic group $\mathbb{Z}/L$, (ii) 2-dimensional rotations SO(2), and (iii) 3-dimensional rotations SO(3).

Let G *act linearly* on $\mathbb{R}^p$. A linear action means that each group element $g \in G$ has an associated matrix $\rho(g) \in \mathbb{R}^{p\times p}$ by which it acts on $\mathbb{R}^p$ (via matrix multiplication): for $\theta \in \mathbb{R}^p$ we write $g \cdot \theta = \rho(g)\theta$. The matrices must be consistent with the group structure, i.e. $\rho(gh) = \rho(g)\rho(h)$ and $\rho(e) = I$ where $e \in G$ is the identity.

Given a compact group G acting linearly on $\mathbb{R}^p$, we define the associated *orbit recovery* problem [10] as follows. (This has also been called the *group action channel* [2].) For $i = 1,\dots,n$ we observe

$$y_i = g_i \cdot \theta + \xi_i$$

where $\theta \in \mathbb{R}^p$ is the unknown signal, g_i is drawn from *Haar measure* (the "uniform distribution") on G, and $\xi_i \sim \mathcal{N}(0,\sigma^2 I)$. The random variables g_i, ξ_i are all independent. The goal is to estimate θ up to group action, i.e. to output an estimator close to the orbit $\{g \cdot \theta : g \in G\}$ of θ.

The following are some motivating examples of orbit recovery problems.

- **(Discrete) multi-reference alignment (MRA)** [1, 8, 11, 14, 18, 29]: This is the case where G is the cyclic group $\mathbb{Z}/p$ acting on $\mathbb{R}^p$ via cyclic permutation. Formally, for $g \in \mathbb{Z}/p$ (integers mod p), let $(g \cdot \theta)_i = \theta_{i-g \pmod p}$. This captures the problem where we see many noisy copies of the same discrete signal, each with a different offset. This has applications in signal processing [31, 39] and structural biology [19, 37]. We refer to the above problem as *discrete MRA* in contrast to *continuous MRA* which will be defined later.
- **Cryo-electron microscopy (cryo-EM)** [3, 10, 28, 36]: Cryo-EM is a popular biological imaging technique used to deduce the 3-dimensional structure of a large molecule such as a protein. This method was awarded the 2017 Nobel Prize in Chemistry. The method produces data in the form of many noisy 2-dimensional images of the 3-dimensional molecule, but in each image the molecule is rotated to an unknown orientation in 3-dimensional space. Here we think of $\theta \in \mathbb{R}^p$ as a representation of the molecule in some fixed basis (see [10] for a precise definition). The group is $G = \mathrm{SO}(3)$ acting by rotating the molecule. This is a generalization of the orbit recovery problem where we observe $y_i = \Pi(g_i \cdot \theta) + \xi_i$ where Π is a fixed linear operator, namely the mapping from a 3-dimensional molecule to a 2-dimensional image.

We will consider the *heterogeneous* extension of orbit recovery. This is motivated by cryo-EM in situations where there are multiple molecules (or multiple conformations of the same molecule) and each image contains an unknown one of them. Formally, there are K true signals $\theta^1,\dots,\theta^K \in \mathbb{R}^p$ and each sample takes the form

$$y_i = g_i \cdot \theta^{k_i} + \xi_i$$

where k_i is drawn at random from $[K] = \{1,\dots,K\}$. In general, one can consider an arbitrary distribution over $[K]$, but we will restrict ourselves to the case where k_i is drawn uniformly from $[K]$. In the heterogeneous problem, the goal is to estimate $\theta^1,\dots,\theta^K$ up to permutation and group action.

2.2 Continuous MRA

In this paper we will focus on the (heterogeneous) *continuous MRA* problem, as it is a simple example of an orbit recovery problem over an infinite group. Here we take the group to be $G = \mathrm{SO}(2)$, parametrized by angles $g \in [0, 2\pi)$. (Haar measure is simply the uniform distribution on angles.) Let p be even. The signal is $\theta \in \mathbb{R}^p$ with entries indexed by the "frequencies" $\pm j$ for $j \in [p/2] = \{1, 2, \dots, p/2\}$. We will denote this set of frequencies by $\pm[p/2] = \{-p/2, \dots, -1, 1, \dots, p/2\}$ (note that 0 is not included for convenience). The action of G on $\mathbb{R}^p$ is block-diagonal with 2×2 blocks: $g \in G$ acts on $[\theta_j\ \theta_{-j}]^\top$ (with $j > 0$) via the matrix

$$\begin{pmatrix} \cos(jg) & -\sin(jg) \\ \sin(jg) & \cos(jg) \end{pmatrix}.$$

It will sometimes be convenient to work in the Fourier basis: for $j > 0$,

$$\hat{\theta}_j = \frac{1}{\sqrt{2}}(\theta_j + \mathrm{i}\,\theta_{-j}) \quad \text{and} \quad \hat{\theta}_{-j} = \frac{1}{\sqrt{2}}(\theta_j - \mathrm{i}\,\theta_{-j}) \tag{2}$$

where i is the imaginary unit. If $\theta \sim \mathcal{N}(0, I/p)$, we have $\hat{\theta}_j \sim \mathcal{N}(0, 1/(2p)) + \mathrm{i}\,\mathcal{N}(0, 1/(2p))$ with $\hat{\theta}_{-j} = \overline{\hat{\theta}_j}$ (complex conjugate). In the Fourier basis, the action of G is diagonal, with g acting on $\hat{\theta}_j$ by the scalar $\exp(\mathrm{i}jg)$.

2.3 Method of Moments

One method for approaching orbit recovery problems is to attempt to learn the unknown group elements g_i. This is the well-studied *synchronization approach* [9, 11, 12, 15, 30, 35, 36].

An alternative approach uses the *method of moments*, which seeks to estimate θ directly from the moments of the samples without attempting to estimate the g_i. This was discovered first in the case of MRA [1, 8, 29] and later extended to all groups [2, 10]. This method is suited to the case where the noise σ on each sample is very large but we get many samples; in this regime we cannot hope to accurately estimate g_i but can still hope to recover θ.

We now describe the method of moments more formally. Consider the heterogeneous problem with signals $\theta^1, \dots, \theta^K \in \mathbb{R}^p$. In the method of moments we use the samples y_i to estimate the moments

$$T_1(\{\theta^k\}) = \mathop{\mathbb{E}}_{k,g}[g \cdot \theta^k] = \frac{1}{K}\sum_{k=1}^{K} \mathop{\mathbb{E}}_{g}[g \cdot \theta^k]$$

$$T_2(\{\theta^k\}) = \mathop{\mathbb{E}}_{k,g}[(g \cdot \theta^k)(g \cdot \theta^k)^\top] = \frac{1}{K}\sum_{k=1}^{K} \mathop{\mathbb{E}}_{g}[(g \cdot \theta^k)(g \cdot \theta^k)^\top]$$

$$\vdots$$

$$T_d(\{\theta^k\}) = \mathop{\mathbb{E}}_{k,g}[(g \cdot \theta^k)^{\otimes d}] = \frac{1}{K}\sum_{k=1}^{K} \mathop{\mathbb{E}}_{g}[(g \cdot \theta^k)^{\otimes d}].$$

Above, the expectation is over k drawn uniformly from $[K]$ and g drawn from Haar measure on G. It is possible to accurately estimate the moments $T_1, \dots, T_d$ given roughly $n \sim \sigma^{2d}$ samples (recall σ is the noise level) [2, 10]. Thus we are interested in an inversion procedure that recovers $\{\theta^k\}$ (up to permutation and orbit) given $T_1, \dots, T_d$, for d as small as possible. General algebraic techniques exist for testing how large d needs to be for this to be possible, but this does not necessarily give a polynomial-time algorithm to actually recover the signal from the moments [10]. For many natural problems such as MRA and cryo-EM, it is known that $d = 3$ is sufficient (and necessary) [1, 8, 10, 29].

It is known that the method of moments is statistically optimal in the limit $\sigma \to \infty$ (with the group, group action, and dimension p fixed) in the following sense [2, 8, 10]. On one hand, $n \sim \sigma^{2d}$ samples are sufficient to estimate the moments $T_1, \dots, T_d$. On the other hand, if two signals θ, θ' (or more generally, two collections of K heterogeneous signals) produce the same $T_1, \dots, T_{d-1}$ then at least $n \sim \sigma^{2d}$ samples are statistically required in order to distinguish between θ and θ'. In other words, if the method of moments requires moments up to d then *any* method requires at least σ^{2d} samples.

For the case of continuous MRA, it is easiest to work with the moments in the Fourier domain: $\hat{T}_d(\{\theta^k\}) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_g[(g \cdot \hat{\theta})^{\otimes d}]$ where the action of g on θ is diagonal: identifying g with an angle $g \in [0, 2\pi)$ we have $(g \cdot \hat{\theta})_j = \exp(\mathrm{i}jg)\hat{\theta}_j$ (where i is the imaginary unit). For $j_1, \dots, j_d \in \pm[p/2]$ we can compute

$$\hat{T}_d(\{\theta^k\})_{j_1,\dots,j_d} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}_{j_1+\cdots+j_d=0}\, \hat{\theta}^k_{j_1} \cdots \hat{\theta}^k_{j_d}. \tag{3}$$

2.4 Efficient Algorithms

We have seen above that the optimal statistical procedure is to compute moments T_i and to use these to solve for $\{\theta^k\}$ consistent with these moments. *A priori*, this is a polynomial system of equations which cannot be solved efficiently. In this section we survey known polynomial-time methods for recovering the signal(s) from the moments in special cases.

