MIT Libraries | DSpace@MIT

MIT Open Access Articles

Community detection in hypergraphs, spiked tensor models, and Sum-of-Squares

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation: Kim, Chiheon, Afonso S. Bandeira, and Michel X. Goemans. "Community Detection in Hypergraphs, Spiked Tensor Models, and Sum-of-Squares." 2017 International Conference on Sampling Theory and Applications (SampTA) (July 2017).

As Published: http://dx.doi.org/10.1109/SAMPTA.2017.8024470

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: http://hdl.handle.net/1721.1/116076

Version: Original manuscript: author's manuscript prior to formal peer review

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike

Massachusetts Institute of Technology

arXiv:1705.02973v1 [cs.DS] 8 May 2017

Community Detection in Hypergraphs, Spiked Tensor Models, and Sum-of-Squares

Chiheon Kim[*]
chiheonk@math.mit.edu

Afonso S. Bandeira[†]
bandeira@cims.nyu.edu

Michel X. Goemans[‡]
goemans@math.mit.edu

May 9, 2017

Abstract

We study the problem of community detection in hypergraphs under a stochastic block model. Similarly to how the stochastic block model in graphs suggests studying spiked random matrices, our model motivates investigating statistical and computational limits of exact recovery in a certain spiked tensor model. In contrast with the matrix case, the spiked model naturally arising from community detection in hypergraphs is different from the one arising in the so-called tensor Principal Component Analysis model. We investigate the effectiveness of algorithms in the Sum-of-Squares hierarchy on these models. Interestingly, our results suggest that these two apparently similar models exhibit significantly different computational to statistical gaps.

1 Introduction

Community detection is a central problem in many fields of science and engineering. It has received much attention for its various applications to sociological behaviours [1, 2, 3], protein-to-protein interactions [4, 5], DNA 3D conformation [6], recommendation systems [7, 8, 9], and more. In many networks with community structure one may expect that the groups of nodes within the same community are more densely connected. The *stochastic block model* (SBM)

[*] MIT, Department of Mathematics, 77 Massachusetts Ave., Cambridge, MA 02139

[†] Department of Mathematics and Center for Data Science, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012. Part of this work was done while A. S. Bandeira was with the Mathematics Department at MIT and supported by NSF Grant DMS-1317308.

[‡] MIT, Department of Mathematics, Room 2-474, 77 Massachusetts Ave., Cambridge, MA 02139. Partially supported by ONR grants N00014-14-1-0072 and N00014-17-1-2177.

[10] is arguably the simplest model that attempts to capture such community structure.

Under the SBM, each pair of nodes is connected randomly and independently with a probability decided by the community membership of the nodes. The SBM has received much attention for its sharp phase transition behaviours [11], [12], [13], computational versus information-theoretic gaps [14], [15], and as a test bed for many algorithmic approaches including semidefinite programming [13], [16], spectral methods [17], [18] and belief-propagation [19]. See [20] for a good survey on the subject.

Let us illustrate a version of the SBM with two equal-sized communities. Let $y \in \{\pm 1\}^n$ be a vector indicating community membership of an even number n of nodes and assume that the size of two communities are equal, i.e., $\mathbf{1}^T y = 0$. Let p and q be in $[0, 1]$ indicating the density of edges within and across communities, respectively. Under the model, a random graph G is generated by connecting nodes i and j independently with probability p if $y_i = y_j$ or with probability q if $y_i \neq y_j$. Let $\widehat{y}$ be any estimator of y given a sample G. We say that $\widehat{y}$ exactly recovers y if $\widehat{y}$ is equal to y or $-y$ with probability $1 - o_n(1)$. In the asymptotic regime of $p = a \log n/n$ and $q = b \log n/n$ where $a > b$, it has been shown that the maximum likelihood estimator (MLE) recovers y when $\sqrt{a} - \sqrt{b} > \sqrt{2}$ and the MLE fails when $\sqrt{a} - \sqrt{b} < \sqrt{2}$, showing a sharp phase transition behaviour [13]. Moreover, it was subsequently shown in [16] and [21] that the standard semidefinite programming (SDP) relaxation of the MLE achieves the optimal recovery threshold.

A fruitful way of studying phase transitions and effectiveness of different algorithms for the stochastic block model is to consider a Gaussian analogue of the model [22]. Let G be a graph generated by the SBM and let A_G be the adjacency matrix of G. We have

$$\mathbb{E} A_G = \frac{p+q}{2} \mathbf{1}\mathbf{1}^T + \frac{p-q}{2} y y^T - pI.$$

It is then useful to think A_G as a perturbation of the signal $\mathbb{E} A_G$ under centered noise.

This motivates a model with Gaussian noise. Given a vector $y \in \{\pm 1\}^n$ with $\mathbf{1}^T y = 0$, a random matrix T is generated as

$$T_{ij} = y_i y_j + \sigma W_{ij}$$

where $W_{ij} = W_{ji} \sim N(0, 1)$ for all $i, j \in [n]$. This model is often referred to $\mathbb{Z}_2$-synchronization [21, 22]. It is very closely related to the *spiked Wigner model* (or spiked random matrix models in general) which has a rich mathematical theory [23, 24, 25].

In many applications, however, nodes exhibit complex interactions that may not be well captured by pairwise interactions [26, 27]. One way to increase the descriptive ability of these models is to consider k-wise interactions, giving rise to generative models on random hypergraphs and tensors. Specifically, a hypergraphic version of the SBM was considered in [28, 29], and a version of the censored block model in [30].

For the sake of exposition we restrict our attention to $k = 4$ in the sequel. Most of our results however easily generalize to any k and will be presented in a subsequent publication. In the remaining of this section, we will introduce a hypergraphic version of the SBM and its Gaussian analogue.

1.1 Hypergraphic SBM and its Gaussian analogue

Here we describe a hypergraphic version of the stochastic block model. Let y be a vector in $\{\pm 1\}^n$ such that $\mathbf{1}^T y = 0$. Let p and q be in $[0, 1]$. Under the model, a random 4-uniform hypergraph H on n nodes is generated so that each $\{i, j, k, l\} \subseteq [n]$ is included in $E(H)$ independently with probability

$$\begin{cases} p & \text{if } y_i = y_j = y_k = y_l \text{ (in the same community)} \\ q & \text{otherwise (across communities).} \end{cases}$$

Let $\mathbf{A}_H$, the adjacency tensor of H, be the 4-tensor given by

$$(\mathbf{A}_H)_{ijkl} = \begin{cases} 1 & \text{if } \{i, j, k, l\} \in E(H) \\ 0 & \text{otherwise.} \end{cases}$$

Let $y^{\circledequal 4}$ be the 4-tensor defined as

$$y^{\circledequal 4}_{ijkl} = \begin{cases} 1 & \text{if } y_i = y_j = y_k = y_l \\ 0 & \text{otherwise,} \end{cases}$$

Since $y \in \{\pm 1\}^n$, we have

$$y^{\circledequal 4} = \left(\frac{\mathbf{1} + y}{2} \right)^{\otimes 4} + \left(\frac{\mathbf{1} - y}{2} \right)^{\otimes 4}.$$

Note that for any quadruple of distinct nodes (i, j, k, l) we have

$$(\mathbb{E}\mathbf{A}_H)_{ijkl} = (q\mathbf{1}^{\otimes 4} + (p - q) y^{\circledequal 4})_{ijkl}.$$

In the asymptotic regime of $p = a \log n / \binom{n-1}{3}$ and $q = b \log n / \binom{n-1}{3}$, there is a sharp information-theoretic threshold for exact recovery. Results regarding this hypergraphic stochastic block model will appear in subsequent publication. Here we will focus on the Gaussian counterpart.

Analogously to the relationship between the SBM and the spiked Wigner model, the hypergraphic version of the SBM suggests the following spiked tensor model:

$$\mathbf{T} = y^{\circledequal 4} + \sigma \mathbf{W},$$

where $\mathbf{W}$ is a random 4-tensor with i.i.d. standard Gaussian entries. We note that here the noise tensor $\mathbf{W}$ is not symmetric, unlike in the spiked Wigner model. This is not crucial: assuming $\mathbf{W}$ to be a symmetric tensor will only scale σ by $\sqrt{4!}$.

2 The Gaussian planted bisection model

Given a sample $\mathbf{T}$, our goal is to recover the hidden spike y up to a global sign flip. Let $\widehat{y}$ be an estimator of y computed from $\mathbf{T}$. Let $p(\widehat{y};\sigma)$ be the probability that $\widehat{y}$ successfully recovers y. Since $\mathbf{W}$ is Gaussian, $p(\widehat{y};\sigma)$ is maximized when

$$\widehat{y} = \operatorname*{argmin}_{x\in\{\pm 1\}^n:\mathbf{1}^T x=0} \|\mathbf{T} - x^{\ominus 4}\|_F^2,$$

or equivalently $\widehat{y} = \operatorname{argmax}_x \left\langle x^{\ominus 4}, \mathbf{T}\right\rangle$, the maximum-likelihood estimator (MLE) of y.

Let $f(x) = \left\langle x^{\ominus 4}, \mathbf{T}\right\rangle$ and $\widehat{y}_{ML}$ be the MLE of y. By definition,

$$p(\widehat{y}_{ML};\sigma) = \Pr\left(f(y) > \max_{\substack{x\in\{\pm 1\}\setminus\{y,-y\}\\ \mathbf{1}^T x=0}} f(x)\right).$$

Theorem 1. *Let $\epsilon > 0$ be a constant not depending on n. Then, as n grows, $p(\widehat{y}_{ML};\sigma)$ converges to 1 if $\sigma < (1-\epsilon)\sigma^*$ and $p(\widehat{y}_{ML};\sigma)$ converges to 0 if $\sigma > (1+\epsilon)\sigma^*$, where*

$$\sigma^* = \sqrt{\frac{1}{8}}\cdot\frac{n^{3/2}}{\sqrt{\log n}}.$$

Here we present a sketch of the proof while deferring the details to the appendix. Observe that $\widehat{y}_{ML}$ is not equal to y if there exists $x \in \{\pm 1\}^n$ distinct from y such that $\mathbf{1}^T x = 0$ and $f(x) \geqslant f(y)$. For each fixed x, the difference $f(x) - f(y)$ is equal to

$$\left\langle y^{\ominus 4}, x^{\ominus 4} - y^{\ominus 4}\right\rangle + \sigma\left\langle \mathbf{W}, x^{\ominus 4} - y^{\ominus 4}\right\rangle$$

which is a Gaussian random variable with mean $\left\langle y^{\ominus 4}, x^{\ominus 4} - y^{\ominus 4}\right\rangle$ and variance $\sigma^2\|x^{\ominus 4} - y^{\ominus 4}\|_F^2$. By definition we have

$$\left\langle x^{\ominus 4}, y^{\ominus 4}\right\rangle = \frac{1}{128}\left((\mathbf{1}^T\mathbf{1} + x^T y)^4 + (\mathbf{1}^T\mathbf{1} - x^T y)^4\right).$$

Let $\phi(t) = \frac{1}{128}((1+t)^4 + (1-t)^4)$. Then, $\left\langle x^{\ominus 4}, y^{\ominus 4}\right\rangle = n^4\phi(x^T y/n)$ so

$$\begin{aligned}\left\langle y^{\ominus 4}, x^{\ominus 4} - y^{\ominus 4}\right\rangle &= -n^4\left(\phi(1) - \phi(x^T y/n)\right),\\ \|x^{\ominus 4} - y^{\ominus 4}\|_F &= n^2\sqrt{2\phi(1) - 2\phi(x^T y/n)}.\end{aligned}$$

Hence, $\Pr\left(f(x) - f(y) \geqslant 0\right)$ is equal to

$$\Pr_{G\sim N(0,1)}\left(G \geqslant \frac{n^2}{\sigma\sqrt{2}}\cdot\sqrt{\phi(1) - \phi(x^T y/n)}\right).$$

This probability is maximized when x and y differs by only two indices, that is, $x^T y = n - 4$. Indeed one can formally prove that the probability that y

maximizes $f(x)$ is dominated by the probability that $f(y) > f(x)$ for all x with $x^T y = n - 4$. By union bound and standard Gaussian tail bounds, the latter probability is at most

$$\left(\frac{n}{2}\right)^2 \exp\left(-\frac{n^4}{4\sigma^2} \cdot (\phi(1) - \phi(1 - 4/n))\right)$$

which is approximately

$$\exp\left(2\log n - \frac{n^3 \cdot \phi'(1)}{\sigma^2}\right) = \exp\left(2\log n - \frac{n^3}{4\sigma^2}\right),$$

and it is $o_n(1)$ if $\sigma^2 < \frac{n^3}{8\log n}$ as in Theorem 1.