2.4.1 Frequency Marching. Both the discrete and continuous MRA problems admit a closed-form solution called *frequency marching* in the *homogeneous* case ($K = 1$). These methods are limited in the sense that they rely heavily on the particular structure of MRA and do not seem to extend to other groups or to the heterogeneous case (even for $K = 2$).

For discrete MRA, the frequency marching approach is described in [14]. An essentially-identical method works for continuous MRA, which we describe here.

Consider the homogeneous continuous MRA problem. The goal is to recover θ from $T_2(\theta)$ and $T_3(\theta)$ under the assumption that all Fourier coefficients of θ are nonzero. Recall the structure of moments (3). From T_2 we learn, for every $j \in [p/2]$, the value $\hat{\theta}_j\hat{\theta}_{-j} = \hat{\theta}_j\overline{\hat{\theta}_j} = |\hat{\theta}_j|$, i.e. we learn the magnitudes of the Fourier coefficients (the *power spectrum*). It suffices to recover the phases. From T_3 we learn the value $\hat{\theta}_{j_1}\hat{\theta}_{j_2}\hat{\theta}_{j_3}$ for every $j_1, j_2, j_3 \in \pm[p/2]$ such that $j_1 + j_2 + j_3 = 0$ (the *bispectrum*). Provided $\hat{\theta}_1 \neq 0$, each orbit has a unique representative such that the phase ϕ_1 of $\hat{\theta}_1$ is 0. Thus we take $\phi_1 = 0$. Now use $\hat{\theta}_{-1}\hat{\theta}_{-1}\hat{\theta}_2$ to learn ϕ_2, use $\hat{\theta}_{-1}\hat{\theta}_{-2}\hat{\theta}_3$ to learn ϕ_3, and so on until we have learned all the phases.

Another problem that admits a similar closed-form solution (in the homogeneous case only) is cryo-ET (cryo-electron tomography), a variant of cryo-EM without the projection step [10]. The cryo-EM

problem remains open (even in the homogeneous case): there are no known polynomial-time algorithms with provable guarantees.

2.4.2 Tensor Decomposition. Note that when G is a finite group, the third moment T_3 takes the form

$$T_3(\{\theta^k\}) = \frac{1}{K|G|}\sum_{k=1}^{K}\sum_{g\in G}(g\cdot\theta^k)^{\otimes 3}$$

which is a low-rank tensor (of rank $K|G|$) and is thus amenable to standard tensor decomposition techniques. For homogeneous discrete MRA, T_3 is undercomplete and can be decomposed using Jennrich's algorithm [29], thus recovering (all shifts of) the signal. For heterogeneous discrete MRA, T_3 is overcomplete (rank exceeds dimension) but Jennrich's algorithm can still be used if we are given a higher order moment tensor. For instance, if $K \le p/2$ then Jennrich's algorithm can be used to decompose T_5 [29]. However, estimating T_5 requires suboptimal sample complexity $n \sim \sigma^{10}$. If we assume θ^k are random (i.i.d. Gaussian) and $K \lesssim \sqrt{p}$, we can avoid this by using overcomplete methods to decompose T_3 [38]. This result is an adaptation of methods for random overcomplete tensor decomposition using the sum-of-squares hierarchy [24]. It is conjectured that $K \lesssim \sqrt{p}$ is optimal for efficient methods that use T_3 [17, 38].

Remark 2.1. One property of (the analysis of) Jennrich's algorithm is that it is only guaranteed to work in cases where the tensor has a unique decomposition. This is a serious barrier to using Jennrich's algorithm for problems over infinite groups. If G is infinite, we might still hope that $T_3 = \mathbb{E}_g[(g\cdot\theta)^{\otimes 3}]$ (or a higher-order moment) has a low-rank decomposition and that this decomposition tells us something about θ. However, even if this were true, we could not use (the existing analysis of) Jennrich's algorithm to find such a decomposition because the decomposition would not be unique: if $T_3 = \sum_{i=1}^{r} a_i^{\otimes 3}$ then we also have $T_3 = \sum_{i=1}^{r}(g\cdot a_i)^{\otimes 3}$ for any $g \in G$. More generally, it seems that any spectral method (which attempts to recover the signal as an eigenvector of some matrix) cannot succeed unless it first breaks the symmetry; otherwise there are infinitely-many solutions but a matrix only has finitely-many eigenvectors. Our method will randomly break the symmetry and then use a spectral method.

3 RESULTS AND TECHNIQUES

3.1 Notation

We say an event occurs with *high probability* if it has probability $1 - o(1)$ (as $p \to \infty$). We say an event occurs with *overwhelming probability* if it occurs with probability $1 - 1/\delta(p)$ where $\delta(p)$ grows faster than any polynomial in p (i.e. for any $k \in \mathbb{N}$, $\delta(p) \ge \omega(p^k)$). The notation $\tilde{O}(\cdot)$ hides factors of $\log(p)$.

We write $[p] = \{1, 2, \ldots, p\}$ and define $\pm[p/2]$ as in Section 2.2. The $p \times p$ identity matrix is denoted I_p or simply I. We use $\|\cdot\|$ to denote the spectral (operator) norm of a matrix. We use $\|\cdot\|_F$ and $\|\cdot\|_\infty$ to denote the Frobenius and L^∞ norms (respectively) of a matrix or tensor. For a tensor T, we use e.g. $\|T\|_{\{a,b\},\{c,d\}}$ to denote the spectral norm of the $(\{a,b\},\{c,d\})$-flattening of T. The $(\{a,b\},\{c,d\})$-flattening of a 4-tensor $T \in (\mathbb{R}^p)^{\otimes 4}$ is the $p^2 \times p^2$ matrix $M_{ab,cd} = T_{abcd}$.

3.2 Main Result

We now state our main result on list recovery for heterogeneous continuous MRA.

Theorem 3.1. *Let $\theta^1, \ldots, \theta^K \in \mathbb{R}^p$ be drawn independently from $\mathcal{N}(0, I_p/p)$. Suppose we are given the tensor $\mathcal{T} = T + E \in (\mathbb{R}^p)^{\otimes 3}$ where $\|E\|_\infty \le K^{-8}p^{-4}/\mathrm{polylog}(p)$ and*

$$T = \sum_{k=1}^{K}\mathop{\mathbb{E}}_{g}\left[(g\cdot\theta^k)^{\otimes 3}\right]$$

with g drawn from Haar (uniform) measure on $\mathrm{SO}(2)$. For any $\varepsilon > 0$, there is an algorithm that runs in time $p^{O(1)/\varepsilon^4}$ and outputs a list of unit vectors $\tau_1, \ldots, \tau_L \in \mathbb{R}^p$ with $L = p^{O(1)/\varepsilon^4}$ that has the following guarantee. Suppose $K \le p^\delta$ for a universal constant $\delta > 0$. With high probability over both $\theta^1, \ldots, \theta^K$ and the algorithm's randomness, for every $k \in [K]$ there exists $i \in [L]$ such that $\langle\tau_i, \theta^k\rangle^2 \ge 1 - \varepsilon - o(1)$.

For any constant ε, our algorithm runs in polynomial time. To the best of our knowledge, this is the first polynomial-time algorithm for a heterogeneous orbit recovery problem over an infinite group. (A few homogeneous problems have frequency marching solutions; see Section 2.4.1.) Moreover by Proposition 7.6 of [10], to compute $\mathcal{T}$ satisfying the above condition on $\|E\|_\infty$, it is sufficient to take $n = \tilde{O}(\sigma^6 K^{18} p^8)$ samples. This exhibits *statistically-optimal* dependence of σ^6 on the noise level. We do not attempt to optimize the constant δ, but we expect that $K \sim \sqrt{p}$ is optimal; see Appendix C of the full version [26].

Our algorithm produces a list of candidate solutions but we do not analyze how to hypothesis test to select the correct solution(s) from the list. We leave this as an open question for future work. Heuristically, in the homogeneous case, one can evaluate a candidate solution τ by comparing $T_2(\tau)$ and $T_3(\tau)$ to our estimates for the true moments $T_2(\theta), T_3(\theta)$. In the heterogeneous case, we want to find vectors $\tau_1, \ldots, \tau_K$ from our list such that $T_d(\{\tau_k\}) = \frac{1}{K}\sum_{k=1}^{K} T_d(\tau_k)$ is close to the true moments $T_d(\{\theta^k\})$ for $d = 2, 3$. This is a linear system subject to a K-sparse constraint, which could perhaps be solved using standard methods such as ℓ_1-minimization.

3.3 Summary of Techniques

Our approach will draw inspiration from prior work on random overcomplete third-order tensor decomposition. This is the problem of recovering $\{a_1, \ldots, a_r\}$ from

$$T = \sum_{i=1}^{r} a_i^{\otimes 3} \tag{4}$$

where the $a_i \in \mathbb{R}^p$ are drawn independently from $\mathcal{N}(0, I/p)$. The state-of-the-art theoretical results for this problem are a close-to-linear-time spectral method that succeeds when $r \lesssim p^{4/3}$ [21] and a polynomial-time sum-of-squares method that succeeds when $r \lesssim p^{3/2}$ [24]. (It seems likely that no efficient algorithm can succeed when r exceeds $p^{3/2}$.)

As a starting point for our techniques, we consider the spectral method of [21] for random overcomplete tensor decomposition. The key step of the algorithm is to construct (from T) the $p^2 \times p^2$

matrix

$$M = \sum_{i,j\in[r]} \langle u, \tilde{T}(a_i \otimes a_j)\rangle \cdot (a_i \otimes a_j)(a_i \otimes a_j)^\top \tag{5}$$

where $u \in \mathbb{R}^p$ is drawn randomly from $\mathcal{N}(0, I)$ and $\tilde{T}$ is obtained by flattening the input tensor to a $p \times p^2$ matrix: $\tilde{T} = \sum_{i\in[r]} a_i(a_i \otimes a_i)^\top$. The idea is that with some decent probability (inverse polynomial), the random vector u will align reasonably well with some a_i, and this causes the top eigenvector of M (after applying a certain "preconditioner") to be close to $a_i \otimes a_i$.

We can re-interpret the matrix M in the graphical language of tensor networks (see e.g. [16]), which we now describe. An order-d tensor $T \in (\mathbb{R}^p)^{\otimes d}$ is represented graphically as having d *legs*; the case $d = 3$ is shown in Figure 1(a). The legs are labeled with the three indices a, b, c that index into T. When two tensor legs are connected by a *wire*, this indicates contraction of the corresponding indices. For instance, the tensor network in Figure 1(b) represents the tensor $B \in (\mathbb{R}^p)^{\otimes 4}$ given by $B_{abcd} = \sum_{i\in[p]} T_{abi}T_{cdi}$. The matrix M from (5) is the $(\{a, b\}, \{c, d\})$-flattening of the tensor $C \in (\mathbb{R}^p)^{\otimes 4}$ that is represented by Figure 1(c). Specifically,

$$M_{ab,cd} = C_{abcd} = \sum_{i,j,k\in[p]} T_{acj}T_{bdk}T_{ijk}u_i. \tag{6}$$

One can check that (6) is equivalent to (5) when T is given by (4).

Now that we have expressed the matrix M from [21] as a tensor network, this opens the door to exploring a whole class of new spectral methods obtained by building various tensor networks out of the input tensor T. For instance, for the continuous MRA problem we will see that the tensor network in Figure 1(c) does not work but that a larger one, shown in Figure 2, does. In Appendix C of the full version [26] we explain in detail some of the considerations involved in choosing this particular tensor network.

We now describe our algorithm in more detail. Similarly to [21], our algorithm takes in a random guess u in order to break symmetry. Instead of a vector, u is now an order-5 tensor (with i.i.d. $\mathcal{N}(0, 1)$ entries). (In Appendix C of the full version [26] we explain the reason for this.) The hope is that u has better-than-random correlation with $\theta^{\otimes 5}$ for some θ in the orbit of one of the true signals $\theta^1, \dots, \theta^K$; if this occurs then we will recover a vector close to θ. Our algorithm takes u and the input tensor $\mathcal{T}$, and constructs a $p^2 \times p^2$ matrix $\tilde{M}(\mathcal{T}, u)$ according to the tensor network in Figure 2. We would like it to be the case that if we correctly guess $u = \theta^{\otimes 5}$ then $\tilde{M}(\mathcal{T}, \theta^{\otimes 5}) \approx (\theta^{\otimes 2})(\theta^{\otimes 2})^\top$, allowing us to recover θ. Due to the combinatorics of the SO(2) structure, this is not the case for $\tilde{M}$; however, luckily it is true after applying a particular simple correction to $\tilde{M}$, resulting in a matrix $M(\mathcal{T}, u)$. (This correction operates entrywise in the Fourier basis.) To extract a candidate solution from M, we symmetrize it and compute its top eigenvector $v \in \mathbb{R}^{p^2}$ (which we hope is close to $\theta^{\otimes 2}$). We then re-shape v into a $p \times p$ matrix, symmetrize it, and take the top eigenvector again in order to produce a candidate solution. We then repeat the entire process L times with fresh randomness u on each trial, in order to obtain a list of L candidate vectors.

Roughly speaking, a key step in our analysis is to show a high-probability upper bound on the spectral norm of our matrix $M = M(\mathcal{T}, u)$. To do this we use the trace moment method, a general-purpose tool from random matrix theory which relies on computing

$$\mathbb{E}[\mathrm{Tr}((MM^\top)^q)]. \tag{7}$$

In general, this computation can be quite difficult for complicated random matrices. However, even though M is quite complicated, the fact that it is represented by a tensor network helps us here. As shown in Figure 3, a tensor network for the quantity (7) can be obtained by connecting $2q$ copies of M in a circle. Since M is itself a tensor network, we need to connect $2q$ copies of *that* network in a circle, creating an expanded tensor network. As a result, the computation of (7) boils down to a combinatorics question involving counting certain labelings of this expanded tensor network.

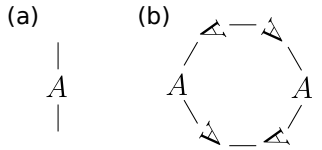

Figure 3: (a) A real-valued rectangular matrix A. (b) The tensor network representation of $\mathrm{Tr}[(AA^\top)^q]$ is formed by connecting $2q$ copies of A in a ring (here $q = 3$). Since A is asymmetric, the orientation of the "A" symbols matters.

4 PROOF FOR CONTINUOUS MRA

4.1 Preliminaries

4.1.1 Concentration. First we have some basic concentration results for random vectors.

Lemma 4.1. *If $\theta \sim \mathcal{N}(0, I_p/p)$ then*

$$\left|\|\theta\|^2 - 1\right| \le \tilde{O}(1/\sqrt{p})$$

with overwhelming probability.

Proof. This follows from Bernstein's inequality for subexponential random variables (see e.g. [33]). □

Lemma 4.2. *If $\theta \sim \mathcal{N}(0, I_p/p)$ then with overwhelming probability we have for all i,*

$$|\theta_i| \le \tilde{O}(1/\sqrt{p}).$$

Proof. This follows from standard Gaussian tail bounds. □

The following concentration bound is a consequence of *hypercontractivity* (see e.g. Theorem 1.10 of [34]).

Theorem 4.3. *Consider a degree-q polynomial $f(Y) = f(Y_1, \dots, Y_n)$ of independent Gaussian random variables $Y_1, \dots, Y_n$. Let σ^2 be the variance of $f(Y)$. There exists an absolute constant $R > 0$ such that*

$$\Pr\left[|f(Y) - \mathbb{E}[f(Y)]| \ge t\right] \le e^2 \cdot e^{-\left(\frac{t^2}{R\sigma^2}\right)^{1/q}}.$$

4.1.2 Fourier Basis. We will largely work in the Fourier domain. Let Δ be the unitary matrix that converts from the Fourier representation to the standard representation of a vector $v \in \mathbb{R}^p$, i.e. $\theta = \Delta\hat{\theta}$; see (2). We define the Fourier transform $\hat{T}$ of a tensor T as depicted in Figure 4(b). One can check that $\Delta^\top\Delta$ is the permutation matrix that swaps indices i and $-i$, i.e. $(\Delta^\top\Delta)_{ij} = \mathbb{1}_{i=-j}$. Thus, as shown in Figure 4(c), when two copies of Δ combine in a tensor network, we denote this by a dotted line which is understood to mean contraction with indices i and $-i$ paired.

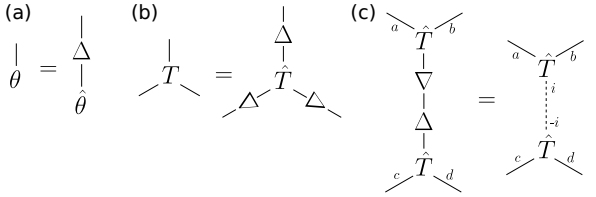

Figure 4: (a) The matrix Δ converts a vector's Fourier representation $\hat{\theta}$ to its standard representation θ. (b) We can convert from $\hat{T}$ to T by attaching three copies of Δ. Note that Δ is asymmetric and so the orientation of the Δ symbols is important. (c) When two Δ's connect as shown, this has the effect of a contraction in which indices i and $-i$ are paired. We abbreviate this as a dotted line with one end labeled i and the other end labeled $-i$. The tensor C shown here is $C_{abcd} = \sum_i \hat{T}_{abi}\hat{T}_{cd(-i)}$.

4.2 Main Technical Theorem

We now begin the proof of our main result (Theorem 3.1).

Our algorithm will build its list of candidate solutions by repeating a certain spectral method L times, with fresh randomness u each time. The following main technical theorem shows that each of these trials has a decent probability of success.

Theorem 4.4 (main technical theorem). *Let $\{\theta^k\}$, δ and $\mathcal{T}$ be as in Theorem 3.1. Let $K \le p^\delta$. Let $u \in (\mathbb{R}^p)^{\otimes 5}$ be drawn from $\mathcal{N}(0, I_{p^5})$. There is a matrix $M(\mathcal{T}, u) \in \mathbb{R}^{p^2\times p^2}$ (computable in time* poly(p) *from $\mathcal{T}$ and u) with the following guarantee. Let $v \in \mathbb{R}^{p^2}$ be the leading eigenvector of $\frac{1}{2}[M(\mathcal{T}, u) + M(\mathcal{T}, u)^\top]$. Re-shape[1] v to a $p\times p$ matrix V and let $\tau \in \mathbb{R}^p$ be the (unit-norm) leading eigenvector[2] of $\frac{1}{2}(V + V^\top)$. There is a deterministic predicate $P(\{\theta^k\})$ (defined in Section 4.7) depending only on $\{\theta^k\}$, that is satisfied with high probability (over $\{\theta^k\}$). For fixed $\{\theta^k\}$ satisfying $P(\{\theta^k\})$, for any $k \in [K]$ and any $\varepsilon > 0$, we have $\langle\tau, \theta^k\rangle^2 \ge 1 - \varepsilon - o(1)$ with probability $p^{-O(1)/\varepsilon^4}$ over the randomness of u.*

We first see how our main technical theorem (Theorem 4.4) implies our main theorem (Theorem 3.1).