3 Efficient recovery

Before we address the Gaussian model, let us describe an algorithm for hypergraph partitioning. Let H be a 4-uniform hypergraph on the vertex set $V = [n]$. Let $\mathbf{A}_H$ be the adjacency tensor of H. Then the problem can be formulated as

$$\max \left\langle \mathbf{A}_H, x^{\ominus 4} \right\rangle \text{ subject to } x \in \{\pm 1\}^n, \mathbf{1}^T x = 0.$$

One approach for finding a partition is to consider the multigraph realization of H, which appears in [28, 29] in different terminology. Let G be the multigraph on $V = [n]$ such that the multiplicity of edge $\{i, j\}$ is the number of hyperedges $e \in E(H)$ containing $\{i, j\}$. One may visualize it as substituting each hyperedge by a 4-clique. Now, one may attempt to solve the reduced problem

$$\max \left\langle A_G, xx^T \right\rangle \text{ subject to } x \in \{\pm 1\}^n, \mathbf{1}^T x = 0$$

which is unfortunately NP-hard in general. Instead, we consider the semidefinite programming (SDP) relaxation

$$\begin{aligned} \max \quad & \langle A_G, X \rangle \\ \text{subject to} \quad & X_{ii} = 1 \text{ for all } i \in [n], \\ & \left\langle X, \mathbf{1}\mathbf{1}^T \right\rangle = 0, \\ & X \succeq 0. \end{aligned}$$

When A_G is generated under the graph stochastic block model, this algorithm recovers the hidden partition down to optimum parameters. On the other hand, it achieves recovery to nearly-optimum parameters when G is the multigraph corresponding to the hypergraph generated by a hypergraphic stochastic block model: there is a constant multiplicative gap between the guarantee and the information-theoretic limit. This will be treated in a future publication.

Here we had two stages in the algorithm: (1) "truncating" the hypergraph down to a multigraph, and (2) relaxing the optimization problem on the truncated objective function.

Now let us return to the Gaussian model $\mathbf{T} = y^{\ominus 4} + \sigma \mathbf{W}$. Let $f(x) = \langle x^{\ominus 4}, \mathbf{T} \rangle$. Our goal is to find the maximizer of $f(x)$. Note that

$$f(x) = \sum_{i_1, \cdots, i_4 \in [n]} \mathbf{T}_{i_1 i_2 i_3 i_4} \cdot \frac{1}{16} \left(\prod_{s=1}^{4} (1 + x_{i_s}) + \prod_{s=1}^{4} (1 - x_{i_s}) \right)$$

so $f(x)$ is a polynomial of degree 4 in variables $x_1, \ldots, x_n$. Let $f_{(2)}(x)$ be the degree 2 truncation of $f(x)$, i.e.,

$$f_{(2)}(x) = \frac{1}{8} \sum_{i_1, \cdots, i_4 \in [n]} \mathbf{T}_{i_1 i_2 i_3 i_4} \left(\sum_{1 \leqslant s < t \leqslant 4} x_{i_s} x_{i_t} \right).$$

Here we have ignored the constant term of $f(x)$ since it does not affect the maximizer. For each $\{s < t\} \subseteq \{1, 2, 3, 4\}$, let Q^{st} be n by n matrix where

$$Q^{st}_{ij} = \sum_{\substack{(i_1, \cdots, i_4) \in [n]^4 \\ i_s = i, i_t = j}} \mathbf{T}_{i_1 i_2 i_3 i_4}.$$

Then,

$$f_{(2)}(x) = \frac{1}{8} \langle Q, xx^T \rangle$$

where $Q = \sum_{1 \leqslant s < t \leqslant 4} Q^{st}$. This Q is analogous to the adjacency matrix of the multigraph constructed above. It is now natural to consider the following SDP:

$$\begin{aligned} \max \quad & \langle Q, X \rangle \\ \text{subject to} \quad & X_{ii} = 1 \text{ for all } i \in [n], \\ & \langle X, \mathbf{1}\mathbf{1}^T \rangle = 0, \\ & X \succeq 0. \end{aligned} \tag{1}$$

Theorem 2. *Let $\epsilon > 0$ be a constant not depending on n. Let $\widehat{Y}$ be a solution of (1) and $p(\widehat{Y}; \sigma)$ be the probability that $\widehat{Y}$ coincide with yy^T. If $\sigma < (1-\epsilon)\sigma^*_{(2)}$ where*

$$\sigma^*_{(2)} = \sqrt{\frac{3}{32}} \cdot \frac{n^{3/2}}{\sqrt{\log n}} = \sqrt{\frac{3}{4}} \cdot \sigma^*,$$

then $p(\widehat{Y}; \sigma) = 1 - o_n(1)$.

We present a sketch of the proof where details are deferred to the appendix. We note that a similar idea was used in [16, 21] for the graph stochastic block model.

We construct a dual solution of (1) which is feasible with high probability, and certifies that yy^T is the unique optimum solution for the primal (1). By complementary slackness, such dual solution must be of the form $S := D_{Q'} - Q'$ where $Q' = \mathrm{diag}(y) Q \mathrm{diag}(y)$ and $D_{Q'}$ is $\mathrm{diag}(Q'\mathbf{1})$. It remains to show that S is positive semidefinite with high probability.

To show that S is positive semidefinite, we claim that the second smallest eigenvalue of $\mathbb{E}S$ is $\Theta(n^3)$ and the operator norm $\|S - \mathbb{E}S\|$ is $O(\sigma n^{3/2}\sqrt{\log n})$ with high probability. The first part is just an easy calculation, and the second part is application of a nonasymptotic bound on Laplacian random matrices [21]. Hence, S is positive semidefinite with high probability if $\sigma n^{3/2}\sqrt{\log n} \lesssim n^3$, matching with the order of $\sigma^*_{(2)}$.

4 Standard Spiked Tensor Model

Montanari and Richard proposed a statistical model for tensor Principal Component Analysis [31]. In the model we observe a random tensor $\mathbf{T} = v^{\otimes 4} + \sigma \mathbf{W}$ where $v \in \mathbb{R}^n$ is a vector with $\|v\| = \sqrt{n}$ (spike), $\mathbf{W}$ is a random 4-tensor with i.i.d. standard Gaussian entries, and $\sigma \geqslant 0$ is the noise parameter. They showed a nearly-tight information-theoretic threshold for approximate recovery: when $\sigma \gg n^{3/2}$ then the recovery is information-theoretically impossible, while if $\sigma \ll n^{3/2}$ then the MLE gives a vector $v' \in \mathbb{R}^n$ with $|v^T v'| = (1 - o_n(1))n$ with high probability. Subsequently, sharp phase transitions for weak recovery, and strong and weak detection were shown in [32].

Those information-theoretic thresholds are achieved by the MLE for which no efficient algorithm is known. Montanari and Richard considered a simple spectral algorithm based on tensor unfolding, which is efficient in both theory and practice. They show that the algorithm finds a solution v' with $|v^T v'| = (1 - o_n(1))n$ with high probability as long as $\sigma = O(n)$ [31]. This is somewhat believed to be unimprovable using semidefinite programming [33, 34], or using approximate message passing algorithms [35].

For clear comparison to the Gaussian planted bisection model, let us consider when spike is in the hypercube $\{\pm 1\}^n$. Let $y \in \{\pm 1\}^n$ and $\sigma \geqslant 0$. Given a tensor

$$\mathbf{T} = y^{\otimes 4} + \sigma \mathbf{W}$$

where $\mathbf{W}$ is a random 4-tensor with independent, standard Gaussian entries, we would like to recover y *exactly*. The MLE is given by the maximizer of $g(x) := \langle x^{\otimes 4}, \mathbf{T} \rangle$ over all vectors $x \in \{\pm 1\}^n$.

Theorem 3. *Let $\epsilon > 0$ be any constant which does not depend on n. Let $\lambda^* = \frac{\sqrt{2}\cdot n^{3/2}}{\sqrt{\log n}}$. When $\sigma > (1+\epsilon)\lambda^*$, exact recovery is information-theoretically impossible (i.e. the MLE fails with $1 - o_n(1)$ probability), while if $\sigma < (1-\epsilon)\lambda^*$ then the MLE recovers y with $1 - o_n(1)$ probability.*

The proof is just a slight modification of the proof of Theorem 1 which appears in the appendix. Note that both λ^* and σ^* are in the order of $n^{3/2}/\sqrt{\log n}$. The standard spiked tensor model and the Gaussian planted bisection model exhibit similar behaviour when unbounded computational resources are given.

4.1 Sum-of-Squares based algorithms

Here we briefly discuss Sum-of-Squares based relaxation algorithms. Given a polynomial $p \in \mathbb{R}[x_1, \cdots, x_n]$, consider the problem of finding the maximum of $p(x)$ over $x \in \mathbb{R}^n$ satisfying polynomial equalities $q_1(x) = 0, \cdots, q_m(x) = 0$. Most hard combinatorial optimization problems can be reduced into this form, including max-cut, k-colorability, and general constraint satisfaction problems. The *Sum-of-Squares hierarchy* (SoS) is a systematic way to relax a polynomial optimization problem to a sequence of increasingly strong convex programs, each leading to a larger semidefinite program. See [36] for a good exposition of the topic.

There are many different ways to formulate the SoS hierarchy [37, 38, 39, 40]. Here we choose to follow the description based on *pseudo-expectation functionals* [36].

For illustration, let us recall the definition of $g(x)$: given a tensor $\mathbf{T} = y^{\otimes 4} + \sigma \mathbf{W}$, we define $g(x) = \left\langle x^{\otimes 4}, \mathbf{T} \right\rangle$. Here $g(x)$ is a polynomial of degree 4, and the corresponding maximum-likelihood estimator is the maximizer

$$\begin{aligned} \max \quad & g(x) \\ \text{subject to} \quad & x_i^2 = 1 \text{ for all } i \in [n], \\ & \sum_{i=1}^{n} x_i = 0, x \in \mathbb{R}^n. \end{aligned} \tag{2}$$

Let μ be a probability distribution over the set $\{x \in \{\pm 1\}^n : \mathbf{1}^T x = 0\}$. We can rewrite (2) as $\max_\mu \mathbb{E}_{x \sim \mu} g(x)$ over such distributions. A linear functional $\widetilde{\mathbb{E}}$ on $\mathbb{R}[x]$ is called *pseudoexpectation* of degree 2ℓ if it satisfies $\widetilde{\mathbb{E}}\, 1 = 1$ and $\widetilde{\mathbb{E}}\, q(x)^2 \geqslant 0$ for any $q \in \mathbb{R}[x]$ of degree at most ℓ. We note that any expectation $\mathbb{E}_\mu$ is a pseudoexpectation, but the converse is not true. So (2) can be relaxed to

$$\begin{aligned} \max \quad & \widetilde{\mathbb{E}}\, g(x) \\ \text{subject to} \quad & \widetilde{\mathbb{E}} \text{ is a pseudoexpectation of degree } 2\ell, \\ & \widetilde{\mathbb{E}} \text{ is zero on } I \end{aligned} \tag{3}$$

where $I \subseteq \mathbb{R}[x]$ is the ideal generated by $\sum_{i=1}^n x_i$ and $\{x_i^2 - 1\}_{i \in [n]}$.

The space of pseudoexpectations is a convex set which can be described as an affine section of the semidefinite cone. As ℓ increases, this space gets smaller and in this particular case, it coincides with the set of true expectations when $\ell = n$.

4.2 SoS on spiked tensor models

We would like to apply the SoS algorithm to the spiked tensor model. In particular, let us consider the degree 4 SoS relaxation of the maximum-likelihood problem. We note that for the spike tensor model with the spherical prior, it is known that neither algorithms using tensor unfolding [31], the degree 4 SoS relaxation of the MLE [33], nor approximate message passing [35] can achieve the statistical threshold, therefore the statistical-computational gap is believed to be present. Moreover, higher degree SoS relaxations were considered in [34]: for any small $\epsilon \geqslant 0$, degree at least $n^{\Omega(\epsilon)}$ is required in order to achieve recovery when $\sigma \approx n^{1+\epsilon}$.

We show an analogous gap of the degree 4 SoS relaxation for $\{\pm 1\}^n$ and $\mathbf{1}^T x$ prior.

Theorem 4. *Let $\mathbf{T} = y^{\otimes 4} + \sigma \mathbf{W}$ be as defined above. Let $g(x) = \left\langle x^{\otimes 4}, \mathbf{T} \right\rangle$. If $\sigma \lesssim \frac{n}{\log^{\Theta(1)} n}$, then the degree 4 SoS relaxation of $\max_x g(x)$ gives the solution y, i.e., $\widetilde{\mathbb{E}} g(x)$ is maximized when $\widetilde{\mathbb{E}}$ is the expectation operator of the uniform distribution on $\{y, -y\}$.*

On the other hand, if $\sigma \gtrsim n \log^{\Theta(1)} n$ then there exists a pseudoexpectation $\widetilde{\mathbb{E}}$ of degree 4 on the hypercube $\{\pm 1\}^n$ satisfying $\mathbf{1}^T x = 0$ such that

$$g(y) < \max_{\widetilde{\mathbb{E}}} \widetilde{\mathbb{E}}\, g(x),$$

so y is not recovered via the degree 4 SoS relaxation.