Proof of Theorem 3.1. To produce the list $\tau_1, \dots, \tau_L$, the algorithm draws independent samples $u_1, \dots, u_L \sim \mathcal{N}(0, I_{p^5})$. For $i \in [L]$, extract τ_i from $M(\mathcal{T}, u_i)$ as in Theorem 4.4. For fixed k, let $\gamma = p^{-O(1)/\varepsilon^4}$ denote the success probability of a single trial. For fixed k, the probability of success after L trials is at least $1-(1-\gamma)^L \ge 1-\exp(-\gamma L)$. Taking a union bound over $[K]$, the overall probability of failure is at most $K\exp(-\gamma L) \le p^\delta \exp(-\gamma L)$. To make this $o(1)$, it is sufficient to take $L = \log^2(p)/\gamma = p^{O(1)/\varepsilon^4}$. □

[1]We will see that $M(\mathcal{T}, u)$ is a flattening of a 4-tensor, with entries $M(\mathcal{T}, u)_{ab,cd}$. Thus v has entries v_{ab} and can be naturally thought of as a $p \times p$ matrix.

[2]Here the *leading eigenvector* is defined to be the one whose eigenvalue is largest in absolute value.

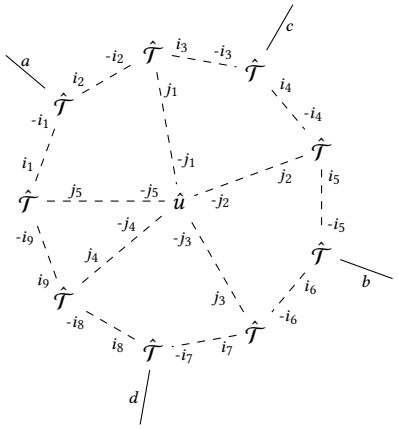

Figure 5: The $p^2 \times p^2$ matrix $\hat{M}(\mathcal{T}, u)$ is obtained by applying $\hat{\mathcal{S}}$ to the $(\{a, b\}, \{c, d\})$-flattening of the tensor shown here. The dotted lines and the Fourier transforms $\hat{\mathcal{T}}$ and $\hat{u}$ are defined as in Figure 4.

We now begin the proof of the main technical theorem (Theorem 4.4). The $p^2\times p^2$ matrix $M(\mathcal{T}, u)$ is the $(\{a, b\}, \{c, d\})$-flattening of the tensor depicted in Figure 2, but with an additional post-processing operator $\mathcal{S}$ applied to it. This operator is easiest to describe in the Fourier domain: let $\hat{\mathcal{S}}$ be the operator that acts entrywise on a 4-tensor by multiplying the $abcd$ entry by a non-negative real number S_{abcd} to be specified later. We will have $S_{abcd} = S_{(-a)(-b)(-c)(-d)}$ and so $\mathcal{S}$ takes real 4-tensors to real 4-tensors. We define

$$M(\mathcal{T}, u) = (\Delta \otimes \Delta)[\hat{M}(\mathcal{T}, u)](\Delta \otimes \Delta)^\top$$

where Δ is as in Section 4.1.2, and where $\hat{M}(\mathcal{T}, u)$ is obtained by applying $\hat{\mathcal{S}}$ to the $(\{a, b\}, \{c, d\})$-flattening of the tensor depicted in Figure 5.

Explicitly, we have

$$\begin{aligned}\hat{M}(\mathcal{T}, u)_{ab,cd} = S_{abcd} \sum_{i_1,\dots,i_9} \sum_{j_1,\dots,j_5} &\hat{u}_{-j_1,-j_2,-j_3,-j_4,-j_5}\hat{\mathcal{T}}_{-i_1,a,i_2} \\ &\times \hat{\mathcal{T}}_{-i_2,j_1,i_3}\hat{\mathcal{T}}_{-i_3,c,i_4}\hat{\mathcal{T}}_{-i_4,j_2,i_5}\hat{\mathcal{T}}_{-i_5,b,i_6} \\ &\times \hat{\mathcal{T}}_{-i_6,j_3,i_7}\hat{\mathcal{T}}_{-i_7,d,i_8}\hat{\mathcal{T}}_{-i_8,j_4,i_9}\hat{\mathcal{T}}_{-i_9,j_5,i_1}.\end{aligned}$$

Recall $\mathcal{T} = T + E$ where

$$T = \sum_{k=1}^K \mathop{\mathbb{E}}_g \left[(g \cdot \theta^k)^{\otimes 3}\right].$$

Let θ^k be the signal we are hoping to recover. Let

$$T^k = \mathop{\mathbb{E}}_g \left[(g \cdot \theta^k)^{\otimes 3}\right].$$

Recall $u \sim \mathcal{N}(0, I_{p^5}) \in (\mathbb{R}^p)^{\otimes 5}$. Write $u = \alpha\,(\theta^k)^{\otimes 5} + \tilde{u}$ with $\tilde{u} \perp (\theta^k)^{\otimes 5}$. We will break down the matrix $M(\mathcal{T}, u)$ into the following terms:

$$\begin{aligned} M(\mathcal{T},u) &= M(T,u) + [M(\mathcal{T},u) - M(T,u)] \\ &= \alpha M(T,(\theta^k)^{\otimes 5}) + M(T,\tilde{u}) + [M(\mathcal{T},u) - M(T,u)] \\ &= \alpha M(T^k,(\theta^k)^{\otimes 5}) + \alpha[M(T,(\theta^k)^{\otimes 5}) - M(T^k,(\theta^k)^{\otimes 5})] \\ &\quad + M(T,\tilde{u}) + [M(\mathcal{T},u) - M(T,u)]. \end{aligned}$$

Here we have used the fact that $M(\mathcal{T}, u)$ is linear in u. We now have four terms to bound separately.

4.3 Signal Term

Here we consider the signal term $M(T^k, (\theta^k)^{\otimes 5})$. Let $\Theta^k = (\theta^k \otimes \theta^k)(\theta^k \otimes \theta^k)^\top$, the matrix we would like to recover. Intuitively, we will show that if we were to correctly guess $u = (\theta^k)^{\otimes 5}$, then the resulting matrix matrix would be close to Θ^k. In order for this to be true, we will need to choose the parameters S_{abcd} appropriately.

Proposition 4.5. *For any $k \in [K]$, with overwhelming probability over θ^k,*

$$\|M(T^k,(\theta^k)^{\otimes 5}) - \Theta^k\| \le o(1).$$

This section is devoted to proving Proposition 4.5. Recall

$$\hat{T}_{i_1 i_2 i_3} = \mathbb{1}_{i_1+i_2+i_3=0} \sum_{k=1}^{K} \hat{\theta}^k_{i_1}\hat{\theta}^k_{i_2}\hat{\theta}^k_{i_3}$$

and so

$$\hat{T}^k_{i_1 i_2 i_3} = \mathbb{1}_{i_1+i_2+i_3=0}\,\hat{\theta}^k_{i_1}\hat{\theta}^k_{i_2}\hat{\theta}^k_{i_3}.$$

Without loss of generality, take $k = 1$. We have

$$\hat{M}(T^1,(\theta^1)^{\otimes 5})_{ab,cd} = S_{abcd}\,s_{abcd}\,\hat{\theta}^1_a\hat{\theta}^1_b\hat{\theta}^1_c\hat{\theta}^1_d \tag{8}$$

where

$$\begin{aligned} s_{abcd} = \sum_{i_1,\dots,i_9}\sum_{j_1,\dots,j_5} &\left(\mathbb{1}_{-i_1+a+i_2=0}\cdots\mathbb{1}_{-i_9+j_5+i_1=0}\right) \\ &\times\left(|\hat{\theta}^1_{i_1}|^2\cdots|\hat{\theta}^1_{i_9}|^2|\hat{\theta}^1_{j_1}|^2\cdots|\hat{\theta}^1_{j_5}|^2\right). \end{aligned}$$

Here the indicator functions enforce that for each copy of $\hat{\mathcal{T}}$ in Figure 5, the three incident labels sum to zero. Define

$$S_{abcd} \triangleq \begin{cases} 0 & \text{if } a = -b \text{ or } c = -d \\ 1/\mathbb{E}[s_{abcd}] & \text{otherwise.} \end{cases}$$

The reason for zeroing out some S_{abcd}'s will not be apparent until later (Section 4.4); this is crucially used in the proof of Lemma 4.19 for bounding the noise term $M(T, \tilde{u})$. The reason for $1/\mathbb{E}[s_{abcd}]$ should be clear from (8).

We will show that s_{abcd} concentrates near its expectation. We start with a basic computation of the moments of $\hat{\theta}^1$ (which of course holds for any $\hat{\theta}^k$).

Lemma 4.6. *$\mathbb{E}|\hat{\theta}^1_i|^{2k} = k!\,p^{-k}$. If $k_1 \neq k_2$ then $\mathbb{E}[(\hat{\theta}^1_i)^{k_1}(\hat{\theta}^1_{-i})^{k_2}] = 0$. If $i \neq \pm j$ then $\hat{\theta}^1_i$ and $\hat{\theta}^1_j$ are independent.*

Proof. The third statement is immediate from (2), since $\theta^1 \sim \mathcal{N}(0, I/p)$. The second statement is immediate from the fact that the complex phase of $\hat{\theta}^1_i$ is a uniformly random angle, and $\hat{\theta}^1_{-i} = \overline{\hat{\theta}^1_i}$. For the first statement, $|\hat{\theta}^1_i|^2 \sim \frac{1}{2p}\chi^2_2$, so use the known formula for chi-squared moments: $\mathbb{E}[(\chi^2_2)^k] = 2^k k!$. □

We next show that for every a, b, c, d we have $\mathbb{E}[s_{abcd}] = \Theta(p^{-9})$, specifically:

Lemma 4.7. *There exist universal positive constants c_1 and c_2 such that for every $a, b, c, d \in \pm[p/2]$,*

$$c_1\,p^{-9} \le \mathbb{E}[s_{abcd}] \le c_2\,p^{-9}.$$

Proof. Fix a, b, c, d. There is a (nonzero) term of s_{abcd} for each choice of indices $i_1, \dots, i_9, j_1, \dots, j_5 \in \pm[p/2]$ such that for each copy of $\hat{\mathcal{T}}$ in Figure 5, the three incident indices sum to zero. There are at most p^5 (nonzero) terms in s_{abcd} because once i_9, j_5, j_4, j_3, j_1 are chosen, the zero-sum constraints uniquely determine at most one possible value for the other indices. (We say "at most one" since only indices in the set $\pm[p/2]$ are valid.) Each term of s_{abcd} has expectation at most $14!\,p^{-14}$ (by Lemma 4.6), so $\mathbb{E}[s_{abcd}] \le 14!\,p^{-9}$. This proves the upper bound.

The idea of the lower bound is to argue that s_{abcd} has $\Omega(p^5)$ terms and each term has expectation at least p^{-14}. We defer the full proof to Appendix A of the full version [26]. □

Lemma 4.8. *There exists a universal positive constant c_3 such that for every $a, b, c, d \in \pm[p/2]$,*

$$\mathrm{Var}[s_{abcd}] \le c_3\,p^{-19}.$$

Proof. The variance of a sum can be broken down as

$$\mathrm{Var}(\sum_i x_i) = \sum_i \mathrm{Var}(x_i) + \sum_{i\neq j}\mathrm{Covar}(x_i, x_j).$$

Each of the $O(p^5)$ terms of s_{abcd} has variance $O(p^{-28})$. There are $O(p^{10})$ ways to choose two distinct terms of s_{abcd}. Only $O(p^9)$ of these ways gives two terms that are dependent, in which case their covariance is $O(p^{-28})$; otherwise they are independent and have covariance zero. This means $\mathrm{Var}[s_{abcd}] \le O(p^5 \cdot p^{-28} + p^9 \cdot p^{-28}) = O(p^{-19})$. □

By hypercontractivity (Theorem 4.3) we have with overwhelming probability, $|s_{abcd} - \mathbb{E}[s_{abcd}]| \le p^{-9.1}$. Thus, when $a \neq -b$ and $c \neq -d$, we have $|S_{abcd}\,s_{abcd} - 1| \le O(p^{-0.1})$ (recall that $S_{abcd}s_{abcd}$ appears in (8)). For the entries with $a = -b$ or $c = -d$ (there are $\le 2p^3$ such entries in $\hat{M}$), we will simply use the bound $|\hat{\theta}^1_i| \le \tilde{O}(1/\sqrt{p})$.

We can now complete the proof of Proposition 4.5. Using (8),

$$\begin{aligned} &\|M(T^1,(\theta^1)^{\otimes 5}) - (\theta^1\otimes\theta^1)(\theta^1\otimes\theta^1)^\top\| \\ &\quad= \|\hat{M}(T^1,(\theta^1)^{\otimes 5}) - (\hat{\theta}^1\otimes\hat{\theta}^1)(\hat{\theta}^1\otimes\hat{\theta}^1)^\top\| \\ &\quad\le \|\hat{M}(T^1,(\theta^1)^{\otimes 5}) - (\hat{\theta}^1\otimes\hat{\theta}^1)(\hat{\theta}^1\otimes\hat{\theta}^1)^\top\|_F \\ &\quad\le \sqrt{p^4\cdot\tilde{O}(p^{-0.1}\cdot p^{-2})^2 + 2p^3\cdot\tilde{O}(p^{-2})^2} \\ &\quad\le o(1). \end{aligned}$$

4.4 Noise Term

We now consider the noise term $M(T, \tilde{u})$, i.e. the term created by the "bad" component of u that is orthogonal to $(\theta^k)^{\otimes 5}$. This term is the crux of the proof, where we will crucially use the assumption $K \le p^\delta$.

Proposition 4.9. *There exists $\delta > 0$ such that if $K \le p^\delta$, we have the following. There is a determinstic predicate $P_1(\{\theta^k\})$ depending only on $\{\theta^k\}$ that is satisfied with high probability. For any fixed $\{\theta^k\}$ satisfying $P_1(\{\theta^k\})$, we have*

$$\|M(T, \tilde{u})\| \le O(\sqrt{\log p})$$

with high probability over the randomness of $\tilde{u}$.

Remark 4.10. Note that $\tilde{u}$ depends on which signal $k \in [K]$ we have chosen as the target, since $\tilde{u} \perp (\theta^k)^{\otimes 5}$. However, P_1 does not depend on k, and the conclusion of Proposition 4.9 holds for any fixed k.

This section is devoted to proving Proposition 4.9. We will use the following result on random contractions of tensors.

Theorem 4.11 ([24] Corollary 6.6). *Let $W \in \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^r$ be an order-3 tensor. Let $\tilde{u} \sim \mathcal{N}(0, \Sigma)$ with $r \times r$ covariance matrix satisfying $0 \le \Sigma \le I$. Then for any $t \ge 0$,*

$$\Pr_{\tilde{u}}\left[\|(I \otimes I \otimes \tilde{u}^\top)W\|_{\{1\},\{2\}} \ge t\left(\|W\|_{\{1\},\{2,3\}} \vee \|W\|_{\{1,3\},\{2\}}\right)\right] \le 4(p+q)\exp(-t^2/2).$$

In our setting, we have $M(T, \tilde{u}) = (I \otimes I \otimes \tilde{u}^\top)W$ where W is given by the tensor network in Figure 6(a) ($\hat{\mathcal{S}}$ is present but not shown). Explicitly, the Fourier transform (defined as in Figure 4) of W is

$$\hat{W}_{ab,cd,j_1j_2j_3j_4j_5} = S_{abcd} \sum_{i_1,\ldots,i_9} \hat{T}_{-i_1,a,i_2}\hat{T}_{-i_2,j_1,i_3}\hat{T}_{-i_3,c,i_4}\hat{T}_{-i_4,j_2,i_5} \times \hat{T}_{-i_5,b,i_6}\hat{T}_{-i_6,j_3,i_7}\hat{T}_{-i_7,d,i_8}\hat{T}_{-i_8,j_4,i_9}\hat{T}_{-i_9,j_5,i_1}.$$

By Theorem 4.11,

$$\Pr_{\tilde{u}}[\|M(T,\tilde{u})\| \ge t\sigma] \le 8p^2\exp(-t^2/2) \tag{9}$$

where

$$\sigma = \max\{\|W\|_{\{a,b\},\{c,d,j_1,j_2,j_3,j_4,j_5\}}, \|W\|_{\{a,b,j_1,j_2,j_3,j_4,j_5\},\{c,d\}}\}.$$

The two flattenings of W that appear in the definition of σ are equivalent due to symmetry, so it suffices to consider just the first one. The corresponding matrix is $\tilde{W} \in \mathbb{R}^{2\times 7}$ given by

$$\tilde{W}_{ab,cdj_1j_2j_3j_4j_5} = W_{ab,cd,j_1j_2j_3j_4j_5}.$$

We will bound $\|\tilde{W}\|$ using the trace moment method:

Theorem 4.12 (e.g. [32] Proposition 5.2). *For any real-valued random matrix Y, for any integer $q \ge 1$ and any $\varepsilon > 0$,*

$$\Pr\left[\|Y\| > \left(\frac{\mathbb{E}[\mathrm{Tr}((YY^\top)^q)]}{\varepsilon}\right)^{\frac{1}{2q}}\right] < \varepsilon.$$

As illustrated in Figures 3 and 6, we can represent $\mathrm{Tr}((\tilde{W}\tilde{W}^\top)^q)$ as a tensor network by connecting (in a ring) $2q$ copies of the tensor network for $\tilde{W}$. (We will take $q = \log p$.) Call this new tensor network $\mathcal{G}_q$ (see Figure 6). The computation of $\mathbb{E}[\mathrm{Tr}((\tilde{W}\tilde{W}^\top)^q)]$ is thus reduced to a combinatorial problem involving labelings of $\mathcal{G}_q$, which we next describe.