The proof of Theorem 4 is very similar to one that appears in [33, 34]. In the proof it is crucial to observe that

$$\frac{n^3}{\log^{\Theta(1)} n} \lesssim \max_{\substack{\widetilde{\mathbb{E}}\text{:degree 4} \\ \text{pseudo-exp.}}} \widetilde{\mathbb{E}} \left\langle \mathbf{W}, x^{\otimes 4} \right\rangle \lesssim n^3 \log^{\Theta(1)} n.$$

The upper bound can be shown via Cauchy-Schwarz inequality for pseudoexpectations and the lower bound, considerably more involved, can be shown by constructing a pseudoexpectation $\widetilde{\mathbb{E}}$ which is highly correlated to the entries of $\mathbf{W}$. We refer the readers to the appendix for details.

4.3 Comparison with the planted bisection model

Here we summarize the statistical-computational thresholds of two models, the planted bisection model and the spiked tensor model.

- For the planted bisection model, there is a constant $c > 0$ not depending on n such that for any $\epsilon > 0$ if $\sigma < (c - \epsilon)\frac{n^{3/2}}{\sqrt{\log n}}$ then recovery is information-theoretically possible, and if $\sigma > (c + \epsilon)\frac{n^{3/2}}{\sqrt{\log n}}$ then the recovery is impossible via any algorithm. Moreover, there is an efficient algorithm and a constant $c' < c$ such that the algorithm recovers y with high probability when $\sigma < c' \frac{n^{3/2}}{\sqrt{\log n}}$.
- For the spiked tensor model, there is a constant $C > 0$ not depending on n such that for any $\epsilon > 0$ if $\sigma < (C - \epsilon)\frac{n^{3/2}}{\sqrt{\log n}}$ then the recovery is information-theoretically possible, and if $\sigma > (C + \epsilon)\frac{n^{3/2}}{\sqrt{\log n}}$ then the

recovery is impossible. In contrast to the planted bisection model, all efficient algorithms known so far only successfully recover y in the asymptotic regime of $\sigma \lesssim n$.

We note that the efficient, nearly-optimal recovery is achieved for the planted bisection model by "forgetting" higher moments of the data. On the other hand, such approach is unsuitable for the spiked tensor model since the signal $y^{\otimes 4}$ does not seem to be approximable by a non-trivial low-degree polynomial. This might shed light into an interesting phenomenon in average complexity, as in this case it seems to be crucial that second moments (or pairwise relations) carry information.

All information-theoretic phase transition exhibited in the paper readily generalize to k-tensor models and this will be discussed in a future publication. In future work we will also investigate the performance of higher degree SoS algorithms for the planted bisection model.

References

[1] A. Goldenberg, A. X. Zheng, S. E. Fienberg, E. M. Airoldi *et al.*, "A survey of statistical network models," *Foundations and Trends® in Machine Learning*, vol. 2, no. 2, pp. 129–233, 2010.

[2] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.

[3] M. E. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. suppl 1, pp. 2566–2572, 2002.

[4] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, "Detecting protein function and protein-protein interactions from genome sequences," *Science*, vol. 285, no. 5428, pp. 751–753, 1999.

[5] J. Chen and B. Yuan, "Detecting functional modules in the yeast protein–protein interaction network," *Bioinformatics*, vol. 22, no. 18, pp. 2283–2290, 2006.

[6] I. Cabreros, E. Abbe, and A. Tsirigos, "Detecting community structures in hi-c genomic data," in *Information Science and Systems (CISS), 2016 Annual Conference on.* IEEE, 2016, pp. 584–589.

[7] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," *IEEE Internet computing*, vol. 7, no. 1, pp. 76–80, 2003.

[8] S. Sahebi and W. W. Cohen, "Community-based recommendations: a solution to the cold start problem," in *Workshop on recommender systems and the social web, RSWEB*, 2011.

[9] R. Wu, J. Xu, R. Srikant, L. Massoulié, M. Lelarge, and B. Hajek, "Clustering and inference from pairwise comparisons," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 43, no. 1. ACM, 2015, pp. 449–450.

[10] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[11] E. Mossel, J. Neeman, and A. Sly, "A proof of the block model threshold conjecture," *arXiv preprint arXiv:1311.4115*, 2013.

[12] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," in *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE, 2015, pp. 670–688.

[13] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 471–487, 2016.

[14] Y. Chen and J. Xu, "Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices," *Journal of Machine Learning Research*, vol. 17, no. 27, pp. 1–57, 2016.

[15] E. Abbe and C. Sandon, "Recovering communities in the general stochastic block model without knowing the parameters," in *Advances in neural information processing systems*, 2015, pp. 676–684.

[16] B. Hajek, Y. Wu, and J. Xu, "Achieving exact cluster recovery threshold via semidefinite programming," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2788–2797, 2016.

[17] L. Massoulié, "Community detection thresholds and the weak ramanujan property," in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. ACM, 2014, pp. 694–703.

[18] V. A. N. Vu, "A simple SVD algorithm for finding hidden partitions," pp. 1–12, 2014. [Online]. Available: http://arxiv.org/abs/1404.3918v1

[19] E. Abbe and C. Sandon, "Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap," *arXiv preprint arXiv:1512.09080*, 2015.

[20] E. Abbe, "Community detection and the stochastic block model: recent developments," 2016.

[21] A. S. Bandeira, "Random Laplacian Matrices and Convex Relaxations," pp. 1–35, 2016.

[22] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi, “Phase transitions in semidefinite relaxations,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 16, pp. E2218–E2223, 2016.

[23] D. Féral and S. Péché, “The largest eigenvalue of rank one deformation of large wigner matrices,” *Communications in mathematical physics*, vol. 272, no. 1, pp. 185–228, 2007.

[24] A. Montanari and S. Sen, “Semidefinite programs on sparse random graphs and their application to community detection,” *arXiv preprint arXiv:1504.05910*, 2015.

[25] A. Perry, A. S. Wein, A. S. Bandeira, and A. Moitra, “Optimality and sub-optimality of pca for spiked random matrices and synchronization,” *arXiv preprint arXiv:1609.05573*, 2016.

[26] S. Agarwal, J. Lim, L. Zelnik-Manor, P. Perona, D. Kriegman, and S. Belongie, “Beyond pairwise clustering,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 838–845.

[27] D. Zhou, J. Huang, and B. Schölkopf, “Learning with hypergraphs: Clustering, classification, and embedding,” in *NIPS*, vol. 19, 2006, pp. 1633–1640.

[28] D. Ghoshdastidar and A. Dukkipati, “Consistency of spectral partitioning of uniform hypergraphs under planted partition model,” in *Advances in Neural Information Processing Systems*, 2014, pp. 397–405.

[29] L. Florescu and W. Perkins, “Spectral thresholds in the bipartite stochastic block model,” *arXiv preprint arXiv:1506.06737*, 2015.

[30] K. Ahn, K. Lee, and C. Suh, “Community recovery in hypergraphs,” in *Allerton Conference on Communication, Control and Computing.* UIUC, 2016.

[31] A. Montanari and E. Richard, “A statistical model for tensor PCA,” p. 30, nov 2014. [Online]. Available: http://arxiv.org/abs/1411.1076

[32] A. Perry, A. S. Wein, and A. S. Bandeira, “Statistical limits of spiked tensor models,” p. 39, dec 2016. [Online]. Available: http://arxiv.org/abs/1612.07728

[33] S. Hopkins, “Tensor principal component analysis via sum-of-squares proofs,” *arXiv:1507.03269*, vol. 40, pp. 1–51, 2015.

[34] V. Bhattiprolu, V. Guruswami, and E. Lee, “Certifying Random Polynomials over the Unit Sphere via Sum of Squares Hierarchy,” 2016. [Online]. Available: http://arxiv.org/abs/1605.00903

[35] T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, and L. Zdeborová, "Statistical and computational phase transitions in spiked tensor estimation," *arXiv preprint arXiv:1701.08010*, 2017.

[36] B. Barak and D. Steurer, "Sum-of-squares proofs and the quest toward optimal algorithms," in *Proceedings of the International Congress of Mathematicians*, 2014.

[37] N. Z. Shor, "An approach to obtaining global extremums in polynomial mathematical programming problems," *Cybernetics*, vol. 23, no. 5, pp. 695–700, 1987.

[38] P. A. Parrilo, "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization," Ph.D. dissertation, California Institute of Technology, 2000.

[39] Y. Nesterov, "Squared functional systems and optimization problems," in *High performance optimization*. Springer, 2000, pp. 405–440.

[40] J. B. Lasserre, "Global optimization with polynomials and the problem of moments," *SIAM Journal on Optimization*, vol. 11, no. 3, pp. 796–817, 2001.

[41] P. Terwilliger, "The subconstituent algebra of an association scheme,(part i)," *Journal of Algebraic Combinatorics*, vol. 1, no. 4, pp. 363–388, 1992.

[42] A. Schrijver, "New code upper bounds from the terwilliger algebra and semidefinite programming," *IEEE Transactions on Information Theory*, vol. 51, no. 8, pp. 2859–2866, 2005.

[43] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

A Proof of Theorem 1

In this section, we prove Theorem 1 for general k.

Theorem 5 (Theorem 1, general k)**.** *Let k be a positive integer with $k \geqslant 2$. Let $y \in \{\pm 1\}^n$ with $\mathbf{1}^T y = 0$ and $\mathbf{T}$ be a k-tensor defined as $\mathbf{T} = y^{\circledequal k} + \sigma \mathbf{W}$ where $\mathbf{W}$ is a random k-tensor with i.i.d. standard Gaussian entries. Let $\widehat{y}_{ML}$ be the maximum-likelihood estimator of y, i.e.,*

$$\widehat{y}_{ML} = \operatorname*{argmax}_{x \in \{\pm 1\}^n : \mathbf{1}^T x} \left\langle \mathbf{T}, x^{\circledequal k} \right\rangle .$$

For any positive ϵ,

(i) $\widehat{y}_{ML}$ is equal to y with probability $1 - o_n(1)$ if $\sigma < (1-\epsilon)\sigma_*$, and

(ii) $\widehat{y}_{ML}$ is not equal to y with probability $1 - o_n(1)$ if $\sigma > (1+\epsilon)\sigma_*$

where

$$\sigma_* = \sqrt{\frac{k}{2^k}} \cdot \frac{n^{\frac{k-1}{2}}}{\sqrt{2 \log n}}.$$

First we prove the following lemma.

Lemma 6. *Let ϕ be a function defined as $\phi(t) = \frac{1}{2^{2k-1}} \left((1-t)^k + (1+t)^k \right)$. Then,*

$$\left\langle x^{\circledequal k}, y^{\circledequal k} \right\rangle = n^k \phi \left(\frac{x^T y}{n} \right)$$

for any $x, y \in \{\pm 1\}^n$ such that $\mathbf{1}^T x = \mathbf{1}^T y = 0$.

Proof. Note that $x^{\circledequal k} = \frac{1}{2^k} \left((\mathbf{1} + x)^{\otimes k} + (\mathbf{1} - x)^{\otimes k} \right)$. Hence,

$$\left\langle x^{\circledequal k}, y^{\circledequal k} \right\rangle = \frac{1}{2^{2k}} \sum_{s,t \in \{\pm 1\}} \langle \mathbf{1} + sx, \mathbf{1} + ty \rangle^k .$$

Since $\mathbf{1}^T x = \mathbf{1}^T y = 0$, we have

$$\begin{aligned}
\left\langle x^{\circledequal k}, y^{\circledequal k} \right\rangle &= \frac{1}{2^{2k}} \left(2(\mathbf{1}^T \mathbf{1} + x^T y)^k + 2(\mathbf{1}^T \mathbf{1} - x^T y)^k \right) \\
&= \frac{n^k}{2^{2k-1}} \left(\left(1 + \frac{x^T y}{n} \right)^k + \left(1 - \frac{x^T y}{n} \right)^k \right) \\
&= n^k \phi \left(\frac{x^T y}{n} \right)
\end{aligned}$$

as desired. □

Proof of Theorem 5. Let

$$f(x) = \left\langle x^{\circledequal k}, \mathbf{T} \right\rangle = \left\langle x^{\circledequal k}, y^{\circledequal k} + \sigma \mathbf{W} \right\rangle .$$

By definition, $\widehat{y}_{ML}$ is not equal to y if there exists $x \in \{\pm 1\}^n$ distinct from y or $-y$ such that $\mathbf{1}^T x = 0$ and $f(x)$ is greater than or equal to $f(y)$. For each fixed $x \in \{\pm 1\}^n$ with $\mathbf{1}^T x = 0$, note that

$$f(x) - f(y) = \left\langle y^{\circledequal k}, x^{\circledequal k} - y^{\circledequal k} \right\rangle + \sigma \left\langle \mathbf{W}, x^{\circledequal k} - y^{\circledequal k} \right\rangle$$

is a Gaussian random variable with mean $-n^k(\phi(1) - \phi(x^T y/n)$ and variance

$$\sigma^2 \|x^{\circledequal k} - y^{\circledequal k}\|_F^2 = 2\sigma^2 n^k(\phi(1) - \phi(x^T y/n)).$$

Hence, $\Pr\left(f(x) - f(y) \geqslant 0\right)$ is equal to

$$\Pr_{G \sim N(0,1)} \left(G \geqslant \frac{n^{k/2}}{\sigma\sqrt{2}} \cdot \sqrt{\phi(1) - \phi\left(\frac{x^T y}{n}\right)} \right).$$

Let σ_* be

$$\sigma_* = \sqrt{\phi'(1)} \cdot \frac{n^{\frac{k-1}{2}}}{\sqrt{2\log n}}.$$

Since $\phi'(1) = \frac{k}{2^k}$, it matches with the definition in the statement of the Theorem. *Upper bound.* Let us prove that $p(\widehat{y}_{ML}; \sigma) = 1 - o_n(1)$ if $\sigma < (1-\epsilon)\sigma_*$. By union bound, we have

$$\begin{aligned} 1 - p(\widehat{y}_{ML}; \sigma) \quad &\leqslant \quad \sum_{\substack{x \in \{\pm 1\}^n \setminus \{y, -y\} \\ \mathbf{1}^T x = 0}} \Pr\left(f(x) - f(y) \geqslant 0\right) \\ &= \quad \sum_{\substack{x \in \{\pm 1\}^n \setminus \{y, -y\} \\ \mathbf{1}^T x = 0}} \Pr_{G \sim N(0,1)} \left(G \geqslant \frac{n^{k/2}}{\sigma\sqrt{2}} \cdot \sqrt{\phi(1) - \phi\left(\frac{x^T y}{n}\right)} \right) \\ &\leqslant \quad \sum_{\substack{x \in \{\pm 1\}^n \setminus \{y, -y\} \\ \mathbf{1}^T x = 0}} \exp\left(-\frac{n^k}{4\sigma^2} \left(\phi(1) - \phi\left(\frac{x^T y}{n}\right) \right) \right). \end{aligned}$$

The last inequality follows from a standard Gaussian tail bound $\Pr_G(G > t) \leqslant \exp(-t^2/2)$.