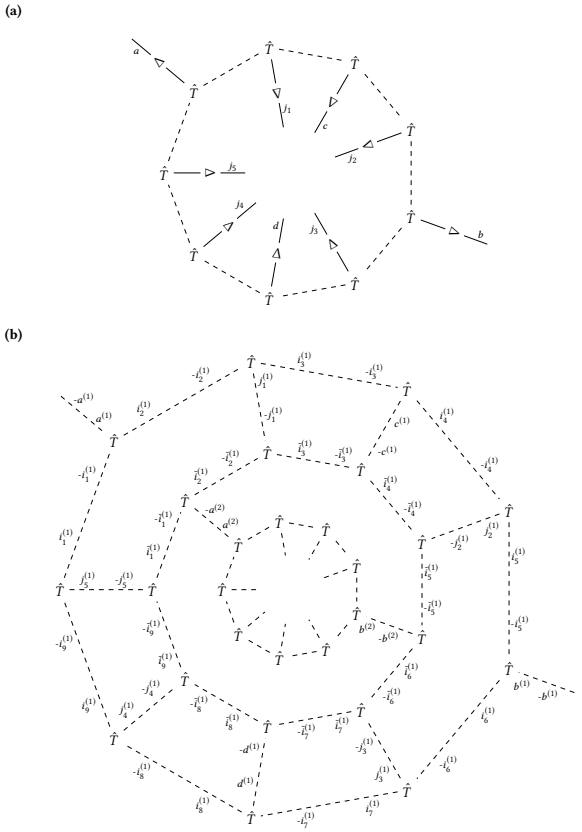

Figure 6: (a) The tensor network for W. (The operator $\hat{S}$ is not shown but can be thought of as living on the appropriate four edges.) The matrix $\tilde{W}$ is the $(\{a,b\},\{c,d,j_1,j_2,j_3,j_4,j_5\})$-flattening. (b) Here we see (part of) the tensor network $\mathcal{G}_q$ that computes $\mathrm{Tr}((\tilde{W}\tilde{W}^\top)^q)$. There are $2q$ copies of the tensor network from (a) (3 copies are shown here) connected in a ring as in Figure 3. The outermost copy connects back to the innermost copy, so that the entire network can be visualized as living on the surface of a torus. Again, $\hat{\mathcal{S}}$ is not shown (it will be unimportant since we will bound its contribution separately). Recall that connecting two "opposing" copies of Δ results in a dotted edge, as shown in Figure 4.

Definition 4.13. A *labeling* $\mathcal{L}$ of $\mathcal{G}_q$ is described by the following. For each edge e, label one end with a value $i_e \in \pm[p/2]$ and label the other end with $-i_e$. Call each copy of $\hat{T}$ in $\mathcal{G}_q$ a *vertex*, and label each vertex v with a value $k_v \in [K]$. Let $\mathcal{L}(v) = \mathbb{1}_{i_1+i_2+i_3=0}\,\hat{\theta}^{k_v}_{i_1}\hat{\theta}^{k_v}_{i_2}\hat{\theta}^{k_v}_{i_3}$ where i_1, i_2, i_3 are the three edge labels incident[3] to v.

Recall that $\hat{T}_{i_1i_2i_3} = \mathbb{1}_{i_1+i_2+i_3=0}\sum_{k=1}^K \hat{\theta}^k_{i_1}\hat{\theta}^k_{i_2}\hat{\theta}^k_{i_3}$. The vertex labels k_v correspond to the terms in this sum.

[3]Suppose an edge e is incident to a vertex v in a tensor network. There is a label $\pm i_e$ at each end of e. The label at the v-end of e is considered *incident* to v.

As shown in Figure 6, there are q *layers*, and layer ℓ (for $\ell = 1, 2, \dots, q$) has labels

$$a^{(\ell)}, b^{(\ell)}, c^{(\ell)}, d^{(\ell)}, i_1^{(\ell)}, \dots, i_9^{(\ell)}, \tilde{i}_1^{(\ell)}, \dots, \tilde{i}_9^{(\ell)}, j_1^{(\ell)}, \dots, j_5^{(\ell)}.$$

Layer numbers are defined modulo q, i.e. layer $q+1$ refers to layer 1.

We now have

$$\mathbb{E}[\mathrm{Tr}((\tilde{W}\tilde{W}^\top)^q)] = \mathbb{E}\sum_{\mathcal{L}} S_{\mathcal{L}} \prod_v \mathcal{L}(v)$$

where: $\mathcal{L}$ ranges over all labelings of $\mathcal{G}_q$; v ranges over all vertices of $\mathcal{G}_q$; the expectation is over the randomness of $\theta^1, \dots, \theta^K$; and the contributions from $\mathcal{S}$ are captured by

$$S_{\mathcal{L}} \triangleq \prod_{\ell=1}^{q} S_{a^{(\ell)}b^{(\ell)}c^{(\ell)}d^{(\ell)}} S_{(-a^{(\ell+1)})(-b^{(\ell+1)})(-c^{(\ell)})(-d^{(\ell)})}.$$

Definition 4.14. Define the *number of repeated labels* in a labeling to be

$$c(\mathcal{L}) = \sum_{i \in [p/2]} \max\{0, [\#\text{ edges labeled with } \pm i] - 1\}.$$

Definition 4.15. For any k, call the set of vertices $\{v : k_v = k\}$ a *region*. The *number of regions* in a labeling is

$$r(\mathcal{L}) = |\{k_v\}_v| = [\#\text{ distinct } k_v \text{ values}].$$

Definition 4.16. Call $\mathcal{L}$ a *valid labeling* if $S_{\mathcal{L}}\, \mathbb{E} \prod_v \mathcal{L}(v) \neq 0$.

We have the following straightforward characterization of valid labelings.

Lemma 4.17. *$\mathcal{L}$ is valid if and only if*

(i) *for every vertex v, the three edge labels i_1, i_2, i_3 incident to v satisfy $i_1 + i_2 + i_3 = 0$,*
(ii) *for every ℓ, $a^{(\ell)} \neq -b^{(\ell)}$ and $c^{(\ell)} \neq -d^{(\ell)}$, and*
(iii) *for every i, each region has as many incident[4] i labels as incident $-i$ labels.*

Proof. Condition (i) is due to the factor $\mathbb{1}_{i_1+i_2+i_3}$ in $\mathcal{L}(v)$. Condition (ii) is due to the definition of S_{abcd} (recall that we set certain S_{abcd} values to zero). Condition (iii) comes from aggregating all the $\hat{\theta}_i^v$ factors in $\prod_v \mathcal{L}(v)$ and applying Lemma 4.6. □

Lemma 4.18. *For any valid labeling $\mathcal{L}$ with $r(\mathcal{L}) > 1$, we have $c(\mathcal{L}) \geq r(\mathcal{L})/2$.*

Proof. Since $r(\mathcal{L}) > 1$, every region has at least two edges crossing its boundary[5], so there are at least $r(\mathcal{L})$ such boundary edges total. In a valid labeling, each of these boundary edges must have the same label as at least one other boundary edge. This results in at least $r(\mathcal{L})/2$ repeated labels. □

The following key lemma is proved in Appendix A of the full version [26].

[4]If R is a region and $e = (u, v)$ is an edge with $u \in R$ and $v \notin R$ then the label at the u-end of e is considered *incident* to R.

[5]It is immediate from Lemma 4.17(iii) that a region must have an even number of edges crossing its boundary. In fact, it is not hard to see that $\mathcal{G}_q$ has no cuts of size two and so at least four edges must cross.

Lemma 4.19. *The number of valid edge-labelings[6] with exactly $c(\mathcal{L})$ repeated labels is at most $3[2(27q)^2]^{c(\mathcal{L})} p^{1+9q-c(\mathcal{L})/25}$.*

The interpretation of this is as follows. The important factor is $p^{1+9q-c(\mathcal{L})/25}$; the rest is lower-order. Without requiring any repeated labels, the number of valid edge-labelings is $\sim p^{1+9q}$. Thus the lemma shows that when repeated labels are required, the number of valid edge-labelings decreases substantially.

Remark 4.20. To achieve heterogeneity $K \lesssim p^\delta$, one needs to show that the number of valid labelings with $r(\mathcal{L})$ regions is $\lesssim p^{1+9q-\delta r(\mathcal{L})}$. Together, Lemmas 4.18 and 4.19 show this for some constant $\delta > 0$, but we have not attempted to optimize δ. See Appendix C of the full version [26] for more on this.

Using the above lemmas, we are able to bound $\mathbb{E}[\mathrm{Tr}((\tilde{W}\tilde{W}^\top)^q)]$ as desired. We prove the following in Appendix A of the full version [26].

Lemma 4.21. *With $q = \log p$,*

$$\mathbb{E}[\mathrm{Tr}((\tilde{W}\tilde{W}^\top)^q)] \leq O(1)^q p^2$$

and so

$$\{\mathbb{E}[\mathrm{Tr}((\tilde{W}\tilde{W}^\top)^q)]\}^{1/2q} \leq O(p^{1/q}) \leq O(1).$$

By Theorem 4.12, with probability at least $1 - 1/p$ (over the randomness of $\theta^1, \dots, \theta^K$) we have $\|\tilde{W}\| \leq O(1)$; let this be the predicate $P_1(\{\theta^k\})$. Provided $P_1(\{\theta^k\})$ holds, we then have by (9) that

$$\|M(T, \tilde{u})\| \leq c\sqrt{\log p}$$

with probability at least $1 - 8p^{2-\Omega(c^2)}$ (over the randomness of $\tilde{u}$). Taking c to be a sufficiently large constant completes the proof.

4.5 Heterogeneous Signal Term

Here we consider the term $M(T, (\theta^k)^{\otimes 5}) - M(T^k, (\theta^k)^{\otimes 5})$. This term is relatively benign compared to the previous one and we bound it using a much simpler variant of the argument in the previous section. For convenience we use $K \leq p^\delta$ here but we expect that the previous term (not this one) would be the bottleneck if we wanted to optimize δ.

Proposition 4.22. *There exists $\delta > 0$ such that if $K \leq p^\delta$ then with high probability over $\{\theta^k\}$ we have for every $k \in [K]$,*

$$\|M(T, (\theta^k)^{\otimes 5}) - M(T^k, (\theta^k)^{\otimes 5})\| \leq o(1).$$

This section is devoted to proving Proposition 4.22. By Markov's inequality, it is sufficient to show

$$\mathbb{E}\|M(T, (\theta^1)^{\otimes 5}) - M(T^1, (\theta^1)^{\otimes 5})\|_F^2 \leq o(1/K)$$

so that we can take a union bound over all $k \in [K]$.

The value $\|M(T, (\theta^1)^{\otimes 5})\|_F^2$ is depicted by the tensor network in Figure 7. Similarly to the previous section, we consider labelings of Figure 7. As before, each edge gets a label i_e and each vertex gets a label k_v. (Each $\hat{\theta}^1$ is also considered a vertex.)

[6]By *edge-labelings* we mean choices for the edge labels i_e but not the vertex labels v_k. Here *valid* means that (i) and (ii) in Lemma 4.17 are satisfied.

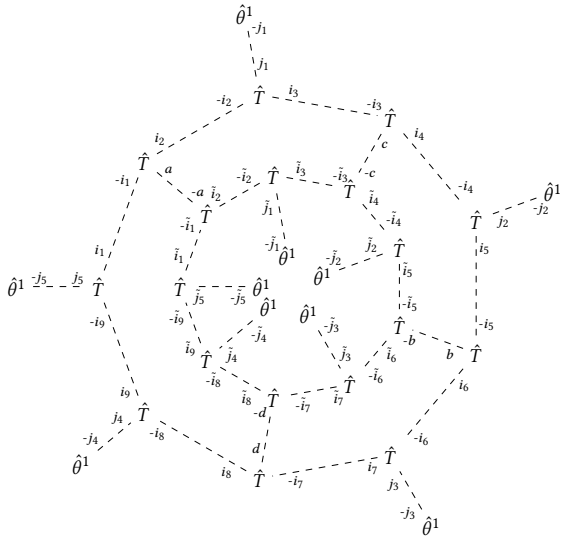

Figure 7: The tensor network for $\|M(T,(\theta^1)^{\otimes 5})\|_F^2$. The error term in (10) is obtained by only considering the terms corresponding to particular labelings (see Definition 4.23).

Definition 4.23. In addition to the requirements in Lemma 4.17, we define a *valid labeling* of Figure 7 to have two additional constraints: (i) each $\hat{\theta}^1$ has vertex label $k_v = 1$, and (ii) the inner and outer ring of $\hat{T}$'s each have a vertex for which $k_v \neq 1$.

By restricting to these valid labelings, we get an expression for the error term that we want. Formally,

$$\begin{aligned}&\|M(T,(\theta^1)^{\otimes 5}) - M(T^1,(\theta^1)^{\otimes 5})\|_F^2 \\ &\quad = \sum_{\text{valid } \mathcal{L}} S_{abcd} S_{(-a)(-b)(-c)(-d)} \prod_v \mathcal{L}(v) \end{aligned} \tag{10}$$

where for a $\hat{T}$ vertex, $\mathcal{L}(v)$ is defined as in the previous section, and for a $\hat{\theta}^1$ vertex, $\mathcal{L}(v) = \hat{\theta}^1_{-j}$ where $-j$ is the incident label. The proof of the following result can be found in the full version [26].

Lemma 4.24. *The number of valid labelings of Fig. 7 is $O(K^{18}p^{13})$.*

Let $S_{\max} = \max_{abcd} S_{abcd} \leq O(p^9)$. Since $0 \leq \mathbb{E}\prod_v \mathcal{L}(v) \leq O(p^{-32})$, we have

$$\begin{aligned}\mathbb{E}\|M(T,(\theta^1)^{\otimes 5}) - M(T^1,(\theta^1)^{\otimes 5})\|_F^2 &\leq S_{\max}^2 \sum_{\text{valid } \mathcal{L}} \mathbb{E}\prod_v \mathcal{L}(v) \\ &\leq O(p^{18}) \cdot O(K^{18}p^{13}) \cdot O(p^{-32}) \leq O(K^{18}/p)\end{aligned}$$

which is $o(1/K)$ provided $K \leq p^{1/20}$.

4.6 Error Term

Here we consider the term $M(\mathcal{T},u) - M(T,u)$. This is the error term due to the small error E that we allow in our input tensor. The following result follows from crude and easy bounds; see the full version [26] for the proof.

Proposition 4.25. *There is a deterministic predicate $P_2(\{\theta^k\})$ depending only on $\{\theta^k\}$ that occurs with high probability. If $P_2(\{\theta^k\})$ is satisfied then*

$$\|M(\mathcal{T},u) - M(T,u)\| \leq \tilde{O}(K^8 p^4 \|E\|_\infty)$$

with overwhelming probability over u.

4.7 Putting it all Together

Here we complete the proof of Theorem 4.4. Recall $u \sim \mathcal{N}(0,I) \in \mathbb{R}^{p^5}$, $\Theta^k = (\theta^k \otimes \theta^k)(\theta^k \otimes \theta^k)^\top$ and $u = \alpha(\theta^k)^{\otimes 5} + \tilde{u}$ with $\tilde{u} \perp (\theta^k)^{\otimes 5}$. As above, we write

$$\begin{aligned}M(\mathcal{T},u) = \alpha M(T^k,(\theta^k)^{\otimes 5}) + \alpha[M(T,(\theta^k)^{\otimes 5}) - M(T^k,(\theta^k)^{\otimes 5})] \\ + M(T,\tilde{u}) + [M(\mathcal{T},u) - M(T,u)].\end{aligned}$$

Let the predicate $P(\{\theta^k\})$ be the intersection of the following high-probability events:

- the conclusion of Proposition 4.5 holds for every $k \in [K]$,
- $P_1(\{\theta^k\})$ (from Proposition 4.9) holds,
- the conclusion of Proposition 4.22 holds,
- $P_2(\{\theta^k\})$ (from Proposition 4.25) holds,
- for every $k \in [K]$, $1 - \tilde{O}(1/\sqrt{p}) \leq \|\theta^k\| \leq 1 + \tilde{O}(1/\sqrt{p})$.

For fixed $\theta^1,\ldots,\theta^K$ satisfying $P(\{\theta^k\})$, we have for any k,

- $\|M(T^k,(\theta^k)^{\otimes 5}) - \Theta^k\| \leq o(1)$,
- $\|M(T,(\theta^k)^{\otimes 5}) - M(T^k,(\theta^k)^{\otimes 5})\| \leq o(1)$,
- $\|M(T,\tilde{u})\| \leq O(\sqrt{\log p})$ with high probability over $\tilde{u}$,
- $\|M(\mathcal{T},u) - M(T,u)\| \leq o(1)$ with overwhelming probability over u.

We have

$$\begin{aligned}\alpha &= \langle u,(\theta^k)^{\otimes 5}\rangle / \|(\theta^k)^{\otimes 5}\|^2 \\ &= \langle u,(\theta^k)^{\otimes 5}/\|(\theta^k)^{\otimes 5}\|\rangle / \|(\theta^k)^{\otimes 5}\| \triangleq \tilde{\alpha}/\|\theta^k\|^5\end{aligned}$$

where $\tilde{\alpha} \sim \mathcal{N}(0,1)$ independently from $\tilde{u}$.

We have the Gaussian lower tail bound

$$\Pr\{\tilde{\alpha} \geq t\} \geq \frac{1}{2\sqrt{2\pi}} t^{-1} \exp(-t^2/2)$$

and so

$$\Pr\left\{\tilde{\alpha} \geq C\sqrt{\log p}\right\} \geq \frac{1}{2\sqrt{2\pi}} \frac{1}{C\sqrt{\log p}} p^{-C^2/2}. \tag{11}$$

We can write $M_{\text{sym}}(\mathcal{T},u) \triangleq \frac{1}{2}[M(\mathcal{T},u) + M(\mathcal{T},u)^\top] = \alpha\,\Theta^k + B = \tilde{\alpha}\,\Theta^k/\|\theta^k\|^5 + B$ where

$$\|B\| \cdot \|\theta^k\| \triangleq \beta \leq o(1)\,\alpha + O(\sqrt{\log p}) \leq o(1)\,\tilde{\alpha} + O(\sqrt{\log p}). \tag{12}$$

Let $w = (\theta^k \otimes \theta^k)/\|\theta^k\|^2 \in (\mathbb{R}^p)^{\otimes 2}$ so that $ww^\top = \Theta^k/\|\theta^k\|^4$. Let $v \in (\mathbb{R}^p)^{\otimes 2}$ be the leading eigenvector of $M_{\text{sym}}(\mathcal{T},u)$ (with $\|v\| = 1$), which is also the leading eigenvector of $\tilde{M}_{\text{sym}}(\mathcal{T},u) = \|\theta^k\| M_{\text{sym}}(\mathcal{T},u) = \tilde{\alpha} ww^\top + \|\theta^k\| B$. We have

$$\tilde{\alpha}\langle v,w\rangle^2 + \beta \geq v^\top \tilde{M}_{\text{sym}}(\mathcal{T},u)\,v \geq w^\top \tilde{M}_{\text{sym}}(\mathcal{T},u)\,w \geq \tilde{\alpha} - \beta$$

and so $\langle v,w\rangle^2 \geq 1 - \frac{2\beta}{\tilde{\alpha}}$. Re-shape v into a $p \times p$ matrix $\tilde{V}$ and let $V = \frac{1}{2}(\tilde{V} + \tilde{V}^\top)$. Let τ be the eigenvector of V corresponding to the eigenvalue of largest absolute value. Let $y = \theta^k/\|\theta^k\|$. Write

$$V = \langle V, yy^\top\rangle yy^\top + B'$$

where $\|B'\|^2 \le \|B'\|_F^2 = \|V\|_F - \langle V, yy^\top\rangle^2 \le 1 - \langle V, yy^\top\rangle^2$. We have

$$|\langle V, yy^\top\rangle| = |y^\top V y| \le |\tau^\top V \tau| \le |\langle V, yy^\top\rangle| \cdot \langle \tau, y\rangle^2 + \|B'\|$$

and so

$$\langle \tau, y\rangle^2 \ge 1 - \frac{\|B'\|}{|\langle V, yy^\top\rangle|} \ge 1 - \frac{\sqrt{1 - \langle V, yy^\top\rangle^2}}{|\langle V, yy^\top\rangle|}.$$