A simple counting argument shows that the number of $x \in \{\pm 1\}^n$ with $\mathbf{1}^T x = 0$ and $x^T y = n - 4r$ is exactly $\binom{n/2}{r}^2$ for $r \in \{0, 1, \cdots, n/2\}$. Moreover, for any $t \geqslant 0$ we have $\phi(t) = \phi(-t)$. Hence,

$$1 - p(\widehat{y}_{ML}; \sigma) \leqslant 2 \sum_{r=1}^{\lceil n/4 \rceil} \binom{n/2}{r}^2 \exp\left(-\frac{n^k}{4\sigma^2} \left(\phi(1) - \phi\left(1 - \frac{4r}{n}\right) \right) \right).$$

Note that $\phi(1) - \phi(1-x) \geqslant \phi'(1)x - O(x^2)$. Hence, there exists an absolute constant $C > 0$ such that $\phi(1) - \phi(1 - 4r/n)$ is at least $(1-\epsilon)\phi'(1) \cdot 4r/n$ if $r < Cn$ and is at least $\Omega(1)$ otherwise. Since $\sigma < (1-\epsilon)\sigma^*$ we have

$$-\frac{n^k}{4\sigma^2} \left(\phi(1) - \phi\left(1 - \frac{4r}{n}\right) \right) \leqslant -\frac{2r\log n}{1-\epsilon} \leqslant -2(1+\epsilon) r \log n$$

if $r < Cn$, and $-\frac{n^k}{4\sigma^2}\left(\phi(1) - \phi\left(1 - \frac{4r}{n}\right)\right) = -\Omega(n \log n)$ otherwise. It implies that

$$\begin{aligned} 1 - p(\widehat{y}_{ML}; \sigma) &\leqslant 2\left(\sum_{r=1}^{Cn} \exp(-2\epsilon r \log n) + n \exp(-\Omega(n \log n))\right) \\ &\lesssim n^{-2\epsilon} + n \exp(-\Omega(n \log n)) = o_n(1). \end{aligned}$$

Lower bound. Now we prove that $p(\widehat{y}_{ML}; \sigma) = o_n(1)$ if $\sigma > (1 - \epsilon)\sigma^*$. Let $A = \{i \in [n] : y_i = +1\}$ and $B = [n] \backslash A$. For each $a \in A$ and $b \in B$, let E_{ab} be the event that $f(y^{(ab)})$ is greater than $f(y)$ where $y^{(ab)}$ is the ± 1-vector obtained by flipping the signs of y_a and y_b. For any $H_A \subseteq A$ and $H_B \subseteq B$, note that

$$\begin{aligned} 1 - p(\widehat{y}_{ML}; \sigma) &\geqslant \Pr\left(\bigcup_{a \in H_A, b \in H_B} E_{ab}\right) \\ &= \Pr\left(\max_{a \in H_A, b \in H_B} \left(f(y^{(ab)}) - f(y)\right) > 0\right) \end{aligned}$$

since any of the event E_{ab} implies that $\widehat{y}_{ML} \neq y$. Recall that

$$f(y^{(ab)}) - f(y) = -n^k\left(\phi(1) - \phi\left(1 - \frac{4}{n}\right)\right) + \sigma\left\langle \mathbf{W}, (y^{(ab)})^{\circledequal k} - y^{\circledequal k}\right\rangle.$$

So, $1 - p(\widehat{y}_{ML}; \sigma)$ is at least

$$\Pr\left(\max_{\substack{a \in H_A \\ b \in H_B}} \left\langle \mathbf{W}, (y^{(ab)})^{\circledequal k} - y^{\circledequal k}\right\rangle > \frac{n^k}{\sigma}\left(\phi(1) - \phi\left(1 - \frac{4}{n}\right)\right)\right).$$

Fix $H_A \subseteq A$ and $H_B \subseteq B$ with $|H_A| = |H_B| = h$ where $h = \frac{n}{\log^2 n}$. Let $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ be the partition of $[n]^k$ defined as

$$\begin{aligned} \mathcal{X} &= \{\alpha \in [n]^k : \alpha^{-1}(H_A \cup H_B) = \varnothing\}, \\ \mathcal{Y} &= \{\alpha \in [n]^k : |\alpha^{-1}(H_A \cup H_B)| = 1\}, \\ \mathcal{Z} &= \{\alpha \in [n]^k : |\alpha^{-1}(H_A \cup H_B)| \geqslant 2\}. \end{aligned}$$

Let $\mathbf{W}_{\mathcal{X}}$, $\mathbf{W}_{\mathcal{Y}}$ and $\mathbf{W}_{\mathcal{Z}}$ be the k-tensor supported on $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ respectively. For each $a \in H_A$ and $b \in H_B$, let

$$\begin{aligned} X_{ab} &= \left\langle \mathbf{W}_{\mathcal{X}}, (y^{(ab)})^{\circledequal k} - y^{\circledequal k}\right\rangle, \\ Y_{ab} &= \left\langle \mathbf{W}_{\mathcal{Y}}, (y^{(ab)})^{\circledequal k} - y^{\circledequal k}\right\rangle, \\ Z_{ab} &= \left\langle \mathbf{W}_{\mathcal{Z}}, (y^{(ab)})^{\circledequal k} - y^{\circledequal k}\right\rangle. \end{aligned}$$

Claim. *The followings are true:*

(i) $X_{ab} = 0$ *for any* $a \in H_A$ *and* $b \in H_B$.

(ii) *For fixed* $a \in H_A$ *and* $b \in H_B$*, the variables* Y_{ab} *and* Z_{ab} *are independent.*

(iii) *Each* Y_{ab} *can be decomposed into* $Y_a + Y_b$ *where* $\{Y_a\}_{a\in H_A} \cup \{Y_b\}_{b\in H_B}$ *is a collection of i.i.d. Gaussian random variables.*

Proof of Claim. First note that $\left((y^{(ab)})^{⊜k} - y^{⊜k}\right)_\alpha$ is non-zero if and only if $|\alpha^{-1}(a)|$ and $|\alpha^{-1}(b)|$ have the same parity. This implies (i) since when $\alpha \in \mathcal{X}$ we have $|\alpha^{-1}(a)| = |\alpha^{-1}(b)| = 0$. (ii) holds because $\mathcal{Y} \cap \mathcal{Z} = 0$.

For $s \in H_A \cup H_B$, let $\mathcal{Y}_s$ be the subset of $\mathcal{Y}$ such that

$$\mathcal{Y}_s = \{\alpha \in \mathcal{Y} : |\alpha^{-1}(s)| = 1\}.$$

By definition, $\mathcal{Y}_s$ are disjoint and $\mathcal{Y} = \bigcup_{s\in H_A\cup H_B} \mathcal{Y}_s$. Hence,

$$Y_{ab} = \sum_{s\in H_A\cup H_B} \left\langle \mathbf{W}_{\mathcal{Y}_s}, (y^{(ab)})^{⊜k} - y^{⊜k} \right\rangle.$$

Moreover, $\left\langle \mathbf{W}_{\mathcal{Y}_s}, (y^{(ab)})^{⊜k} - y^{⊜k} \right\rangle$ is zero when $s \notin \{a, b\}$. So,

$$Y_{ab} = \sum_{\alpha\in\mathcal{Y}_a\cup\mathcal{Y}_b} \mathbf{W}_\alpha((y^{(ab)})^{⊜k} - y^{⊜k})_\alpha.$$

Note that for $\alpha \in \mathcal{Y}_a$

$$((y^{(ab)})^{⊜k} - y^{⊜k})_\alpha = \begin{cases} +1 & \text{if } |\alpha^{-1}(A\backslash H_A)| = 0 \\ -1 & \text{if } |\alpha^{-1}(B\backslash H_B)| = 0 \\ 0 & \text{otherwise.} \end{cases}$$

So,

$$\sum_{\alpha\in\mathcal{Y}_a} \mathbf{W}_\alpha((y^{(ab)})^{⊜k} - y^{⊜k})_\alpha = \sum_{\substack{\alpha\in\mathcal{Y}_a \\ |\alpha^{-1}(A\backslash H_A)|=0}} \mathbf{W}_\alpha - \sum_{\substack{\alpha\in\mathcal{Y}_a \\ |\alpha^{-1}(B\backslash H_B)|=0}} \mathbf{W}_\alpha$$

which does not depend on the choice of b. Let

$$Y_a = \sum_{\substack{\alpha\in\mathcal{Y}_a \\ |\alpha^{-1}(A\backslash H_A)|=0}} \mathbf{W}_\alpha - \sum_{\substack{\alpha\in\mathcal{Y}_a \\ |\alpha^{-1}(B\backslash H_B)|=0}} \mathbf{W}_\alpha$$

and Y_b respectively. Then $Y_{ab} = Y_a + Y_b$ and $\{Y_s\}_{s\in H_A\cup H_B}$ is a collection of independent Gaussian random variables. Moreover, the variance of Y_s is equal to $2k\left(\frac{n}{2} - h\right)^{k-1}$, independent of the choice of s. □

By the claim, we have $\left\langle \mathbf{W}, (y^{(ab)})^{⊜k} - y^{⊜k} \right\rangle = Y_a + Y_b + Z_{ab}$. Moreover,

$$\begin{aligned} \max_{\substack{a\in H_A \\ b\in H_B}} (Y_a + Y_b + Z_{ab}) &\geqslant \max_{\substack{a\in H_A \\ b\in H_B}} (Y_a + Y_b) - \max_{\substack{a\in H_A \\ b\in H_B}} (-Z_{ab}) \\ &= \max_{a\in H_A} Y_a + \max_{b\in H_B} Y_b - \max_{\substack{a\in H_A \\ b\in H_B}} (-Z_{ab}). \end{aligned}$$

We need the following tail bound on the maximum of Gaussian random variables.

Lemma 7. *Let $G_1, \ldots, G_N$ be (not necessarily independent) Gaussian random variables with variance* 1*. Let $\epsilon > 0$ be a constant which does not depend on N. Then,*

$$\Pr\left(\max_{i=1,\cdots,N} G_i > (1+\epsilon)\sqrt{2\log N} \right) \leqslant N^{-\epsilon}.$$

On the other hand,

$$\Pr\left(\max_{i=1,\cdots,N} G_i < (1-\epsilon)\sqrt{2\log N} \right) \leqslant \exp(-N^{\Omega(\epsilon)})$$

if G_i's are independent.