Note that

$$\langle V, yy^\top\rangle^2 = \langle \tilde{V}, yy^\top\rangle^2 = \langle v, w\rangle^2 \ge 1 - \frac{2\beta}{\tilde{\alpha}}$$

and so, provided $2\beta/\tilde{\alpha} \le 1/2$,

$$\langle \tau, y\rangle^2 \ge 1 - \frac{\sqrt{2\beta/\tilde{\alpha}}}{\sqrt{1/2}} = 1 - 2\sqrt{\frac{\beta}{\tilde{\alpha}}}.$$

Recall from (12) that if $\tilde{\alpha} \ge C\sqrt{\log p}$ then $\beta/\tilde{\alpha} \le o(1) + O(1)/C$. Thus, to have $\langle \tau, y\rangle^2 \ge 1 - \varepsilon - o(1)$, it is sufficient to take $C = O(1)/\varepsilon^2$. Using (11), the success probability is $\ge p^{-O(1)/\varepsilon^4}$.

ACKNOWLEDGMENTS

This work was inspired by ideas of Amelia Perry, who hoped to use tensor networks to find improved sum-of-squares algorithms for the planted clique problem (before this was shown to be impossible [13]). Tragically, Amelia passed away in January, 2018.

We thank the anonymous referees for their helpful comments. A. Moitra was supported in part by NSF CAREER Award CCF-1453261, NSF Large CCF-1565235, a David and Lucile Packard Fellowship, an Alfred P. Sloan Fellowship and an ONR Young Investigator Award. A. S. Wein was supported in part by NSF grant DMS-1712730 and by the Simons Collaboration on Algorithms and Geometry.

REFERENCES

[1] Emmanuel Abbe, João M. Pereira, and Amit Singer. 2017. Sample complexity of the Boolean multireference alignment problem. *arXiv preprint arXiv:1701.07540* (2017).

[2] Emmanuel Abbe, João M Pereira, and Amit Singer. 2018. Estimation in the group action channel. *arXiv preprint arXiv:1801.04366* (2018).

[3] Marc Adrian, Jacques Dubochet, Jean Lepault, and Alasdair W. McDowall. 1984. Cryo-electron microscopy of viruses. *Nature* 308, 5954 (1984), 32–36.

[4] Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-Kai Liu. 2012. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*. 917–925.

[5] Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham M Kakade. 2014. A tensor approach to learning mixed membership community models. *The Journal of Machine Learning Research* 15, 1 (2014), 2239–2312.

[6] Anima Anandkumar, Rong Ge, and Majid Janzamin. 2014. Analyzing tensor power method dynamics: Applications to learning overcomplete latent variable models. *CoRR abs/1411.1488* 17 (2014).

[7] Animashree Anandkumar, Rong Ge, and Majid Janzamin. 2017. Analyzing tensor power method dynamics in overcomplete regime. *Journal of Machine Learning Research* 18, 22 (2017), 1–40.

[8] Afonso Bandeira, Philippe Rigollet, and Jonathan Weed. 2017. Optimal rates of estimation for multi-reference alignment. *arXiv preprint arXiv:1702.08546* (2017).

[9] Afonso S. Bandeira. 2015. *Convex Relaxations for Certain Inverse Problems on Graphs*. Ph.D. Dissertation. Princeton University.

[10] Afonso S Bandeira, Ben Blum-Smith, Amelia Perry, Jonathan Weed, and Alexander S Wein. 2017. Estimation under group actions: recovering orbits from invariants. *arXiv preprint arXiv:1712.10163* (2017).

[11] Afonso S. Bandeira, Moses Charikar, Amit Singer, and Andy Zhu. 2014. Multireference alignment using semidefinite programming. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*. ACM, 459–470.

[12] Afonso S. Bandeira, Yutong Chen, and Amit Singer. 2015. Non-unique games over compact groups and orientation estimation in cryo-EM. *arXiv preprint arXiv:1505.03840* (2015).

[13] Boaz Barak, Samuel B Hopkins, Jonathan Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. 2016. A nearly tight sum-of-squares lower bound for the planted clique problem. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*. IEEE, 428–437.

[14] Tamir Bendory, Nicolas Boumal, Chao Ma, Zhizhen Zhao, and Amit Singer. 2017. Bispectrum inversion with application to multireference alignment. *IEEE Transactions on Signal Processing* 66, 4 (2017), 1037–1050.

[15] Tejal Bhamre, Teng Zhang, and Amit Singer. 2015. Orthogonal matrix retrieval in cryo-electron microscopy. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE, 1048–1052.

[16] Jacob Biamonte and Ville Bergholm. 2017. Tensor networks in a nutshell. *arXiv preprint arXiv:1708.00006* (2017).

[17] Nicolas Boumal, Tamir Bendory, Roy R Lederman, and Amit Singer. 2017. Heterogeneous multireference alignment: a single pass approach. *arXiv preprint arXiv:1710.02590* (2017).

[18] Hua Chen, Mona Zehni, and Zhizhen Zhao. 2018. A spectral method for stable bispectrum inversion with application to multireference alignment. *IEEE Signal Processing Letters* 25, 7 (2018), 911–915.

[19] Robert Diamond. 1992. On the multiple simultaneous superposition of molecular structures by rigid body transformations. *Protein Science* 1, 10 (October 1992), 1279–1287.

[20] Rong Ge and Tengyu Ma. 2015. Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms. *arXiv preprint arXiv:1504.05287* (2015).

[21] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. 2016. 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*. ACM, 178–191.

[22] Daniel Hsu and Sham M Kakade. 2013. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*. ACM, 11–20.

[23] Prateek Jain and Sewoong Oh. 2014. Learning mixtures of discrete product distributions using spectral decompositions. In *Conference on Learning Theory*. 824–856.

[24] Tengyu Ma, Jonathan Shi, and David Steurer. 2016. Polynomial-time tensor decompositions with sum-of-squares. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*. IEEE, 438–446.

[25] Ankur Moitra. 2018. *Algorithmic Aspects of Machine Learning*. Cambridge University Press.

[26] Ankur Moitra and Alexander S Wein. 2018. Spectral methods from tensor networks. *arXiv preprint arXiv:1811.00944* (2018).

[27] Elchanan Mossel and Sébastien Roch. 2005. Learning nonsingular phylogenies and hidden Markov models. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*. ACM, 366–375.

[28] Eva Nogales. 2016. The development of cryo-EM into a mainstream structural biology technique. *Nature Methods* 13, 1 (2016), 24–27.

[29] Amelia Perry, Jonathan Weed, Afonso Bandeira, Philippe Rigollet, and Amit Singer. 2017. The sample complexity of multi-reference alignment. *arXiv preprint arXiv:1707.00943* (2017).

[30] Amelia Perry, Alexander S. Wein, Afonso S. Bandeira, and Ankur Moitra. 2016. Message-passing algorithms for synchronization problems over compact groups. *arXiv preprint arXiv:1610.04583* (2016).

[31] R. Gil Pita, M. Rosa Zurera, P. Jarabo Amores, and F. López Ferreras. 2005. Using Multilayer Perceptrons to Align High Range Resolution Radar Signals. In *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005*. Springer Berlin Heidelberg, 911–916. https://doi.org/10.1007/11550907_144

[32] Aaron Potechin and David Steurer. 2017. Exact tensor completion with sum-of-squares. *arXiv preprint arXiv:1702.06237* (2017).

[33] Philippe Rigollet and Jan-Christian Hütter. 2018. High-dimensional statistics. *Lecture notes* (2018).

[34] Warren Schudy and Maxim Sviridenko. 2012. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 437–446.

[35] Amit Singer. 2011. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis* 30, 1 (2011), 20–36.

[36] Amit Singer and Yoel Shkolnisky. 2011. Three-dimensional structure determination from common lines in cryo-EM by eigenvectors and semidefinite programming. *SIAM Journal on Imaging Sciences* 4, 2 (2011), 543–572.

[37] Douglas L. Theobald and Phillip A. Steindel. 2012. Optimal simultaneous superpositioning of multiple structures with missing data. *Bioinformatics* 28, 15 (2012), 1972–1979.

[38] Alexander S. Wein. 2018. *Statistical Estimation in the Presence of Group Actions*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[39] J. P. Zwart, R. van der Heiden, S. Gelsema, and F. Groen. 2003. Fast translation invariant classification of HRR range profiles in a zero phase representation. *Radar, Sonar and Navigation, IEE Proceedings* 150, 6 (2003), 411–418.