By the Lemma, we have

$$\begin{aligned}
\max_{a\in H_A} Y_a &\geqslant (1-0.01\epsilon)\sqrt{2\log h \cdot 2k\left(\frac{n}{2}-h\right)^{k-1}} \\
\max_{b\in H_A} Y_b &\geqslant (1-0.01\epsilon)\sqrt{2\log h \cdot 2k\left(\frac{n}{2}-h\right)^{k-1}} \\
\max_{\substack{a\in H_A\\ b\in H_B}} Z_{ab} &\lesssim \sqrt{\log h \cdot \max \operatorname{Var}(Z_{ab})}
\end{aligned}$$

with probability $1 - o_n(1)$. Note that

$$\begin{aligned}
\operatorname{Var}(Z_{ab}) &= \|(y^{(ab)})^{\circledequal k} - y^{\circledequal k}\|_F^2 - (\operatorname{Var}(Y_A) + \operatorname{Var}(Y_B)) \\
&= 2n^k\left(\phi(1) - \phi\left(1-\frac{4}{n}\right)\right) - 4k\left(\frac{n}{2}-h\right)^{k-1} \\
&\leqslant 8n^{k-1}\left(\phi'(1) - (1-o(1))\frac{k}{2^k}\right)
\end{aligned}$$

which is $o(n^{k-1})$. Hence,

$$\begin{aligned}
\max_{\substack{a\in H_A\\ b\in H_B}} \left\langle \mathbf{W}, (y^{(ab)})^{\circledequal k} - y^{\circledequal k} \right\rangle &\geqslant 2(1-0.01\epsilon - o(1))\sqrt{2\log n \cdot 2k\left(\frac{n}{2}-h\right)^{k-1}} \\
&\geqslant (1-0.01\epsilon - o(1))\sqrt{\frac{kn^{k-1}\log n}{2^{k-5}}}.
\end{aligned}$$

On the other hand, since $\sigma > (1+\epsilon)\sigma^*$ we have

$$\begin{aligned}
\frac{n^k}{\sigma}\left(\phi(1) - \phi\left(1-\frac{4}{n}\right)\right) &< \frac{n^k}{1+\epsilon}\cdot\frac{4\phi'(1)}{n}\cdot\sqrt{\frac{2\log n}{n^{k-1}\phi'(1)}} \\
&< \frac{1}{1+\epsilon}\sqrt{\frac{n^{k-1}\log n}{2^{k-5}}}
\end{aligned}$$

which is less than

$$\max_{\substack{a \in H_A \\ b \in H_B}} \left\langle \mathbf{W}, (y^{(ab)})^{\circledequal k} - y^{\circledequal k} \right\rangle$$

with probability $1 - o_n(1)$. Thus, $1 - p(\widehat{y}_{ML}) \geqslant 1 - o_n(1)$. □

B Proof of Theorem 2

In this section, we prove Theorem 2 for general k.

Let k be a positive integer with $k > 2$. Let $y \in \{\pm 1\}^n$ with $\mathbf{1}^T y = 0$ and $\mathbf{T}$ be a k-tensor defined as $\mathbf{T} = y^{\circledequal k} + \sigma \mathbf{W}$ where $\mathbf{W}$ is a random k-tensor with i.i.d. standard Gaussian entries. Let $f(x) = \left\langle x^{\circledequal k}, \mathbf{T} \right\rangle$ and let $f_{(2)}(x)$ be the degree 2 truncation of $f(x)$, i.e.,

$$f_{(2)}(x) = \sum_{\alpha \in [n]^k} \mathbf{T}_\alpha \left(\frac{1}{2^{k-1}} \sum_{1 \leqslant s < t \leqslant k} x_{\alpha(s)} x_{\alpha(t)} \right).$$

For each $\{s < t\} \subseteq [k]$, let Q^{st} be n by n matrix where

$$Q^{st}_{ij} = \frac{1}{2} \left(\sum_{\substack{\alpha \in [n]^k \\ \alpha(s)=i, \alpha(t)=j}} \mathbf{T}_\alpha + \sum_{\substack{\alpha \in [n]^k \\ \alpha(s)=j, \alpha(t)=i}} \mathbf{T}_\alpha \right)$$

Then,

$$f_{(2)}(x) = \frac{1}{2^{k-1}} \left\langle Q, xx^T \right\rangle$$

where $Q = \sum_{1 \leqslant s < t \leqslant k} Q^{st}$. We consider the following SDP relaxation for $\max_x f_{(2)}(x)$:

$$\begin{aligned} \max \quad & \langle Q, X \rangle \\ \text{subject to} \quad & X_{ii} = 1 \text{ for all } i \in [n], \\ & \left\langle X, \mathbf{1}\mathbf{1}^T \right\rangle = 0, \\ & X \succeq 0. \end{aligned} \tag{4}$$

Theorem 8 (Theorem 2, general k)**.** *Let $\epsilon > 0$ be a constant not depending on n. Let $\widehat{Y}$ be a solution of (4) and $p(\widehat{Y}; \sigma)$ be the probability that $\widehat{Y}$ coincide with yy^T. Let $\sigma_{(2)}$ be*

$$\sigma_{(2)} = \sqrt{\frac{k(k-1)}{2^{2k-1}}} \cdot \frac{n^{\frac{k-1}{2}}}{\sqrt{2 \log n}}$$

If $\sigma < (1 - \epsilon)\sigma_{(2)}$, then $p(\widehat{Y}; \sigma) = 1 - o_n(1)$.

Proof. First note that $Q^{st} = \frac{n^{k-2}}{2^{k-1}} yy^T + \sigma W^{st}$ where

$$W^{st}_{ij} = \sum_{\substack{\alpha \in [n]^k \\ \alpha(s)=i, \alpha(t)=j}} \mathbf{W}_\alpha.$$

So we have

$$Q = n^{k-2} \cdot \frac{k(k-1)}{2^k} yy^T + \sigma \overline{W}$$

where $\overline{W} = \sum_{1 \leqslant s < t \leqslant k} W^{st}$.

The dual program of (4) is

$$\begin{array}{rl} \max & \operatorname{tr}(D) \\ \text{subject to} & D + \lambda \mathbf{1}\mathbf{1}^T - Q \succeq 0, \\ & D \text{ is diagonal.} \end{array} \tag{5}$$

By complementary slackness, yy^T is the unique optimum solution of (4) if there exists a dual feasible solution (D, λ) such that

$$\left\langle D + \lambda \mathbf{1}\mathbf{1}^T - Q, yy^T \right\rangle = 0$$

and the second smallest eigenvalue of $D + \lambda \mathbf{1}\mathbf{1}^T - Q$ is greater than zero. For brevity, let $S = D + \lambda \mathbf{1}\mathbf{1}^T - Q$. Since S is positive semidefinite, we must have $Sy = 0$, that is,

$$D_{ii} = \sum_{j=1}^{n} Q_{ij} y_i y_j$$

for any $i \in [n]$.

For a symmetric matrix M, we define the *Laplacian* $\mathcal{L}(M)$ of M as $\mathcal{L}(M) = \operatorname{diag}(M\mathbf{1}) - M$ (See [21]). Using this language, we can express S as

$$S = \operatorname{diag}(y) \left(\mathcal{L}(\operatorname{diag}(y) Q \operatorname{diag}(y)) + \lambda yy^T \right) \operatorname{diag}(y).$$

Note that

$$\operatorname{diag}(y) Q \operatorname{diag}(y) = n^{k-2} \cdot \frac{k(k-1)}{2^k} \mathbf{1}\mathbf{1}^T + \sigma \operatorname{diag}(y) \overline{W} \operatorname{diag}(y).$$

Hence, the Laplacian of $\operatorname{diag}(y)Q\operatorname{diag}(y)$ is equal to

$$\underbrace{n^{k-1} \cdot \frac{k(k-1)}{2^k} \left(I_{n \times n} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right)}_{\text{deterministic part}} + \underbrace{\sigma \mathcal{L} \left(\operatorname{diag}(y) \overline{W} \operatorname{diag}(y) \right)}_{\text{noisy part}}.$$

This matrix is positive semidefinite if

$$\sigma \left\| \mathcal{L}(\operatorname{diag}(y) \overline{W} \operatorname{diag}(y) \right\| \leqslant n^{k-1} \cdot \frac{k(k-1)}{2^k}.$$

Moreover, if the inequality is strict then the second smallest eigenvalue of S is greater than zero.

By triangle inequality, we have

$$\begin{aligned} \left\| \mathcal{L}(\operatorname{diag}(y) \overline{W} \operatorname{diag}(y)) \right\| &\leqslant \max_{i \in [n]} \sum_{j=1}^{n} \overline{W}_{ij} y_i y_j + \sum_{1 \leqslant s < t \leqslant k} \left\| \operatorname{diag}(y) W^{st} \operatorname{diag}(y) \right\| \\ &\leqslant \max_{i \in [n]} \sum_{j=1}^{n} \overline{W}_{ij} y_i y_j + \binom{k}{2} \sqrt{2n^{k-1}}. \end{aligned}$$

The second inequality holds with high probability since W^{st} has independent Gaussian entries. Since $\sum_{j=1}^{n} \overline{W}_{ij} y_i y_j$ is Gaussian and centered, with high probability we have that

$$\begin{aligned}\max_{i\in[n]} \sum_{j=1}^{n} \overline{W}_{ij} y_i y_j &\leqslant (1+0.1\epsilon)\sqrt{2\log n} \cdot \left(\max_{i\in[n]} \operatorname{Var}\left(\sum_{j=1}^{n} \overline{W}_{ij} y_i y_j\right)\right)^{1/2} \\ &\leqslant (1+0.1\epsilon)\sqrt{2\log n} \cdot \sqrt{\binom{k}{2} n^{k-1}}.\end{aligned}$$

Hence,

$$\left\|\mathcal{L}(\operatorname{diag}(y)\overline{W}\operatorname{diag}(y))\right\| \leqslant (1+0.1\epsilon+o(1))\sqrt{2\binom{k}{2} n^{k-1}\log n}$$

with high probability. So, $S \geq 0$ as long as

$$\sigma(1+0.1\epsilon+o(1))\sqrt{2\binom{k}{2} n^{k-1}\log n} < n^{k-1} \cdot \frac{k(k-1)}{2^k}$$

or simply

$$(1+0.1\epsilon+o(1))\sigma < \sqrt{\frac{k(k-1)}{2^{2k-1}}} \cdot \frac{n^{\frac{k-1}{2}}}{\sqrt{2\log n}} = \sigma_{(2)}.$$

So, when $\sigma < (1-\epsilon)\sigma_{(2)}$ with high probability $\widehat{Y} = yy^T$. □

C Pseudo-expectation and its moment matrix

C.1 Notation and preliminaries

Let $\mathcal{V}$ be the space of real-valued functions on the n-dimensional hypercube $\{-1,+1\}^n$. For each $S \subseteq [n]$, let $x_S = \prod_{i\in S} x_i$. Note that $\{x_S : S \subseteq [n]\}$ is a basis of $\mathcal{V}$. Hence, any function $f : \{\pm 1\}^n \to \mathbb{R}$ can be written as a unique linear combination of multilinear monomials, say

$$f(x) = \sum_{S\subseteq[n]} f_S x_S.$$

The degree of f is defined as the maximum size of $S \subset [n]$ such that f_S is nonzero.

Let ℓ be a positive integer. Let us denote the collection of subsets of $[n]$ of size at most ℓ by $\binom{[n]}{\leqslant \ell}$, and the size $\left|\binom{[n]}{\leqslant \ell}\right|$ of it by $\binom{n}{\leqslant \ell}$.

Let M be a square symmetric matrix of size $\binom{n}{\leqslant \ell}$. The rows and the columns of M are indexed by the elements in $\binom{[n]}{\leqslant \ell}$. To avoid confusion, we use $M[S,T]$ to denote the entry of M at row S and column T.

We say M is *SoS-symmetric* if $M[S,T] = M[S',T']$ whenever $x_S x_T = x_{S'} x_{T'}$ on the hypercube. Since $x \in \{\pm 1\}^n$, it means that $S \oplus T = S' \oplus T'$ where $S \oplus T$ denotes the symmetric difference of S and T. Given $f \in \mathcal{V}$ with degree at most 2ℓ, we say M *represents* f if

$$f_U = \sum_{\substack{S,T \in \binom{[n]}{\leqslant \ell}: \\ S \oplus T = U}} M[S,T].$$

We use M_f to denote the unique SoS-symmetric matrix representing f.

Let $\mathcal{L}$ be a linear functional on $\mathcal{V}$. By linearity, $\mathcal{L}$ is determined by $(\mathcal{L}[x_S] : S \subseteq [n])$. Let $X_{\mathcal{L}}$ be the SoS-symmetric matrix of size $\binom{n}{\leqslant \ell}$ with entries $X[S,T] = \mathcal{L}[x_{S \oplus T}]$. We call $X_{\mathcal{L}}$ the *moment matrix of* $\mathcal{L}$ *of degree* 2ℓ.

By definition, we have

$$\begin{aligned} \mathcal{L}[f] &= \sum_{U \in \binom{[n]}{\leqslant 2\ell}} f_U \mathcal{L}[x_U] \\ &= \sum_{S,T \in \binom{[n]}{\leqslant \ell}} M_f[S,T] X_{\mathcal{L}}[S,T] \\ &= \langle X_{\mathcal{L}}, M_f \rangle \end{aligned}$$

for any f of degree at most 2ℓ.

C.2 A quick introduction to pseudo-expectations

For our purpose, we only work on pseudo-expectations defined on the hypercube $\{\pm 1\}^n$. See [36] for general definition.

Let ℓ be a positive integer and $d = 2\ell$. A *pseudo-expectation of degree* d *on* $\{\pm 1\}^n$ is a linear functional $\widetilde{\mathbb{E}}$ on the space $\mathcal{V}$ of functions on the hypercube such that

(i) $\widetilde{\mathbb{E}}[1] = 1$,

(ii) $\widetilde{\mathbb{E}}[q^2] \geqslant 0$ for any $q \in \mathcal{V}$ of degree at most $\ell = d/2$.

We say $\widetilde{\mathbb{E}}$ *satisfies* the system of equalities $\{p_i(x) = 0\}_{i=1}^m$ if $\widetilde{\mathbb{E}}[f] = 0$ for any $f \in \mathcal{V}$ of degree at most d which can be written as

$$f = p_1 q_1 + p_2 q_2 + \cdots + p_m q_m$$

for some $q_1, q_2, \ldots, q_m \in \mathcal{V}$.

We note the following facts:

- If $\widetilde{\mathbb{E}}$ is a pseudo-expectation of degree d, then it is also a pseudo-expectation of any degree smaller than d.
- If $\widetilde{\mathbb{E}}$ is a pseudo-expectation of degree $2n$, then $\widetilde{\mathbb{E}}$ defines a valid probability distribution supported on $P := \{x \in \{\pm 1\}^n : p_i(x) = 0 \text{ for all } i \in [m]\}$.

The second fact implies that maximizing $f(x)$ on P is equivalent to maximizing $\widetilde{\mathbb{E}}[f]$ over all pseudo-expectations of degree $2n$ satisfying $\{p_i(x) = 0\}_{i=1}^m$. Now, let d be an even integer such that

$$d > \max\{\deg(f), \deg(p_1), \cdots, \deg(p_m)\}.$$

We relax the original problem to

$$\begin{aligned} \max \quad & \widetilde{\mathbb{E}}[f] \\ \text{subject to} \quad & \widetilde{\mathbb{E}} \text{ is degree-}d \text{ pseudo-expectation on } \{\pm 1\}^n \\ & \text{satisfying } \{p_i(x) = 0\}_{i=1}^m. \end{aligned} \tag{SoS_d}$$

We note that the value of (SoS_d) decreases as d grows, and it reaches the optimum value $\max_{x \in P} p_0(x)$ of the original problem at $d = 2n$.

C.3 Matrix point of view

Let $\widetilde{\mathbb{E}}$ be a pseudo-expectation of degree $d = 2\ell$ for some positive integer ℓ. Suppose that $\widetilde{\mathbb{E}}$ satisfies the system $\{p_i(x) = 0\}_{i=1}^m$. Let $X_{\widetilde{\mathbb{E}}}$ be the moment matrix of $\widetilde{\mathbb{E}}$ of degree 2ℓ, hence the size of $X_{\widetilde{\mathbb{E}}}$ is $\binom{n}{\leqslant \ell}$.

Conditions for $\widetilde{\mathbb{E}}$ being pseudo-expectation translate as the following conditions for $X_{\widetilde{\mathbb{E}}}$:

(i) $X_{\varnothing,\varnothing} = 1$.

(ii) X is positive semidefinite.

Moreover, $\widetilde{\mathbb{E}}$ satisfies $\{p_i(x) = 0\}_{i=1}^m$ if and only if

(iii) Let $\mathcal{U}$ be the space of functions in $\mathcal{V}$ of degree at most d which can be written as $\sum_{i=1}^m p_i q_i$ for some $q_i \in \mathcal{V}$. Then, $\left\langle M_f, X_{\widetilde{\mathbb{E}}} \right\rangle = 0$ for any $f \in \mathcal{U}$.

Hence, (SoS_d) can be written as the following semidefinite program

$$\begin{aligned} \max \quad & \langle M_f, X \rangle \\ \text{subject to} \quad & X_{\varnothing,\varnothing} = 1 \\ & \langle M_q, X \rangle = 0 \text{ for all } q \in \mathcal{B} \\ & X \succeq 0, \end{aligned} \tag{SDP_d}$$

where $\mathcal{B}$ is any finite subset of $\mathcal{U}$ which spans $\mathcal{U}$, for example,

$$\mathcal{B} = \{x_S p_i(x) : i \in [m], |S| \leqslant d - \deg(p_i)\}.$$

D Proof of Theorem 4

Let $y \in \{\pm 1\}^n$ such that $\mathbf{1}^T y = 0$, $\sigma > 0$, and $\mathbf{W} \in (\mathbb{R}^n)^{\otimes 4}$ be 4-tensor with independent, standard Gaussian entries. Given a tensor $\mathbf{T} = y^{\otimes 4} + \sigma \mathbf{W}$, we would like to recover the planted solution y from $\mathbf{T}$. Let $f(x) = \langle \mathbf{T}, x^{\otimes 4} \rangle$. The maximum-likelihood estimator is given by the optimum solution of

$$\max_{x \in \{\pm 1\}^n : \mathbf{1}^T x = 0} f(x).$$

Consider the SoS relaxation of degree 4

$$\begin{aligned} \max \quad & \widetilde{\mathbb{E}}[f] \\ \text{subject to} \quad & \widetilde{\mathbb{E}} \text{ is a pseudo-expectation of degree 4 on } \{\pm 1\}^n \\ & \text{satisfying } \sum_{i=1}^n x_i = 0. \end{aligned}$$

Let $\mathbb{E}_{U(\{y,-y\})}$ be the expectation operator of the uniform distribution on $\{y, -y\}$, i.e., $\mathbb{E}_{U(\{y,-y\})}[x_S] = y_S$ if $|S|$ is even, and $\mathbb{E}_{U(\{y,-y\})}[x_S] = 0$ if $|S|$ is odd.

If $\mathbb{E}_{U(\{y,-y\})}$ is the optimal solution of the relaxation, then we can recover y from it up to a global sign flip. First we give an upper bound on σ to achieve it with high probability.

Theorem 9 (Part one of Theorem 4)**.** *Let* $\mathbf{T} = y^{\otimes 4} + \sigma \mathbf{W}$ *and* $f(x) = \langle \mathbf{T}, x^{\otimes 4} \rangle$ *be as defined above. If* $\sigma \lesssim \frac{n}{\sqrt{\log n}}$, *then the relaxation* $\max_{\widetilde{\mathbb{E}}} \widetilde{\mathbb{E}}[f]$ *over pseudo-expectation* $\widetilde{\mathbb{E}}$ *of degree 4 satisfying* $\sum_{i=1}^n x_i = 0$ *is maximized when* $\widetilde{\mathbb{E}} = \mathbb{E}_{U(\{y,-y\})}$ *with probability* $1 - o_n(1)$.

We can reduce Theorem 9 to the matrix version of the problem via flattening [31]. Given a 4-tensor $\mathbf{T}$, the canonical flattening of $\mathbf{T}$ is defined as $n^2 \times n^2$ matrix T with entries $T_{(i,j),(k,\ell)} = \mathbf{T}_{ijk\ell}$. Then, $T = \widetilde{y}\widetilde{y}^T + \sigma W$ where $\widetilde{y}$ is the vectorization of yy^T, and W is the flattening of $\mathbf{W}$. Note that this is an instance of $\mathbb{Z}_2$-synchronization model with Gaussian noises. It follows that with high probability the exact recovery is possible when $\sigma \lesssim \frac{n}{\sqrt{\log n}}$ (see Proposition 2.3 in [21]).

We complement the result by providing a lower bound on σ which is off by polylog factor.

Theorem 10 (Part two of Theorem 4)**.** *Let* $c > 0$ *be a small constant. If* $\sigma \geqslant n(\log n)^{1/2+c}$, *then there exists a pseudo-expectation* $\widetilde{\mathbb{E}}$ *of degree 4 on the hypercube* $\{\pm 1\}^n$ *satisfying* $\sum_{i=1}^n x_i = 0$ *such that* $\widetilde{\mathbb{E}}[f] > f(y)$ *with probability* $1 - o_n(1)$.

D.1 Proof of Theorem 10

Let $g(x)$ be the noise part of $f(x)$, i.e., $g(x) = \langle \mathbf{W}, x^{\otimes 4} \rangle$. Let $\widetilde{\mathbb{E}}$ be a pseudo-expectation of degree 4 on the hypercube which satisfies the equality $\sum_{i=1}^n x_i = 0$. We have $\widetilde{\mathbb{E}}[f] \geqslant \sigma \widetilde{\mathbb{E}}[g]$ since $\widetilde{\mathbb{E}}[(x^T y)^4] \geqslant 0$.

Lemma 11. *Let $g(x)$ be the polynomial as defined above. Then, there exists a pseudo-expectation of degree 4 on the hypercube satisfying $\mathbf{1}^T x = 0$ such that*

$$\widetilde{\mathbb{E}}[g] \gtrsim \frac{n^3}{(\log n)^{1/2+o(1)}}.$$

We prove Theorem 10 using the lemma.

Proof of Theorem 10. Note that $g(y) = \left\langle \mathbf{W}, y^{\otimes 4} \right\rangle$ is a Gaussian random variable with variance n^4. So, $g(y) \leqslant n^2 \log n$ with probability $1 - o(1)$. Let $\widetilde{\mathbb{E}}$ be the pseudo-expectation satisfying the conditions in the lemma. Then, with probability $1 - o(1)$ we have

$$\begin{aligned} \widetilde{\mathbb{E}}[f] - f(y) &= -(n^4 - \widetilde{\mathbb{E}}[(y^T x)^4]) + \sigma(\widetilde{\mathbb{E}}[g] - g(y)) \\ &\geqslant -n^4 + \sigma\left(\frac{n^3}{\log n} + n^2 \log n\right) \\ &\geqslant -n^4 + (1 - o(1))\frac{\sigma n^3}{(\log n)^{1/2+o(1)}}. \end{aligned}$$

Since $\sigma > n(\log n)^{1/2+c}$ for some fixed constant $c > 0$, we have $\widetilde{\mathbb{E}}[f] - f(y) > 0$ as desired. □

In the remainder of the section, we prove Lemma 11.

D.1.1 Outline

We note that our method shares a similar idea which appears in [33] and [34].

We are given a random polynomial $g(x) = \left\langle \mathbf{W}, x^{\otimes 4} \right\rangle$ where $\mathbf{W}$ has independent standard Gaussian entries. We would like to construct $\widetilde{\mathbb{E}} = \widetilde{\mathbb{E}}_{\mathbf{W}}$ which has large correlation with $\mathbf{W}$. If we simply let

$$\widetilde{\mathbb{E}}[x_{i_1} x_{i_2} x_{i_3} x_{i_4}] = \frac{1}{24} \sum_{\pi \in \mathcal{S}_4} \mathbf{W}_{i_{\pi(1)}, i_{\pi(2)}, i_{\pi(3)}, i_{\pi(4)}}$$

for $\{i_1 < i_2 < i_3 < i_4\} \subseteq [n]$ and $\widetilde{\mathbb{E}}[x_T]$ be zero if $|T| \leqslant 3$, then

$$\widetilde{\mathbb{E}}[g] = \frac{1}{24} \sum_{1 \leqslant i_1 < i_2 < i_3 < i_4 \leqslant n} \left(\sum_{\pi \in \mathcal{S}_4} \mathbf{W}_{i_{\pi(1)}, i_{\pi(2)}, i_{\pi(3)}, i_{\pi(4)}} \right)^2$$

so the expectation of $\widetilde{\mathbb{E}}[g]$ over $\mathbf{W}$ would be equal to $\binom{n}{4} \approx \frac{n^4}{24}$. However, in this case $\widetilde{\mathbb{E}}$ does not satisfies the equality $\mathbf{1}^T x = 0$ nor the conditions for pseudo-expectations.

To overcome this, we first project the $\widetilde{\mathbb{E}}$ constructed above to the space of linear functionals which satisfy the equality constraints ($x_i^2 = 1$ and $\mathbf{1}^T x = 0$). Then, we take a convex combination of the projection and a pseudo-expectation to control the spectrum of the functional. The details are following:

(1) (Removing degeneracy) We establish the one-to-one correspondence between the collection of linear functionals on n-variate, even multilinear polynomials of degree at most 4 and the collection of linear functionals on $(n-1)$-variate multilinear polynomials of degree at most 4 by posing $x_n = 1$. This correspondence preserves positivity.

(2) (Description of equality constraints) Let ψ be a linear functional on $(n-1)$-variate multilinear polynomials of degree at most 4. We may think ψ as a vector in $\mathbb{R}^{\binom{n-1}{\leqslant 4}}$. Then, the condition that ψ satisfies $\sum_{i=1}^{n-1} x_i + 1 = 0$ can be written as $A\psi = 0$ for some matrix A.

(3) (Projection) Let $w \in \mathbb{R}^{\binom{n-1}{\leqslant 4}}$ be the coefficient vector of $g(x)$. Let Π be the projection matrix to the space $\{x : Ax = 0\}$. In other words,

$$\Pi = Id - A^T(AA^T)^\dagger A$$

where Id is the identity matrix of size $\binom{n-1}{\leqslant 4}$ and $(\cdot)^\dagger$ denotes the pseudo-inverse. Let e be the first column of Π and $\psi_1 = \frac{\Pi w}{e^T w}$. Then $(\psi_1)_\varnothing = 1$ and $A\psi_1 = 0$ by definition.

(4) (Convex combination) Let $\psi_0 = \frac{e}{e^T e}$. We note that ψ_0 corresponds to the expectation operator of uniform distribution on $\{x \in \{\pm 1\}^n : \mathbf{1}^T x = 0\}$.

We will construct ψ by

$$\psi = (1-\epsilon)\psi_0 + \epsilon\psi_1$$

with an appropriate constant ϵ. Equivalently,

$$\psi = \psi_0 + \frac{\epsilon}{e^T w} \cdot \left(\Pi - \frac{ee^T}{e^T e}\right) w.$$

(5) (Spectrum analysis) We bound the spectrum of the functional $\left(\Pi - \frac{ee^T}{e^T e}\right) w$ to decide the size of ϵ for ψ being positive semidefinite.

D.1.2 Removing degeneracy

Recall that

$$g(x) = \sum_{i,j,k,l \in [n]} \mathbf{W}_{ijkl} x_i x_j x_k x_\ell.$$

Observe that g is even, i.e., $g(x) = g(-x)$ for any $x \in \{\pm 1\}^n$. To maximize such an even function, we claim that we may only consider the pseudo-expectations such that whose odd moments are zero.

Proposition 12. *Let $\widetilde{\mathbb{E}}$ be a pseudo-expectation of degree 4 on hypercube satisfying $\sum_{i=1}^n x_i = 0$. Let p be a degree 4 multilinear polynomial which is even. Then, there exists a pseudo-expectation $\widetilde{\mathbb{E}}'$ of degree 4 such that $\widetilde{\mathbb{E}}[p] = \widetilde{\mathbb{E}}'[p]$ and $\widetilde{\mathbb{E}}'[x_S] = 0$ for any $S \subseteq [n]$ of odd size.*

Proof. Let $\widetilde{\mathbb{E}}$ be a pseudo-expectation of degree 4 on hypercube satisfying $\sum_{i=1}^n x_i = 0$. Let us define a linear functional $\widetilde{\mathbb{E}}_0$ on the space of multilinear polynomials of degree at most 4 so that $\widetilde{\mathbb{E}}_0[x_S] = (-1)^{|S|}\widetilde{\mathbb{E}}[x_S]$ for any $S \in \binom{[n]}{\leqslant 4}$. Then, for any multilinear polynomial q of degree at most 2, we have

$$\widetilde{\mathbb{E}}_0[q(x)^2] = \widetilde{\mathbb{E}}[q(-x)^2] \geqslant 0.$$

Also, $\widetilde{\mathbb{E}}_0$ satisfies $\widetilde{\mathbb{E}}_0[1] = 1$ and

$$\widetilde{\mathbb{E}}_0\left[\left(\sum_{i=1}^n x_i\right) q(x)\right] = -\widetilde{\mathbb{E}}\left[\left(\sum_{i=1}^n x_i\right) q(-x)\right] = 0$$

for any q of degree 3. Thus, $\widetilde{\mathbb{E}}_0$ is a valid pseudo-expectation of degree 4 satisfying $\sum_{i=1}^n x_i = 0$.

Let $\widetilde{\mathbb{E}}' = \frac{1}{2}(\widetilde{\mathbb{E}} + \widetilde{\mathbb{E}}_0)$. This is again a valid pseudo-expectation, since the space of pseudo-expectations is convex. We have $\widetilde{\mathbb{E}}'[p(x)] = \widetilde{\mathbb{E}}[p(x)] = \widetilde{\mathbb{E}}_0[p(x)]$ since p is even, and $\widetilde{\mathbb{E}}'[x_S] = (1 + (-1)^{|S|})\widetilde{\mathbb{E}}[x_S] = 0$ for any S of odd size. □

Let $\mathcal{E}$ be the space of all pseudo-expectations of degree 4 on n-dimensional hypercube with zero odd moments. Let $\mathcal{E}'$ be the space of all pseudo-expectations of degree 4 on $(n-1)$-dimensional hypercube. We claim that there is a bijection between two spaces.

Proposition 13. *Let $\psi \in \mathcal{E}$. Let us define a linear functional ψ' on the space of $(n-1)$-variate multilinear polynomials of degree at most 4 so that for any $T \subseteq [n-1]$ with $|T| \leqslant 4$*

$$\psi'[x_T] = \begin{cases} \psi[x_{T \cup \{n\}}] & \text{if } |T| \text{ is odd} \\ \psi[x_T] & \text{otherwise.} \end{cases}$$

Then, $\psi \mapsto \psi'$ is a bijective mapping from $\mathcal{E}$ to $\mathcal{E}'$.

Proof. We say linear functional ψ on the space of polynomials of degree at most 2ℓ is *positive semidefinite* if $\psi[q^2] \geqslant 0$ for any q of degree ℓ.

Note that the mapping $\psi' \mapsto \psi$ where $\psi[x_S] = \psi'[x_{S\setminus\{n\}}]$ for any $S \subseteq [n]$ of even size is the inverse of $\psi \mapsto \psi'$. Hence, it is sufficient to prove that ψ is positive semidefinite if and only if ψ' is positive semidefinite.

($\Rightarrow$) Let q be an n-variate polynomial of degree at most 2. Let q_0 and q_1 be polynomials in $x_1, \cdots, x_{n-1}$ such that

$$q(x_1, \cdots, x_n) = q_0(x_1, \cdots, x_{n-1}) + x_n q_1(x_1, \cdots, x_{n-1}).$$

We get $\psi'[q^2] = \psi'[(q_0 + x_n q_1)^2] = \psi'[(q_0^2 + q_1^2) + 2x_n q_0 q_1]$. For $i = 1, 2$, let q_{i0} and q_{i1} be the even part and the odd part of q_i, respectively. Then we have

$$\begin{aligned} \psi'[q^2] &= \psi'[(q_{00}^2 + q_{01}^2 + q_{10}^2 + q_{11}^2) + 2x_n(q_{00}q_{11} + q_{01}q_{10})] \\ &= \psi[(q_{00}^2 + q_{01}^2 + q_{10}^2 + q_{11}^2) + 2(q_{00}q_{11} + q_{01}q_{10})] \\ &= \psi[(q_{00} + q_{11})^2 + (q_{10} + q_{01})^2] \geqslant 0. \end{aligned}$$

The first equality follows from that $\psi'[q] = 0$ for odd q. Hence, ψ' is positive semidefinite.
($\Leftarrow$) Let q be an $(n-1)$-variate polynomial of degree at most 2. Let q_0 and q_1 be the even part and the odd part of q, respectively. Then,

$$\psi[q^2] = \psi[(q_0^2 + q_1^2) + 2q_0q_1].$$

Note that $q_0^2 + q_1^2$ is even and q_0q_1 is odd. So,

$$\psi[q^2] = \psi'[(q_0^2 + q_1^2) + 2x_nq_0q_1] = \psi'[(q_0 + x_nq_1)^2] \geqslant 0.$$

Hence ψ is positive semidefinite. $\square$

In addition to the proposition, we note that ψ satisfies $\sum_{i=1}^n x_i = 0$ if and only if ψ' satisfies $1+\sum_{i=1}^{n-1} x_i = 0$. Hence, maximizing $\widetilde{\mathbb{E}}[g]$ over $\widetilde{\mathbb{E}} \in \mathcal{E}$ satisfying $\sum_{i=1}^n x_i = 0$ is equivalent to

$$\max_{\psi' \in \mathcal{E}'} \psi'[g'] \quad \text{subject to} \quad \psi' \text{ satisfies } 1 + \sum_{i=1}^{n-1} x_i = 0,$$

where $g'(x_1, \cdots, x_{n-1}) = g(x_1, \cdots, x_{n-1}, 1)$.

D.1.3 Matrix expression of linear constraints

Let $\mathcal{F}$ be the set of linear functional on the space of $(n-1)$-variate multilinear polynomials of degree at most 4. We often regard a functional $\psi \in \mathcal{F}$ as a $\binom{n-1}{\leqslant 4}$ dimensional vector with entries $\psi_S = \psi[x_S]$ where S is a subset of $[n-1]$ of size at most 4. The space $\mathcal{E}'$ of pseudo-expectations of degree 4 (on $(n-1)$-variate multilinear polynomials) is a convex subset of $\mathcal{F}$.

Observe that $\psi \in \mathcal{F}$ satisfies $1 + \sum_{i=1}^{n-1} x_i = 0$ if and only if

$$\psi\left[\left(1 + \sum_{i=1}^{n-1} x_i\right) x_S\right] = 0$$

for any $S \subseteq [n-1]$ with $|S| \leqslant 3$.

Let s, t and u be integers such that $0 \leqslant s, t \leqslant 4$ and $0 \leqslant u \leqslant \min(s,t)$. Let $M_{s,t}^u$ be the matrix of size $\binom{n-1}{\leqslant 4}$ such that

$$(M_{s,t}^u)_{S,T} = \begin{cases} 1 & \text{if } |S| = s,\ |T| = t, \text{ and } |S \cap T| = u \\ 0 & \text{otherwise} \end{cases}$$

for $S, T \in \binom{[n-1]}{\leqslant 4}$. Then, the condition that $\psi \in \mathcal{F}$ satisfying $1 + \sum_{i=1}^{n-1} x_i = 0$ can be written as $A\psi = 0$ where

$$A = M_{0,0}^0 + M_{0,1}^0 + \sum_{s=1}^{3} (M_{s,s-1}^{s-1} + M_{s,s}^s + M_{s,s+1}^s).$$

D.1.4 Algebra generated by $M^u_{s,t}$

Let m be a positive integer greater than 8. For nonnegative integers s, t, u, let $M^u_{s,t}$ be the $\binom{m}{\leqslant 4} \times \binom{m}{\leqslant 4}$ matrix with

$$(M^u_{s,t})_{S,T} = \begin{cases} 1 & \text{if } |S| = s,\ |T| = t, \text{ and } |S \cap T| = u \\ 0 & \text{otherwise,} \end{cases}$$

for $S, T \subseteq [m]$ with $|S|, |T| \leqslant 4$. Let $\mathcal{A}$ be the algebra of matrices

$$\sum_{0 \leqslant s,t \leqslant 4} \sum_{u=0}^{s \wedge t} x^u_{s,t} M^u_{s,t}$$

with complex numbers $x^u_{s,t}$. This algebra $\mathcal{A}$ is a C*-algebra: it is a complex algebra which is closed under taking complex conjugate. $\mathcal{A}$ is a subalgebra of the Terwilliger algebra of the Hamming cube $H(m, 2)$ [41], [42].

Note that $\mathcal{A}$ has dimension 55 which is the number of triples (s, t, u) with $0 \leqslant s, t \leqslant 4$ and $0 \leqslant u \leqslant s \wedge t$.

Define

$$\beta^u_{s,t,r} := \sum_{p=0}^{s \wedge t} (-1)^{p-t} \binom{p}{u} \binom{m-2r}{p-r} \binom{m-r-p}{s-p} \binom{m-r-p}{t-p}$$

for $0 \leqslant s, t \leqslant 4$ and $0 \leqslant r, u \leqslant s \wedge t$. The following theorem says that matrices in the algebra $\mathcal{A}$ can be written in a block-diagonal form with small sized blocks.

Theorem 14 ([42]). *There exists an orthogonal $\binom{m}{\leqslant 4} \times \binom{m}{\leqslant 4}$ matrix U such that for $M \in \mathcal{A}$ with*

$$M = \sum_{s,t=0}^{4} \sum_{u=0}^{s \wedge t} x^u_{s,t} M^u_{s,t},$$

the matrix $U^T M U$ is equal to the matrix

$$\begin{pmatrix} C_0 & 0 & 0 & 0 & 0 \\ 0 & C_1 & 0 & 0 & 0 \\ 0 & 0 & C_2 & 0 & 0 \\ 0 & 0 & 0 & C_3 & 0 \\ 0 & 0 & 0 & 0 & C_4 \end{pmatrix}$$

where each C_r is a block diagonal matrix with $\binom{m}{r} - \binom{m}{r-1}$ repeated, identical blocks of order $5 - r$:

$$C_r = \begin{pmatrix} B_r & 0 & \cdots & 0 \\ 0 & B_r & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_r \end{pmatrix},$$

and

$$B_r = \left(\sum_u \binom{m-2r}{s-r}^{-1/2} \binom{m-2r}{t-r}^{-1/2} \beta^u_{s,t,r} x^u_{s,t} \right)^4_{s,t=r}.$$

For brevity, let us denote this block-diagonalization of M by the tuple of matrices $(B_0, B_1, B_2, B_3, B_4)$ with $(5-r) \times (5-r)$ matrix B_r's.

D.1.5 Projection

Recall that $g'(x_1, \cdots, x_{n-1})$ is equal to

$$\sum_{i,j,k,\ell \in [n]} \mathbf{W}_{ijk\ell} x_{\{i,j,k,\ell\} \setminus \{n\}}.$$

Let $c \in \mathbb{R}^{\binom{n-1}{\leqslant 4}}$ be the coefficient vector of g'. Since entries of $\mathbf{W}$ are independent Standard Gaussians, c is a Gaussian vector with a diagonal covariance matrix $\Sigma = \mathbb{E}[cc^T]$ where

$$\Sigma_{S,S} = \begin{cases} n & \text{if } |S| = 0 \\ 12n - 16 & \text{if } |S| = 1 \text{ or } |S| = 2 \\ 24 & \text{if } |S| = 3 \text{ or } |S| = 4. \end{cases}$$

Let $w = \Sigma^{-1/2} c$. By definition, w is a random vector with i.i.d. standard normal entries.

Let $\Pi = Id - A^T(AA^T)^+ A$ where $(AA^T)^+$ is the Moore-Penrose pseudoinverse of AA^T and Id is the identity matrix of order $\binom{n-1}{\leqslant 4}$. Then, Π is the orthogonal projection matrix onto the nullspace of A. Since A, A^T and Id are all in the algebra $\mathcal{A}$, the projection matrix Π is also in $\mathcal{A}$.

Let e be the first column of Π and

$$\psi_0 := \frac{e}{e^T e} \quad \text{and} \quad \psi_1 := \frac{\Pi w}{e^T w}.$$

We have $A\psi_0 = A\psi_1 = 0$ by definition of Π, and $(\psi_0)_\varnothing = (\psi_1)_\varnothing = 1$ since $(\Pi w)_\varnothing = e^T w$.

Let ϵ be a real number with $0 < \epsilon < 1$ and $\psi = (1-\epsilon)\psi_0 + \epsilon\psi_1$. This functional still satisfies $A\psi = 0$ and $\psi_\varnothing = 1$, regardless of the choice of ϵ. We would like to choose ϵ such that ψ is positive semidefinite with high probability.

D.1.6 Spectrum of ψ

Consider the functional $\psi_0 = \frac{e}{e^T e}$. It has entries

$$(\psi_0)_S = \begin{cases} 1 & \text{if } S = \varnothing \\ -\frac{1}{n-1} & \text{if } |S| = 1 \text{ or } 2 \\ \frac{3}{(n-1)(n-3)} & \text{if } |S| = 3 \text{ or } 4, \end{cases}$$

for $S \subseteq [n-1]$ of size at most 4. We note that this functional corresponds to the degree 4 or less moments of the uniform distribution on the set of vectors $x \in \{\pm 1\}^{n-1}$ satisfying $\sum_{i=1}^{n-1} x_i + 1 = 0$.

Proposition 15. *Let ψ be a vector in $\mathbb{R}^{\binom{n-1}{\leqslant 4}}$ such that $A\psi = 0$ and p be an $(n-1)$-variate multilinear polynomial of degree at most 2. Suppose that $\psi_0[p^2] = 0$. Then, $\psi[p^2] = 0$.*

Proof. Let $\mathcal{U} = \{x \in \{\pm 1\}^{n-1} : \sum_{i=1}^{n-1} x_i + 1 = 0\}$. Note that ψ_0 is the expectation functional of the uniform distribution on $\mathcal{U}$ as we seen above. Hence, $\psi_0[p^2] = 0$ if and only if $p(x)^2 = 0$ for any $x \in \mathcal{U}$.

On the other hand, the functional ψ is a linear combination of functionals $\{p \mapsto p(x) : x \in \mathcal{U}\}$ since $A\psi = 0$. Hence, if $\psi_0[p^2] = 0$ then $\psi[p^2] = 0$ as $p(x)^2 = 0$ for any $x \in \mathcal{U}$. □

Recall that $\psi = (1-\epsilon)\psi_0 + \epsilon\psi_1$ where $\psi_0 = \frac{e}{e^T e}$ and $\psi_1 = \frac{\Pi w}{e^T w}$. Let $\psi_1' = e^T w \cdot (\psi_1 - \psi_0)$. Then,

$$\begin{aligned}\psi_1' &= \Pi w - \frac{e^T w}{e^T e} e \\ &= \left(\Pi - \frac{ee^T}{e^T e}\right) w\end{aligned}$$

and $\psi = \psi_0 + \frac{\epsilon}{e^T w}\psi_1'$. We note that $A\psi_1' = 0$ since ψ_1' is a linear combination of ψ_0 and ψ_1.

Let X_{ψ_0} and $X_{\psi_1'}$ be the moment matrix of ψ_0 and ψ_1' respectively. Let X_ψ be the moment matrix of ψ. Clearly,

$$X_\psi = X_{\psi_0} + \frac{\epsilon}{e^T w} X_{\psi_1'}.$$

Moreover, for any $p \in \mathbb{R}^{\binom{n-1}{\leqslant 2}}$ satisfying $X_{\psi_0} p = 0$, we have $X_{\psi_1'} p = 0$ by the proposition. Hence, $X_\psi \succeq 0$ if

$$\frac{\epsilon}{|e^T w|} \|X_{\psi_1'}\| \leqslant \lambda_{\min,\neq 0}(X_{\psi_0})$$

where $\lambda_{\min,\neq 0}$ denotes the minimum nonzero eigenvalue.

We note that $e^T w$ and $\|X_{\psi_1'}\|$ are independent random variables. It follows from that w is a gaussian vector with i.i.d. standard entries, and that e and $\left(\Pi - \frac{ee^T}{e^T e}\right)$ are orthogonal. Hence, we can safely bound $e^T w$ and $\|X_{\psi_1'}\|$ separately.

To bound $\|X_{\psi_1'}\|$ we need the following theorem.

Theorem 16 (Matrix Gaussian ([43]))**.** *Let $\{A_k\}$ be a finite sequence of fixed, symmetric matrices with dimension d, and let $\{\xi_k\}$ be a finite sequence of independent standard normal random variables. Then, for any $t \geqslant 0$,*

$$\Pr\left[\left\|\sum_k \xi_k A_k\right\| \geqslant t\right] \leqslant d \cdot e^{-t^2/2\sigma^2} \quad \textit{where} \quad \sigma^2 := \left\|\sum_k A_k^2\right\|.$$

For each $U \subseteq [n-1]$ with size at most 4, let Y_U be the $\binom{n-1}{\leqslant 2} \times \binom{n-1}{\leqslant 2}$ matrix with entries

$$(Y_U)_{S,T} = \begin{cases} 1 & \text{if } S \oplus T = U \\ 0 & \text{otherwise.} \end{cases}$$

We can write $X_{\psi_1'}$ as

$$X_{\psi_1'} = \sum_{\substack{U \subseteq [n-1] \\ |U| \leqslant 4}} (\psi_1')_U Y_U.$$

Since $\psi_1' = \left(\Pi - \frac{ee^T}{e^T e}\right) w$, we have

$$\begin{aligned} X_{\psi_1'} &= \sum_{\substack{U \subseteq [n-1] \\ |U| \leqslant 4}} \sum_{\substack{V \subseteq [n-1] \\ |V| \leqslant 4}} \left(\Pi - \frac{ee^T}{e^T e}\right)_{U,V} w_V Y_U \\ &= \sum_V w_V \left(\sum_U \left(\Pi - \frac{ee^T}{e^T e}\right)_{U,V} Y_U \right). \end{aligned}$$

By Theorem 16, $\|X_{\psi_1'}\|$ is roughly bounded by $(\|\Sigma_X\| \log n)^{1/2}$ where

$$\Sigma_X := \sum_V \left(\sum_U \left(\Pi - \frac{ee^T}{e^T e}\right)_{U,V} Y_U \right)^2.$$

Proposition 17. *For each $I, J \in \binom{[n-1]}{\leqslant 2}$, the (I,J) entry of Σ_X only depends on $|I|$, $|J|$ and $|I \cap J|$, i.e., Σ_X is in the algebra $\mathcal{A}$.*

Proof. Note that

$$\begin{aligned} \Sigma_X &= \sum_V \sum_{U_1, U_2} \left(\Pi - \frac{ee^T}{e^T e}\right)_{U_1,V} \left(\Pi - \frac{ee^T}{e^T e}\right)_{V,U_2} Y_{U_1} Y_{U_2} \\ &= \sum_{U_1,U_2} \left(\left(\Pi - \frac{ee^T}{e^T e}\right)^2 \right)_{U_1,U_2} Y_{U_1} Y_{U_2} \\ &= \sum_{U_1,U_2} \left(\Pi - \frac{ee^T}{e^T e}\right)_{U_1,U_2} Y_{U_1} Y_{U_2}. \end{aligned}$$

Hence,

$$(\Sigma_X)_{I,J} = \sum_{K \in \binom{[n-1]}{\leqslant 2}} \left(\Pi - \frac{ee^T}{e^T e}\right)_{I \oplus K, J \oplus K},$$

which is invariant under any permutation π on $[n-1]$ as $\Pi - \frac{ee^T}{e^T e}$ is. It implies that $\Sigma_X \in \mathcal{A}$. □

The block-diagonalization of Σ_X is $(u_0u_0^T, u_1u_1^T, u_2u_2^T, 0, 0)$ where

$$\begin{aligned} u_0 &= \sqrt{\frac{n(n-3)(n-5)}{3n-14}}\begin{bmatrix}1 & -\frac{1}{\sqrt{n-1}} & -\sqrt{\frac{n-2}{2(n-1)}} & 0 & 0\end{bmatrix}^T \\ u_1 &= \sqrt{\frac{(n-6)(3n^4-24n^3+59n^2-66n+32)}{2(n-1)(n-2)(3n-14)}}\begin{bmatrix}1 & -\frac{1}{\sqrt{n-3}} & 0 & 0\end{bmatrix}^T \\ u_2 &= \sqrt{\frac{(n-6)(3n^4-24n^3+59n^2-66n+32)}{2(n-1)(n-3)(3n-14)}}\begin{bmatrix}1 & 0 & 0\end{bmatrix}^T. \end{aligned}$$

Hence, $\|\Sigma_X\|$ is equal to the maximum of $\|u_i\|^2$ among $i = 0, 1, 2$, which is at most $(1/2 + o_n(1))n^2$. We get the following result:

Proposition 18. *If $\epsilon < o_n\left(\frac{1}{n\sqrt{\log n}}\right)$, then with probability $1-o_n(1)$ the moment matrix X_ψ is positive-semidefinite.*

Proof. By theorem 16, we have

$$\Pr\left(\|X_{\psi_1'}\| \geqslant t\right) \leqslant \binom{n-1}{\leqslant 2} \cdot e^{-t^2/2\|\Sigma_X\|}.$$

Let $t = 3n\sqrt{\log n}$. Since $\|\Sigma_X\| \leqslant (1/2 + o(1))n^2$, we have $\|X_{\psi_1'}\| \leqslant 3n\sqrt{\log n}$ with probability $1 - n^{-\Omega(1)}$. On the other hand, note that

$$\Pr\left(|e^Tw| \leqslant t\right) \leqslant \frac{t}{\sqrt{2\pi}}.$$

It implies that $|e^Tw| > g(n)$ with probability $1 - o_n(1)$ for any $g(n) = o_1(1)$. Thus,

$$\frac{\|X_{\psi_1'}\|}{|e^Tw|} \lesssim \frac{n\sqrt{\log n}}{g(n)}$$

almost asymptotically surely. Together with the fact that $\lambda_{\min,\neq 0}(X_{\psi_0}) = 1 - o_n(1)$, we have $X_\psi \geq 0$ whenever $\epsilon < \frac{g(n)}{n\sqrt{\log n}}$ for some $g(n) = o_n(1)$.

□

D.1.7 Putting it all together

We have constructed a linear functional ψ on the space of $(n-1)$-variate multilinear polynomials of degree at most 4, which satisfies (i) $\psi[1] = 1$, (ii) ψ satisfies $\sum_{i=1}^{n-1} x_i + 1 = 0$, and (iii) $\psi[p^2] \geqslant 0$ for any p of degree 2. It implies that ψ is a valid pseudo-expectation of degree 4.

Now, let us compute the evaluation of

$$g'(x) = \sum_{i,j,k,\ell \in [n]} \mathbf{W}_{ijk\ell} x_{\{i,j,k,\ell\}\setminus\{n\}}$$

by the functional ψ. Recall that c is the coefficient vector of g' and $w = \Sigma^{-1/2}c$ where $\Sigma = \mathbb{E}[cc^T]$. We have

$$\begin{aligned}\psi[g'] = c^T\psi &= w^T\Sigma^{1/2}\left(\frac{e}{e^Te} + \frac{\epsilon}{e^Tw}\left(\Pi - \frac{ee^T}{e^Te}\right)w\right) \\ &= \frac{e^T\Sigma^{1/2}w}{e^Te} + \epsilon \cdot \frac{w^T\Sigma^{1/2}\left(\Pi - \frac{ee^T}{e^Te}\right)w}{e^Tw}.\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}\left[w^T\Sigma^{1/2}\left(\Pi - \frac{ee^T}{e^Te}\right)w\right] &= \left\langle \Sigma^{1/2}\left(\Pi - \frac{ee^T}{e^Te}\right), \mathbb{E}[ww^T]\right\rangle \\ &= \operatorname{tr}\left(\Sigma^{1/2}\left(\Pi - \frac{ee^T}{e^Te}\right)\right),\end{aligned}$$

which is at least $(\sqrt{6}/12 - o_n(1))n^4$. Also, $|e^Tw| = O(1)$ and $|e^T\Sigma^{1/2}w| = O(n)$ with high probability. Hence, with probability $1 - o_n(1)$, we have

$$\psi[g'] \gtrsim O(n) + \frac{n^4}{n(\log n)^{1/2+o(1)}} \gtrsim \frac{n^3}{(\log n)^{1/2+o(1)}}.$$